# Что такое семантическая цифровая библиотека

© В.А. Серебряков
Вычислительный Центр РАН,
Москва
serebr@ccas.ru

## Аннотация

В последние годы в литературе значительное внимание уделяется так называемым «семантическим цифровым библиотекам». Что это такое? В работе на основе анализа проектов и публикаций делается попытка определить такие понятия, как «электронная библиотека», «цифровая библиотека», «семантическая цифровая библиотека».

## 1 Что такое библиотека

Прежде всего необходимо определить, что такое цифровые библиотеки (в отличие от «электронных» библиотек, под которыми будем понимать программное обеспечение обычных, «книжных» библиотек, (часто называемое АБИС – автоматизированная библиотечная информационна система). Переходя от электронных библиотек к цифровым, можно было бы сказать, что цифровая библиотека – это электронная с цифровым контентом. Это было так в первое время, однако затем контент стал включать живопись, видео и т.д. Так что такое определение устарело.

Электронная библиотека сегодня – это прежде всего формат MARC. Машиночитаемая каталогизация (MARC) – это идея разработки общей системы описания ресурсов библиотек. Она берет начало от работ библиотеки Конгресса еще в 1960-е годы по разработке формата LC MARC для собственных нужд, когда начали использовать компьютеры. MARC-запись стала электронным аналогом бумажного каталога и карточки, который может быть создан в Библиотеке Конгресса и продаваться в библиотеки, которым не придется тратить свои ресурсы для создания почти идентичного набора уже предложенной информации. Даже если библиотека не имеет собственной компьютерной системы, совместимой с MARC, она может приобрести напечатанные на компьютере каталожные карточки, заполненные в соответствии с библиографическими записями в MARC файлах Библиотеки Конгресса. Формат MARC введен в 1987 году, а в 1999 году появился формат MARC21, созданный в результате слияния библиографических форматов США и Канады, и призванный стать «Библиографическим форматом 21 века». MARC 21 является эволюцией исходного LC MARC. Последующие издания были опубликованы в 1990 году, 1994 и 2000 года. MARC21 поддерживается библиотекой конгресса США, и используется в основном в США и Великобритании. В настоящее время существуют две группы, ответственные за рассмотрение и пересмотр формата MARC 21: Комитет Marbi (машиночитаемая библиографическая информация) и Консультативный комитет MARC. Каждый год появляется новое официально опубликованное издание MARC 21 в Интернете с учетом изменений в библиотечной среде.

В 1977 году был выпущен формат UNIMARC, который был призван стать посредником между любыми национальными стандартами MARC. Формат UNIMARC включает поля, необходимые для описания монографий, сериальных изданий, нотных записей, видео, изображений и прочих документов. Эти поля делятся на общие, использующиеся при описании любого вида документа, и специфические, используемые только для описания их определенных видов. Этот формат поддерживается международной организацией IFLA, и используется в основном в Европе и Азии.

Программное обеспечение обычной, «книжной» библиотеки состоит из нескольких базовых компонент, которые можно разделить на два блока: блок работы с читателями, включающий проверку обслуживание прав читателя, выдачу, прием и заказ книг и т.д., и блок обслуживания фонда, куда условно можно отнести заказ и покупку литературы, списание, постановку на учет (включая подготовку библиографических записей) и т.д.

### 1.1 Что такое цифровая библиотека

Самое простое определение, которое можно дать – «Цифровая библиотека – это электронная библиотека с цифровым контентом». Более неопределенно можно сказать, что ЦБ – это информационная система, основным назначением которой является доступ к цифровым материалам. Здесь подчеркивается, что организация самой информационной системы может быть произвольной, важно, что вся эта организация нацелена на доступ к цифровому контенту (тексты, видео, аудио и т.д.). Wikipedia определяет цифровую библиотеку так: «Цифровая библиотека – это библиотека, в которой коллекции хранятся в

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

цифровых форматах (в отличие от печатного, микроформата или другого носителя) и собираются с помощью компьютеров».

Еще одним предшественником ЦБ были библиотеки программ. Изначально они были предназначены для размещения и использования объектов операционных систем: библиотеки для связывания объектного кода, библиотеки исходного кода, компилированного объектного кода для повторного использования. Они возникли из потребностей ОС, чтобы находить и загружать компоненты, и того факта, что существующие файловые системы не обеспечивали работу в реальном времени. Возникшая структура остается по-существу и сегодня; за справочником элементов библиотеки, который дает имена и другие метаданные содержащихся объектов, следуют в том же наборе данных или файле двоичные данные для каждого объекта, на который ссылается элементы каталогов.

Часто в связи с ЦБ используется термин «коллекция», под которым имеют в виду определенным образом организованный набор как правило однородных цифровых объектов.

«Основой цифровой библиотеки является коллекция цифровых объектов, которые представляют интерес как таковые (в первую очередь для чтения, прослушивания, просмотра людьми, но и для использования программами), а не просто указания на другие объекты. Примеры:

- Коллекция оцифрованных книг (в отличие от просто интернет-каталога),
- Коллекция биографий (в отличие от базы данных персонала),
- Коллекция устных историй,
- Набор программных модулей (многие так и рассматривают DL)» [5].

Рассмотрим теперь определения цифровых библиотек, приводимые различными авторами и их критику.

ru.wikipedia.org дает такое определение.

**Электро́нная библиоте́ка** – упорядоченная коллекция разнородных электронных документов (в том числе книг, журналов), снабженных средствами навигации и поиска. Может быть веб-сайтом, где постепенно накапливаются различные тексты (чаще литературные, но также научные и любые другие, вплоть до компьютерных программ) и медиафайлы, каждый из которых самодостаточен и в любой момент может быть востребован читателем. Электронные библиотеки могут быть универсальными, стремящимися к наиболее широкому выбору материала (как Библиотека Библиотека Максима Мошкова или Либрусек), и более специализированными, как Фундаментальная электронная библиотека или проект Сетевая Словесность.

Возникает ряд вопросов. Что значит упорядоченная коллекция? Что значит разнородных? Средствами навигации снабжены документы? Только навигации и поиска, т.е. никаких связей нет. Что значит «Может быть веб-сайтом»? А может не быть? Т.е. может быть не привязана к Интернет. «Постепенно накапливаются», а если не постепенно? Почему чаще литературные? Самодостаточен, т.е. в некотором роде отсутствуют связи между ресурсами. Библиотека Мошкова и Фундаментальная электронная библиотека радикально отличаются друг от друга: первая просто набор файлов, вторая пронизана ссылками (HTML).

«Под электронными библиотеками понимаются информационные системы, которые автоматизируют решение основных проблем организации работы с документами» [7].

В соответствии с таким определенем наилучшей электронной библиотекой является система документооборота.

В [4] дается следующее определение ЦБ (в оригинале «Электронной библиотеки»).

«Электронные библиотеки – это организации, в том числе специализированный персонал, представляющие доступ читателей к электронным ресурсам. Кроме того они выполняют отбор, структурирование, предоставление интеллектуального доступа, интерпретацию, распространение, сохранение целостности и обеспечение сохранности в течение длительного времени наборов электронных документов для удобного доступа к ним определенным сообществам специалистов.

В соответствии с данным определением основными компонентами ЭБ являются: специалисты, информационные ресурсы (документы) и информационные технологии.

Электронные библиотеки реализуют набор функций для обеспечения читателям полного доступа к множеству распределенных и разнородных документов, содержащих информацию и знания, интегрируя их в единое информационное пространство».

### 1.2 Цифровые библиотеки или информационные системы?

С другой стороны, ясно, что цифровую библиотеку можно считать информационной системой. А почему бы не считать любую информационную систему цифровой библиотекой? Любая информационная система в конце концов имеет дело с цифромым контентом. Есть ли все-таки разделительная линия, выделяющая цифровые библиотеки из общего класса информационных систем?

В [1, 2, 4] описаны некоторые проблемы ЭБ, основными из которых являются следующие:

- Проблема интеграции разнородной информации (электронных ресурсов,

пользовательских профилей, таксономий) на основе различных метаданных, содержащих выразительные семантические описания.

- Проблема поддержки взаимодействия с другими информационными системами (и не толькоЭБ) либо с помощью метаданных, либо на уровне коммуникации или с помощью обеих возможностей. При этом в качестве единого языка взаимодействия между системами может использоватьсяязык RDF (Resource Description Framework).

- Проблема обеспечения надежного, удобного и адаптируемого поиска и интерфейсов просмотра электронных документов, усиленных работой с семантикой [7].

«ЦБ можно охарактеризовать диапазоном целей, которым она служит, или областью в которой он работает, например, обучение, образование, электронное правительство, электронная коммерция (B2B или B2C), развлечения, и более специфические цели, такие как обеспечение информации, связанной с работой, поддержка домашних заданий студентов, поддерживая внутренней работы организации, поддержка клиентов организации, поддержка связи между пользователями и т.д.» [5].

«1. ЦБ имеет много функций и должна интегрировать поддержку информационного поиска, задачи пользовательской работы, производство информации и сотрудничество.

2. ЦБ связывает многие виды информационных объектов в различных форматах (в том числе документы и базы данных) во всех средствах массовой информации в сложную структуру» [5].

Рассмотрим еще несколько определений ЦБ в контекте информационных систем.

«Термин Цифровая библиотека (ЦБ) используется для диапазона систем, от цифрового объекта и хранилищ метаданных, системы ссылка-связь, архивов и систем управления контентом до сложных систем, которые объединяют в себе передовые цифровые библиотечные услуги и поддержку научных исследований и практических сообществ» [5].

*Ничего специфичного для ЦБ в этих определениях нет, это все также относится и к информационным системам*.

Рассмотрим, как некоторые авторы определяют функции ЦБ [5].

«Цифровые библиотеки сталкиваются со многими проблемами, в том числе:

- Поиск текста, изображения, звука и составных объектах мультимедиа.
- Семантически улучшенный поиск для извлечения из свободного текста и изображения и лучшего использования проставленных пользователем меток.
- Многоязычный поиск.
- Поиск во многих системах синтаксического и семантического взаимодействия.
- Нахождение ответов, а не только документов; рассуждения и логический вывод».
- Интеграция многих форматов сохранения.
- Интеграция библиотек, архивов, музеев а также баз данных и других информационных систем.
- Интеграция чтение / просмотр / прослушивание, доступ к базе данных, обработка данных и создание.
- Интеграция издательских и коммуникационных платформ».
- Сервисы Распространение и уведомления. Современные цифровые библиотеки должны помочь своим пользователям в доступе к метаданным в различных форматах, позволяющих, среди других, построения мэшапы сервисов и контента.
- Сервисы безопасности и политики Assurance. Библиотека должна приспосабливаться к различным усилениям политики; она должна обеспечить гибкие механизмы аутентификации и контроля доступа.
- Сервисы сохранения. Цифровая библиотека должна обеспечить управление версиями, архивирования (резервного копирования и восстановления) а также, отслеживания происхождения (особенно в контексте открытого мирового подхода семантических и социальных технологий), и отслеживание истории событий, связанных с информационными объектами. Должно быть обеспечено, что отношения между объектами и дополненая информация поддерживаются сервисами сохранения.
- Сервисы обеспечения качества. Особое внимание следует уделять качеству сервисов на основе метаданных; семантическая цифровая библиотека должны обеспечить эффективность, безопасность и семантику поддержки метаданных. Эффективность может быть достигнута, например, путем жесткого кодирования части метаданных; ограничений на действия, которые могут быть выполнены над метаданными, могут повысить уровень безопасности. Семантика метаданных можно определить через значения новых концепций».

*Из вышеприведенного можно видеть, что при таких определениях любую информационную систему можно рассматривать как цифровую библиотеку.*

## 2. Что такое семантическая цифровая библиотека

Само по себе слово «семантический» означает не более, чем «смысловой», т.е. в отрыве от контекста не означает ничего. Этот термин (когда-то используемый в теории языков программирования) стал активно употребляться в контексте «семантический WEB» в противовес

«несемантическому WEB», основанному на гиперссылках. Фактически сегодня под «семантической моделью WEB» имеется в виду использование RDF модели для представления информции. Но что такое RDF модель? Это всего навсего использование бинарных отношений, т.е. связей, между объектами и соответствующие словари RDF, обобщающие и стандартизующие их использование. Это внесло колоссальный прорыв в технологии WEB. Но в конце концов, практически все данные, в частности, конечно, и данные цифровых библиотек, хранятся сегодня в реляционных базах данных, также представляющих собой отношения, только вообще говоря, многоместные.

Термин «семантический» не вносит ничего нового в технологии цифровых библиотек. Единственное, что может быть тут стоит отметить, что в обычных цифровых библиотеках эти связи между объектами используются недостаточно активно, хотя в рамках формата MARC, разработанного Библиотекой Конгресса США, предусмотрены так называемые «авторитетные» файлы, хранящие инофрмацию о персонах и организациях. Но эти данные недостаточно формализованы, чтобы их легко можно было использовать для установления всех необходимых связей.

Поэтому термин «Цифровые семантические библиотеки» осмысленно употреблять только в контексте WEB, а именно имея в виду интеграцию цифровых библиотек в контекст семантического WEB». А это означает:

- Разработку стандартов обмена RDF информацией. В качестве примера можно привести онтологии MADS и MODS, разработанные Библиотекой Когресса США для авторитетных файлов и библиографических записей.
- «семантическую» интеграцию библиотек между собой, т.е. возможность, способность цифровых библиотек обмениваться такой информацией.
- «погружение» цифровых библиотек в семантический WEB, т.е. интеграцию с другими, небиблиотечными данными, например, с соцсетями.
- Взаимодействие с данными из Linked Open Data (LOD), например, извлечение данных из LOD в библиотеку и наоборот, публикация собственных данных в LOD.

«Включение семантических данных и обработки в DL предполагает использование метаданных объектов в такой библиотеке и обеспечение доступа пользователей к семантически более мощным поисковым системам. Метаданные, как правило, выражается в Синтаксисе RDF» [3].

*Интересно отметить еще одно обстоятельство. В контексте цифровых семантических библиотек часто упоминают онтологии. Насколько это важно и характерно именно для цифровых семантических библиотек? Опять происходит некая подмена. Онтологии в современном понимании могут использоваться в трех целях: 1) как модель данных более высокого уровня по сравнению с использовавшимися моделями раньше, а именно моделью «сущность-связь» и объектной; 2) для поддержки интеграции данных в пространстве Интернет и 3) для реализации восзможности осуществления логического вывода. Для реализации 1-й цели в ЦСБ онтологии используются в той же мере, в какой они используются для разработки информационных системных в прикладных областях. Для реализации 2-й цели отологии активно используются в той же мере, в какой они используются для интеграции данных в Интернет. Для 3-й цели в приложении к ЦСБ онтологии не используются.*

*В контексте ЦБ упоминаются соц. сети, обучающие системы, архивные системы и их связь с пользователями. Все это было и не называлось ЦБ.*

«Основной целью семантической цифровой библиотеки является предоставление нахождения информации превосходящее решения, обеспечиваемые текущими цифровыми библиотеками. Пользователи должны иметь возможность использовать взаимосвязанную информацию о ресурсах в процессе просмотра, фильтрацию или нахождение подобных информационных объектов. Средства уточнения запроса должны адаптировать свои результаты для решений, соответствующим пользовательским профайлам; средства должны использовать сложные семантические отношения между результатами. Наконец, семантическая цифровая библиотека должна предлагать различные рекомендательные сервисы, например, на основе контекста и ресурса (ресурсов) или аннотации на основе совместной фильтрации. Поисковая система должна позволять использовать информацию о различных типах носителей, сложных объектах, потокового и пространственно-временных ресурсах. В случае ресурсов со сложными аннотациями важно поддерживать поиск на основе содержимого вместе с алгоритмами поиска, основанными на сходстве. В случае гетерогенных конкурентных сетей контент-провайдеров, семантическая цифровая библиотека должна осуществлять алгоритмы запроса, основанные на торговле, для поддержки пользователей в их поиске» [3].

«Одной из наиболее отличительных особенностей семантических цифровых библиотек является дополнительное пополнение аннотаций исходной информации, представляемые в ходе процесса загрузки ресурса. Ожидается, что семантические цифровые библиотеки могут обеспечить как автоматизированные, так и пользовательские аннотации. Последние должны использовать силу социальных сетей, то есть аннотации сообщества, пометки, и рейтинг» [3].

## Заключение

Резюмируя вышеприведенный краткий обзор, можно остановиться на следующих определениях.

Электронная Библиотека (ЭБ, АБИС) – средство автоматизации работы обычных, «книжных» библиотек, основанное как правило на технологиях MARC.

Цифровая Библиотека (ЦБ) – информационная система, ориентированная на действия (поиск, доступ и т.д.) с цифровым контентом (тексты, аудио, видео и т.д.). В этом смысле ЦБ може быть, а может и не быть ЭБ.

Семантическая Цифровая Библиотека (СЦБ) – ЦБ, ориентированная на интеграцию в Semantic Web.

## Литература


[1] Ding Hao. A semantic search framework in peer-to-peer based digital libraries. – NTNU, Norway, 2006.

[2] Sebastian Ryszard Kruk, Adam Westerki, and Ewelina Kruk. Architecture of Semantic Digital // Semantic Digital Libraries / Editors: Sebastian Ryszard Kruk, Bill McDaniel. – Springer, 2009.

[3] Sebastian Ryszard Kruk and Bill McDaniel. Goals of Semantic Digital Libraries // Semantic Digital Libraries / Editors: Sebastian Ryszard Kruk, Bill McDaniel. – Springer, 2009.

[4] A.A. Shiri. Digital library research: current developments and trends // Library Review. – 2003. –Vol. 52. – P. 198–202.

[5] Dagobert Soergel. Digital Libraries and Knowledge Organization // Semantic Digital Libraries / Editors: Sebastian Ryszard Kruk, Bill McDaniel. – Springer, 2009.

[6] Sukhdev Singh. Digital Library: Definition to Implementation [Электронный ресурс]. – http://arizona.openrepository.com/arizona/bitstream/ 10150/106534/1/ lecture_rcc_26jul03.pdf

[7] Ле Хоай, А.Ф. Тузовский, Разработка семантических электронных библиотек. Доклады ТУСУРа, № 2 (24), часть 2, декабрь 2011.


## Semantic digital libraries. What is it?


Vladimir Serebryakov



In recent years, considerable attention is paid to the so-called “semantic digital libraries”. What is it? In this paper, based on analysis of projects and publications an attempt is made to define concepts such as “electronic library”, “digital library”, “semantic digital library”.

# Международная профессиональная ассоциация разработчиков научных информационных систем euroCRIS и ее главный продукт CERIF

© С.И. Паринов
Центральный экономико-математический институт РАН,
Москва
sparinov@gmail.com

**Аннотация**

Международная профессиональная ассоциация разработчиков научных информационных систем euroCRIS с 2000 года занимается сбором, систематизацией и распространением информации о накопленном в разных странах опыте и используемых подходах. Ассоциация через свои рабочие группы разрабатывает, осуществляет поддержку и продвижение единых форматов научных данных (CERIF), унифицированных подходов к построению научных информационных систем (CRIS), а также участвует в формировании условий для эффективной глобальной интероперабельности научных информационных систем.

## 1 Введение

Международная профессиональная ассоциация разработчиков научных информационных систем euroCRIS (www.eurocris.org) была зарегистрирована как некоммерческая организация в 2000 г. Основная миссия ассоциации – продвижение интероперабельности в научном сообществе через CERIF, где CERIF (Common European Research Information Format) по состоянию на данный момент является главным продуктом euroCRIS.

В более широком плане euroCRIS занимается развитием модели современной научной информационной системы (концепция CRIS, [1-2,4-5]), включая создание комплекса необходимых условий (одним из важнейших здесь является формат научных данных CERIF, [3,7]) для полноценных взаимодействий между CRIS, принадлежащих разным организациям независимо от их национальной принадлежности. Фактически euroCRIS является одним из ключевых участников разработки концепции мировой онлайновой научной инфраструктуры и в этом качестве определяет основные подходы к созданию международной системы взаимосвязанных научных информационных систем.

За прошедшие годы ассоциация достигла больших успехов. В настоящее время она объединяет представителей из более 100 организаций из 40 стран, включая Северную Америку, некоторые страны Азии и др. Членами euroCRIS по состоянию на 2014 г. являются также 4 организации из России.

Тьюториал позволит получить представление о деятельность рабочих групп euroCRIS, возможностях участия в их работе и возможной пользе от этого для российских научных организаций и физических лиц, а также о формате научных данных CERIF, который, в частности, рекомендован Европейской Комиссией странам-членам ЕС для создания новых научных информационных систем, а также для развитию существующих в целях придания им свойств интероперабельности и интегрируемости с онлайновой научной инфраструктурой.

## 2 Рабочие группы euroCRIS

Для реализации заявленной миссии в euroCRIS действуют 7 рабочих групп. Кроме этого каждые два года проводятся масштабные научные конференции и два раза в год – рабочие встречи и семинары членов euroCRIS.

### 2.1 Рабочая группа CERIF

Рабочая группа CERIF осуществляет поддержку и дальнейшее развитие общеевропейского формата научных данных, что включает CERIF-XML, представляющего собой формат для обмена данными, и канонические для CERIF словари семантических терминов. Одним из результатов эта работы является поддержка и актуализация некоторого количества документов (спецификации, схемы, скрипты), которые используются в работе других рабочих групп euroCRIS. Поддерживаются постоянные взаимодействия с пользователями

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

CERIF, включая вендоров, основанных на CERIF, коммерческих приложений. Это делается в целях обеспечения CRIS максимально возможную интероперабильность. Руководителем рабочей группы (июль 2014) является Jan Dvořák (Чехия).

### 2.2 Рабочая группа Institutional Repositories (CRIS-IR)

Целью рабочей группы CRIS-IR является дальнейшее развитие подходов и технологий для связывания CRIS систем и репозиториев. В частности, открытых репозиториев научных публикаций, а также репозиториев данных и программного обеспечения. Эта работа включает проработку архитектуры, метаданных и механизмов связывания. Среди нерешенных вопросов в этой области – синтаксис, семантика и программное обеспечение. В частности, рабочая группа обеспечивает коммуникации и согласование точек зрения двух сообществ: разработчиков/менеджеров CRIS и репозиториев. Руководителем рабочей группы (июль 2014) является Danica Zendulková (Словакия).

### 2.3 Рабочая группа Best Practice

Целями рабочей группы Best Practice являются:

(1) регистрация в информационной системе DRIS примеров современной практики создания и использования CRIS систем, включая использование CERIF, примеры интеграции с репозиториями открытого доступа, примеры соединения CRIS с онлайновой научной инфраструктурой, создания научного информационного пространства на основе интегрированного контента CRIS систем, а также и другие связанные с CRIS инициативы;

(2) распространение информации о передовом опыте и помощь разработчикам и пользователям CRIS систем в применение новых концепций, примеров дизайна и технологий;

(3) сведение главных действующих лиц в разработке CRIS систем и онлайновой научной инфраструктуры для достижения взаимных выгод. Руководителем рабочей группы (июль 2014) является Pablo de Castro (Британия).

### 2.4 Рабочая группа Projects

Деятельность рабочей группы «Проекты» направлена на соединение знаний и умений членов euroCRIS в области научных информационных систем как в части контента, так и технологий, для участия в мероприятиях и проектах, инициируемых euroCRIS или другими организациями (партнеры, Европейская Комиссия, национальные / региональные власти и т.п.), на координацию внутренних проектов euroCRIS и на использование возникающих возможностей. Руководителем рабочей группы (июль 2014) является Valérie Brasse (Франция).

### 2.5 Рабочая группа CRIS Architecture and Development

Рабочая группа CRIS Architecture and Development занимается развитием программного обеспечения CRIS, которое может совместно использоваться внутри сообщества CRIS разработчиков. Главный приоритет – развитие референсной версии CRIS системы и определение стандартных API для программного доступа к данным CERIF-CRIS систем. Также в деятельность рабочей группы входят вопросы как анализ задач, спецификация и архитектура программного обеспечения, отбор технологий или продуктов, создаваемых другими организациями, модели взаимодействий пользователей, управление процессом внедрения и тестирование. Руководителем рабочей группы (июль 2014) является Nikos Houssos (Греция).

### 2.6 Рабочая группа Linked Open Data

Миссия рабочей группы Linked Open Data состоит в обеспечении представления в CERIF связанных и семантических данных вместе с развитием и поддержкой необходимых сервисов. В частности это означает: расширение структуры CERIF до конструкций, требующихся для представления связанных и семантических данных; рекомендации по поводу доступных связанных данных, принадлежащих другим областям, но которые могут быть интересны с точки зрения их представления в CERIF; развитие сервисов «open source» для требуемых трансформаций. Руководителем рабочей группы (июль 2014) является Miguel-Angel Sicilia (Испания).

### 2.7 Рабочая группа Indicators

Целью рабочей группы «Индикаторы» является развитие программы исследований и сбора «лучших образцов» в области использования индикаторов (наукометрия, библиометрия) для целей оценки результативности ученых и организаций. Рабочая группа создает каталог известных методик оценивания вместе с анализом их эффективности. Разработка CERIF совместимых сервисов для оценки научной результативности, включая широко используемые национальные и международные методики. Руководителем рабочей группы (июль 2014) является Grete Christina Lingjærde (Норвегия).

## 3 CERIF

Самое общее представление о формате научных данных CERIF дает рисунок 1.

Кроме этого существуют несколько концептуальных уровней: логический уровень (сущности и их связи); физический уровень (определение данных в виде команд к СУБД) и семантический уровень (декларированная семантика).

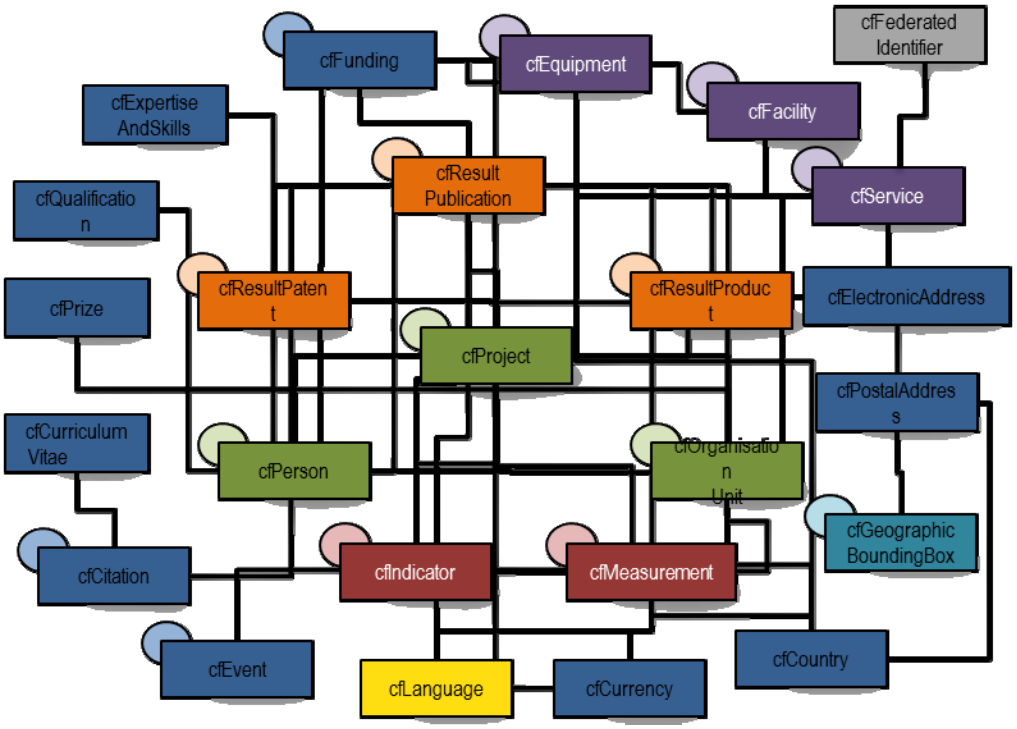



Рис. 1. Диаграмма основных типов объектов CERIF и связей между ними

Источник: Jan Dvorak. CERIF Tutorial, 2013 http://www.eurocris.org/Uploads/Web%20pages/CERIFtutorial/CERIF%20tutorial_Porto_13.11.2013%20-%20Jan%20Dvorak.pptx

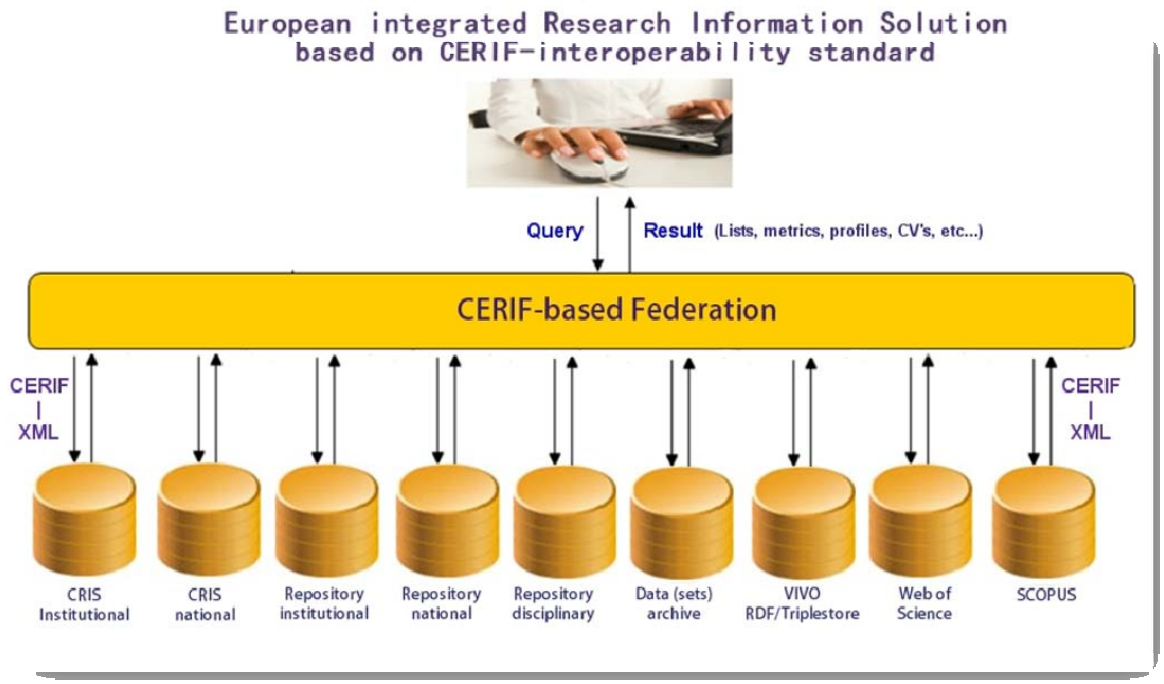



Рис. 2. Примеры использования CERIF для интероперабельности между различными ИС

Источник: Keith Jeffery, Presentation “Current Research Information Systems in Europe (and wider) for Evaluation and Management. Opportunities and Trends”

CERIF является результатом работы международного сообщества экспертов. Поэтому он обобщает разнообразный опыт и представляет собой консенсус взглядов и подходов к стандартизации описания научных данных. CERIF является гибким и расширяемым форматом. Включает богатые семантические средства. Поддерживает многоязыковость.

Все это объясняет растущую популярность CERIF для построения научных информационных систем различного вида (см. рис. 2).

CERIF является национальным стандартом представления научных данных в 10 европейский странах (UK, NO, BE, IT, DE, IS, DK, SE, CZ, SK). На его основе реализованы крупные национальные научные информационные системы (например, CRIStiN (NO); FRIS (BE)). В России CERIF использован для создания подсистемы учета РНТД в АСУ РИД РАН, а также в ряде других проектов (например, в новосибирском RU-CRIS). Система Соционет является CERIF совместимой [6].

## Литература


[1] Zimmerman, E.; Jeffery, K.G.; ‘The Need for a Standardised Current Research Information System (CRIS): A Call to Arms’ in A Nase, G van Grootel (Eds) Proceedings CRIS2004 Conference, Leuven University Press ISBN 90 5867 3839 May 2004 pp 153-160

[2] Jeffery K., Asserson A.: ‘CERIF-CRIS for the European e-Infrastructure. Data Science Journal v9 2010 http://www.codata.org/dsj/special-cris.html

[3] Jeffery K., Asserson A.: ‘The CERIF Model as the Core of a Research Organisation’ Data Science Journal v9 2010 http://www.codata.org/dsj/special-cris.html

[4] Asserson A., Jeffery K.: ‘CRIS and Institutional Repositories’. Data Science Journal v9 2010 http://www.codata.org/dsj/special-cris.html

[5] Joerg, Brigitte, Ivan Ruiz-Rube, Miguel-Angel Sicilia, JAN DVOŘÁK, Keith Jeffery, Thorsten Hoellrigl, Henrik S. Rasmussen, Andreas Engfer, Thomas Vestdam, and Elena Garcia Barriocanal. "Connecting closed world research information systems through the linked open data web." International Journal of Software Engineering and Knowledge Engineering 22, no. 03 (2012): 345-364.

[6] Parinov S, Kogalovsky M and Lyapunov V. A Challenge of Research Outputs in GL Circuit: From Open Access to Open Use. In Proceedings of the Fifteenth International Conference on Grey Literature: The Grey Audit, A Field Assessment in Grey Literature, 2 3 December 2013 / compiled by D. Farace and J. Frantzen; GreyNet International, Grey Literature Network Service. Amsterdam: TextRelease, March 2014, http://socionet.ru/pub.xml?h=RePEc:rus:mqijxk:33

[7] CERIF (Common European Research Information Format), http://www.eurocris.org/Index.php?page=CERIFreleases&t=1


## International Professional Association of Research Information System Specialists euroCRIS and its Main Product CERIF

Sergey Parinov

From 2000 euroCRIS as an international professional association of research information system specialists collects, systemizes and distributes the best practice and approaches developed in different countries. euroCRIS by its task groups develops, maintains and promotes a common research information format (CERIF) and unified approach to build research information systems (CRIS). It works to improve interoperability among research information systems worldwide.

# Social Networks Meet Social Science

© Sergey Chernov
Center for the Study of New Media and Society, New Economic School,
Moscow
schernov@nes.ru

## Abstract

Computer science research community devoted a special attention to the studies of online social networks. In the meantime, we observed that algorithms and data processing techniques are often not enough to fully exploit research potential of social data available. One needs a set of models and methods explaining social interaction to pose meaningful questions on the wealth of data. This is a moment when social sciences come into a picture providing the necessary tools and knowledge. Here we review some research papers on online social networks, which have been published in the fields of economics, sociology, psychology and political science. We hope, such an interdisciplinary view on the social network research might help to eliminate existing gap between information management experts and scholars from social sciences.

## 1 Introduction

Recent boom of online social networks provided computer scientists with a lot of new research directions and inspired numerous efficient algorithms for scalable network analysis. In addition, it motivated scholars from social sciences like economics, sociology, psychology and political science, to catch up with modern data processing techniques. The massive datasets of digitized social data attracted social scientists, who rarely worked with terabytes of information or used scalable data analysis algorithms. Their usual toolbox is tailored to hundreds and thousands of data points, which are rather trustworthy and verifiable. In contrast, contemporary social graphs are built from millions of nodes and billions of edges, the data is highly dynamic, incomplete, unreliable and difficult to interpret.

It is reasonable to assume that online analytical tools from computer science will be widely adopted by social scientists in a short term prospective and current technological gap will be eliminated. On the other hand, the social disciplines will stay superior in posing right questions on social data, which is a key skill in understanding characteristics and mechanisms of a modern society. Such division of competencies creates a need for the interdisciplinary research teams and a number of highly visible research papers were produced by consortia of social and data processing experts.

To provide an overview of problems and challenges addressed by the social sciences using online networking data, we selected a set of publications from four aforementioned branches of social sciences. Each study is summarized into a key research question, a method used and a general outcome. We grouped these works by the field of study, while some papers might be attributed to two or more disciplines. We also made an attempt to mention some national papers where applicable. Finally, some relevant studies completed with support from the Center for the Study of New Media and Society were added to the picture.

## 2 Psychology

An interesting compilation of papers on online addiction is prepared by Kuss and Griffiths [19]. One of the most comprehensive reviews of psychologically oriented research using Facebook was done by Wilson, Gosling, and Graham in 2012 [32]. A large volume of more than 400 articles has been split into five categories including: into 5 categories: descriptive analysis of users, motivations for using Facebook, identity presentation, the role of Facebook in social interactions, and privacy and information disclosure. Descriptive analysis focuses on Facebook users and their typical online behaviour. Motivations part is about why people use Facebook and identity presentation is about how they want to look for the outside world. Social interaction category studies relationships between individuals and groups. Finally, privacy and information disclosure papers target various trade-offs between personal information sharing and associated risks.

Following the proposed categories, we pick few representative papers from each. For example, in-house Facebook research lab published in 2011 one of the largest studies describing inhabitants of the biggest social network [2, 3, 27]. For the motivation study we take an example addressing psychological well-being and increased social capital of the users [23]. A case study of identity representation from [29] confirms old wisdom that attractive photos are more likely to help in getting new friends than unattractive ones. One recent work which caught quite some attention in press is devoted to social interactions, in particular, Backstrom and Kleinberg [4] studied possibility of identifying romantic relationship between two people based on

**Proceedings of the 16th All-Russian Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" — RCDL-2014, Dubna, Russia, October 13–16, 2014.**

their network neighborhood alone. The results are promising and might lead to finding structurally significant people in various online applications.

## 3 Sociology

The methods of sociological research could be considered as the most demanded from the computer science community. Several basic notions in social network analysis have deep roots in sociology, starting from the famous works by Granovetter granovetter1973 or by Travers and Milgram [25]. A good introductions into a sociological prospective on modern social network analysis include work by Marin and Wellman [20] or by Hansen, Shneiderman and Smith [16], while a standard textbook in this area belongs to Faust andWasserman [30]. Among research topics addressed by modern sociologists we observe a popular thema of geo-dependencies of social ties, which has already produced several crossdisciplinary teams [22, 24, 28]. Another hot topic is a cascade-behavior prediction [7, 13], which promises numerous potential benefits for the information diffusion applications. A more specific task in the same direction is measuring the influence on information sharing of each particular node [5].

## 4 Economics

Thanks to the excellent textbooks by Matthew Jackson [18] and by Easley and Kleinberg [9], we currently have a good overview of the role of social networks in economics. For a short intro in this area one is advised to look at a brief summary [17]. However, economics research is often focused on small-scale networks from offline world, rather than on massive online datasets. Situation is changing nowadays and more and more papers address online networks in the economics science, for an up-to-date review one might look into a work by Ravasan, Rouhani and Asgary [21]. A lot of available studies concentrate on applied questions, for example, work by Acquisti and Fong [1] shows how online information could be used for discrimination during in a hiring process. Another interesting topic is connected to online fundrising potential [6]. Some early studies in political economy show dependencies between online activity and offline movements [10].

## 5 Political Science

Modern political science is a lot more about proper math and heavy computing as it was ever before. Researcher perform complex data analysis on millions of tweets and online texts to better understand current political issues and verify classic models and theories of the field. A recent outburst of civil movements supported by visible online activity was recently studied in Egypt [26], Spain [14], USA [8] and Russia [15]. Other interesting topics include debates around political economy of privacy on Facebook [11] and Internet surveillance [12]. Regarding applications, some interesting vizualizations around political activites are developed in [31].

## 6 Conclusions

This survey is a minor step towards better cooperation between computer science and social sciences communities. A set of research problems listed above is not exhaustive by any means, but rather a basic selection for future development. We hope, interested readers will follow the way of diving into alien, but interesting research disciplines, and will find necessary bits of knowledge, which are invisible within their native research circle.

## References


[1] Acquisti, A., and Fong, C.M. An Experiment in Hiring Discrimination Via Online Social Networks. Social Science Research Network Working Paper Series (Apr. 2012).

[2] Backstrom, L. Anatomy of Facebook. https://www.facebook.com/notes/facebookdata-team/anatomy-offacebook/10150388519243859, 2011. [Online; accessed 25-May-2014].

[3] Backstrom, L., Boldi, P., Rosa, M., Ugander, J., and Vigna, S. Four degrees of separation. CoRR abs/1111.4570 (2011).

[4] Backstrom, L., and Kleinberg, J.M. Romantic partnerships and the dispersion of social ties: a network analysis of relationship status on facebook. In CSCW (2014), S. R. Fussell, W. G. Lutters, M. R. Morris, and M. Reddy, Eds., ACM, pp. 831–841.

[5] Bakshy, E., Hofman, J.M., Mason, W.A., and Watts, D.J. Everyone's an influencer: Quantifying influence on twitter. In Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (New York, NY, USA, 2011), WSDM'11, ACM, pp. 65–74.

[6] Castillo, M., Petrie, R., and Wardell, C. Fundraising through online social networks: A field experiment on peer-to-peer solicitation. Journal of Public Economics (2014).

[7] Cheng, J., Adamic, L.A., Dow, P.A., Kleinberg, J. M., and Leskovec, J. Can cascades be predicted? In WWW (2014), C.-W. Chung, A.Z. Broder, K. Shim, and T. Suel, Eds., ACM, pp. 925–936.

[8] Conover, M.D., Davis, C., Ferrara, E., McKelvey, K., Menczer, F., and Flammini, A. The geospatial characteristics of a social movement communication network. PloS one 8, 3 (2013), e55957.

[9] David, E., and Jon, K. Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press, New York, NY, USA, 2010.

[10] Enikolopov, R., Makarin, A., Petrova, M., and Polishchuk, L. Social media and protest participation: Cross-city evidence from Russia.

[11] Fuchs, C. The political economy of privacy on facebook. Television & New Media 13, 2 (2012), 139–159.

[12] Fuchs, C., Boersma, K., Albrechtslund, A., and Sandoval, M. Internet and Surveillance: The Challenges of Web 2.0 and Social Media, 1st ed. Routledge, New York, NY, 10001, 2011.

[13] Goel, S., Watts, D.J., and Goldstein, D.G. The structure of online diffusion networks. In Proceedings of the 13th ACM Conference on Electronic Commerce (New York, NY, USA, 2012), EC'12, ACM, pp. 623–638.

[14] González-Bailón, S., Borge-Holthoefer, J., Rivero, A., and Moreno, Y. The dynamics of protest recruitment through an online network. Scientific reports 1 (2011).

[15] Greene, S.A. Beyond bolotnaia. Problems of Post-Communism 60, 2 (2013), 40–52.

[16] Hansen, D.L., Shneiderman, B., and Smith, M.A. Social Network Analysis: Measuring, Mapping, and Modeling Collections of Connections. In Analyzing Social Media Networks with NodeXL, D.L. Hansen, B. Shneiderman, and M. A. Smith, Eds. Elsevier, 2011, ch. 3, pp. 31–50.

[17] Jackson, M.O. The economics of social networks. In Proceedings of the 9th world congress of the econometric society (2005).

[18] Jackson, M.O. Social and Economic Networks. Princeton University Press, Princeton, NJ, USA, 2008.

[19] Kuss, D.J., and Griffiths, M.D. Online social networking and addiction – a review of the psychological literature. International journal of environmental research and public health 8, 9 (2011), 3528–3552.

[20] Marin, A., and Wellman, B. Social network analysis: An introduction. Scott J, Carrington PJ, editors. The SAGE Handbook of Social Network Analysis. Thousand Oaks, CA: SAGE Publications (2011), 11–25.

[21] Ravasan, A.Z., Rouhani, S., and Asgary, S. A review for the online social networks literature (2005–2011). European Journal of Business and Management 6, 4 (2014), 22–37.

[22] Sadilek, A., Kautz, H., and Bigham, J. P. Finding your friends and following them to where you are. In Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (New York, NY, USA, 2012), WSDM'12, ACM, pp. 723–732.

[23] Steinfield, C., Ellison, N.B., and Lampe, C. Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. Journal of Applied Developmental Psychology 29, 6 (2008), 434–445. Social Networking on the Internet – Developmental Implications.

[24] Takhteyev, Y., Gruzd, A., and Wellman, B. Geography of Twitter networks. Social Networks 34, 1 (Jan. 2012), 73–81.

[25] Travers, J., and Milgram, S. An experimental study of the small world problem. Sociometry 32, 4 (Dec. 1969), 425–443.

[26] Tufekci, Z., and Wilson, C. Social media and the decision to participate in political protest: Observations from tahrir square. Journal of Communication 62, 2 (2012), 363–379.

[27] Ugander, J., Karrer, B., Backstrom, L., and Marlow, C. The anatomy of the facebook social graph. CoRR abs/1111.4503 (2011).

[28] Volkovich, Y., Scellato, S., Laniado, D., Mascolo, C., and Kaltenbrunner, A. The length of bridge ties: Structural and geographic properties of online social interactions. In ICWSM (2012), J.G. Breslin, N.B. Ellison, J.G. Shanahan, and Z. Tufekci, Eds., The AAAI Press.

[29] Wang, S.S., Moon, S.-I., Kwon, K.H., Evans, C.A., and Stefanone, M.A. Face off: Implications of visual cues on initiating friendship on Facebook. Computers in Human Behavior 26, 2 (Mar. 2010), 226–234.

[30] Wasserman, S., and Faust, K. Social network analysis: Methods and applications, vol. 8. Cambridge university press, 1994.

[31] Weber, I., Garimella, V.R.K., and Batayneh, A. Secular vs. islamist polarization in Egypt on twitter. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (New York, NY, USA, 2013), ASONAM'13, ACM, pp. 290–297.

[32] WILSON, R. E., GOSLING, S. D., AND GRAHAM, L. T. A review of Facebook research in the social sciences. Perspectives on Psychological Science 7, 3 (May 2012), 203–220.

## Социальные сети в социальных науках

Сергей Чернов

Стремительный рост социальных сетей придал дополнительный импульс исследованиям для ученых-информатиков в этой области. В то же время, мы наблюдаем, что владение алгоритмами и методами обработки информации не гарантирует максимального раскрытия научного потенциала данных из социальных сетей. Помимо этого, требуется инструментарий моделей и методов для понимания социальных взаимодействий, чтобы иметь возможность формулировать интересные вопросы к многообразию современных данных. В этот момент социальные науки могут помочь процессу исследования, предоставив необходимые инструменты и знания. В данном тьюториале мы представляем избранные исследования онлайновых социальных сетей, опубликованные в научных сообществах экономистов, социологов, психологов и политологов. Надеемся, что подобный междисциплинарный взгляд на исследования социальных сетей поможет сократить существующее разделение между экспертами в обработке информации и учеными из общественных наук.

# Моделирование грид и облачных сервисов как средство повышения эффективности их разработки

© В.В Кореньков © А.В. Нечаевский © Г.А. Ососков
© Д.И. Пряхина © В.В. Трофимов © А.В. Ужинский
Объединенный институт ядерных исследований,
Дубна
korenkov@jinr.ru nechav@jinr.ru ososkov@jinr.ru

**Аннотация**

После введения, показывающего актуальность сетей распределенных вычислений, основанной на грид-облачных структурах в эпоху Больших Данных, описана новая система моделирования грид и облачных сервисов, разрабатываемая в ЛИТ ОИЯИ, ориентированная на повышения эффективности разработки путем учета качества работы уже функционирующей системы при проектировании ее дальнейшего развития за счет объединения самой программы моделирования с системой мониторинга реального (или модельного) грид-облачного сервиса через специальную базу данных. Приведен пример применения программы для моделирования достаточно общей облачной структуры, которая может быть использована вне рамок физического эксперимента, например, как хранилище информации общего доступа типа электронной библиотеки.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 14-07-00215.

## 1 Введение

Впечатляющий прогресс компьютерных технологий, программных средств и взрывное развитие глобального информационного пространства, возникшего с появлением Интернета, объединившего между собой компьютерные сети во всемирную систему передачи информации с помощью информационно-вычислительных ресурсов, — все это ознаменовало вступление человечества в новую эру Больших Данных (Big Data).

Все окружающие нас процессы в технической, научной и, особенно, в социальной сферах постоянно обрушивают на нас непрерывный поток информации, идущий из наших компьютеров и мобильных телефонов, передач различных масс медиа, регулярно перебиваемых назойливой рекламой, различных сенсорных устройств, GPS навигаторов и множества других источников, среди которых, в частности, по оценке Д. Акста, автора книги о будущем электронных библиотек, ожидается, что годовое поступление всех мировых источников цифровой информации: оцифрованных книг, фильмов, оптических и магнитных записей потребует порядка полутора петабайт хранения [1].

Говоря о «больших данных», надо понимать, что это не просто «очень много данных» и, кроме объема, следует учитывать и другие их характеристики. Еще в 2001 г. Мета Групп [2] ввела в качестве определяющих характеристик для больших данных так называемые ***«три V»*:**

1. **объем** (*volume*, в смысле величины физического объёма),
2. **скорость** (*velocity* в смыслах как скорости прироста, так и необходимости высокоскоростной обработки и получения результатов),
3. **многообразие** (*variety*, в смысле возможности одновременной обработки различных типов структурированных и неструктурированных данных

Однако, когда общий поток данных растет экспоненциально, удваиваясь каждый год, за счет революционных технологических изменений, к 2014 г. даже эту «3V» модель предлагают расширить, добавляя новые и новые «V», включая Validity (обоснованность, применимость), Veracity (достоверность), Value (ценность, полезность), и Visibility (обозримость, способность к визуализации) и т.д. [3, 4].

Часто Большие Данные определяют проще, как такие, которые слишком велики и сложны, чтобы их можно было эффективно запомнить, передать и проанализировать стандартными средствами доступных баз данных и иных имеющихся систем хранения, передачи и обработки. В то же время эти Большие Данные, непрерывно поступающие из множества вышеперечисленных источников, должны быть доступны для поисковых систем, проанализированы в центрах бизнеса, производства, медицины, правоохранения, обороны, науки и просто индивидуумов, которые их могут затребовать.

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

Заметим, что влиянию результатов подобного анализа теперь подвержены все мы, в том числе и люди, никак не связанные с компьютером и интернетом. Дело в том, что информация о любом нашем обращении в полицию, финансовые налоговые, медицинские учреждения, использование банковских карт и бонусов сетевых магазинов оседает в соответствующих базах данных и может быть оттуда извлечена по заказу компетентных органов, а также различных поисковых сетевых служб или, наконец, недобросовестными хакерами. Много примеров подобного использования Больших Данных можно найти в популярном ролике А. Сербанта [5]. Обстоятельный обзор революционных изменений, которые вносят Большие Данные в современное общество дан в бестселлере В. Майер-Шенбергера и К. Кукьера [6].

Однако чаще в качестве одного из наиболее впечатляющих примеров больших данных приводят потоки экспериментальных данных физики высоких энергий, поступающие с Большого Адронного Коллайдера (БАК) в ЦЕРНе [7]. За время первого запуска БАК к 2012 г. четыре экспериментальные установки на нем ALICE, ATLAS, CMS и LHCb выдавали каждую секунду один петабайт ($10^{15}$ байт) данных. Запомнить такое количество данных невозможно ни на какой из современных вычислительных систем. Поэтому после сверхбыстрой сложной электронной предобработки, оставлявшей только одно полезное физическое событие из 10 тысяч, выполнялся их анализ в ЦЕРНовском компьютерном центре обработки из многих тысяч процессоров. После этого анализа оставался только 1% событий, возможно содержащих искомый физический феномен. Но даже после такого радикального сокращения потока экспериментальных данных в миллион раз для этих четырех больших экспериментов требовалось хранить поступающие в год 25 петабайт данных в специальных роботизированных ленточных хранилищах, т.к. копии этих данных подлежат передаче в сотни физических центров в 36 странах мира для более тщательного анализа на сотнях тысячах компьютеров, объединенных во Всемирную сеть распределенных вычислений — Worldwide LHC Computing Grid (WLCG). Ежедневно в WLCG обрабатываются полтора миллиона заданий, на что одному даже самому мощному современному компьютеру потребовалось бы 600 лет.

Сравнительная диаграмма на рисунке 1 по общим объемам перерабатываемых в 2012 г. данных в социальных сетях, поисковых системах, разных отраслях бизнеса, медицины, климатических прогнозов и БАК наглядно показывает, что исследования в ЦЕРНе идут в условиях Больших Данных.

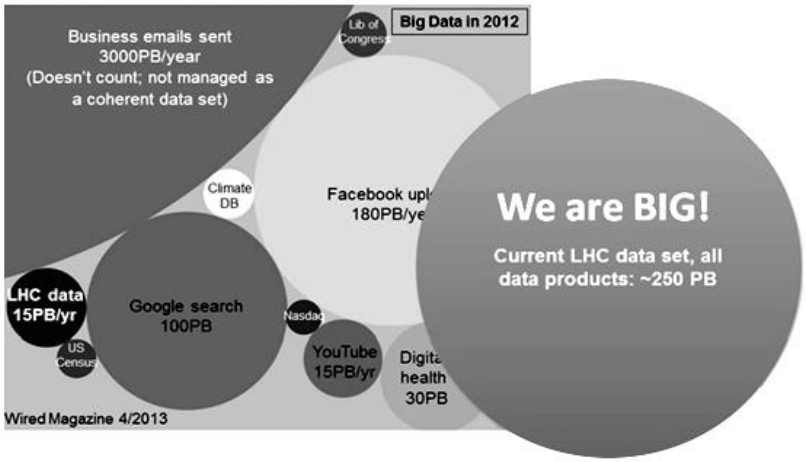



Рис. 1. Сравнительная диаграмма потоков данных в 2012 г. для разных приложений (бизнес, поисковые системы, БАК, медицина, климатические прогнозы) [8]

Известный специалист в области хранения информации Джим Грэй предсказал, что нарастающий поток научной информации должен неизбежно преобразовать практику науки, и назвал это изменение «четвертой парадигмой», в дополнение к трем предыдущим научным парадигмам – экспериментальной, теоретической и вычислительной [9].

На торжестве 4 июля 2012 г. по поводу получения ЦЕРНом нобелевской премии за открытие бозона Хиггса директор ЦЕРНа Рольф Хойер прямо назвал грид-технологии одним из трех столпов успеха (наряду с ускорителем LHC и физическими установками). Этот успех также подтверждает, что ЦЕРН, войдя в эру Больших Данных, эффективно преодолевает проблемы четвертой парадигмы, что является одним из примеров (наряду с созданием в ЦЕРНе WWW-всемирной паутины), когда разработки в области физики частиц начинают влиять на исследования в других научных областях.

Следует отметить, что в 2015 г. предстоит вторичный запуск БАК после его существенной модернизации, когда поток данных возрастет в 2,5 раза при удвоении времени на их обработку. В своих планах на такое развитие компьютинга в ЦЕРНе после 2015 г., которое смогло бы обеспечить потенциально новую физику, помимо значительного

увеличения вычислительных мощностей и ресурсов хранения данных, совершенствования программных средств анализа и моделирования, активного использования распределенных параллельных вычислений, предлагается повысить эффективность распределенных систем вычислений на базе WLCG путем синтеза грид и облачных технологий [8].

Примером уже имеющейся технологии, реализующей подобный синтез для работы с Большими Данными является система PanDA (Production and Distributed Analysis – обработка данных и распределенный анализ) эксперимента ATLAS на LHC. Сегодня PanDA рассматривается как возможная система для российского мегапроекта НИКА в Объединенном институте ядерных исследований (г. Дубна) [10].

Как было показано в недавно защищенной в ОИЯИ диссертации Н.А. Кутовского, включение в грид с его жесткой структурой, ориентированной на интеграцию уже существующих процессорных и программных ресурсов, облачных структур, более гибких за счет использования виртуальных кластеров из виртуальных вычислителей, позволяет сократить время решения широкого круга задач в области физики высоких энергий и повысить эффективность использования ресурсов [11].

Следует подчеркнуть, что в силу своей сложности и высокой стоимости разработка таких сложнейших грид-облачных систем сбора, передачи и распределенной обработки сверхбольших объемов информации требует больших предварительных исследований по выбору оптимальной их структуры с учетом стоимости и предполагаемых ресурсов и загрузки. Подобные исследования должны основываться на тщательном моделировании как потока заданий с учетом их типов и статистических данных о распределении времени их поступления и требуемых компьютерных ресурсов для их выполнения, так и состава моделируемой грид-структуры.

Настоящая работа посвящена разрабатываемой в ЛИТ ОИЯИ новой системе моделирования грид и облачных сервисов, ориентированной на повышения эффективности их разработки путем учета качества работы уже функционирующей системы в прогнозах на ее дальнейшее развитие. Это предлагается сделать за счет объединения самой программы моделирования с системой мониторинга реального (или модельного) грид-облачного сервиса через специальную базу данных, осуществляющую сбор и статистический анализ по вычислению распределений данных мониторинга, используемых затем для динамической коррекции параметров моделирования. Приведен пример применения программы для моделирования достаточно общей облачной структуры, которая может быть использована вне рамок физического эксперимента, например, как хранилище информации общего доступа типа электронной библиотеки.

## 2 Принципы моделирования грид и облачных инфраструктур

В силу очевидной необходимости предварительного моделирования грид-систем были написаны несколько программ для их моделирования, обзор которых можно найти в работе [12], где после сравнительного анализа этих программ был сделан выбор в пользу программы грид-моделирования GridSim [13]. GridSim – это библиотека классов, построенная на стандартной библиотеке SimJava, с помощью которой можно моделировать поток дискретных событий во времени.

В предыдущей работе авторов [14] описана программа моделирования, основанная на расширении классов GridSim и объединении их в программу, которая моделирует обработку потока заданий проектируемой грид-структурой, и алгоритмов планирования потока заданий ALEA [15]. Для запуска программы требуется задать состав и топологию центров обработки моделируемой грид-структуры, а также распределение ресурсов между заданиями. После этого программа выполняет имитационное моделирования процессов прохождения сгенерированного набора заданий через эту грид-структуру. В качестве результатов вычисляются временные оценки искомых параметров потока заданий.

Как уже было отмечено выше, постоянное развитие современных грид-систем требует непрерывных корректировок большинства параметров моделирования. Это необходимо для прогнозирования поведения системы при значительных ее изменениях с учетом статистики эксплуатации системы, получаемой на основе имеющихся программных средств ее мониторинга.

## 3 Мониторинг грид-систем

Система мониторинга – это набор программных и аппаратных средств для анализа и контроля состояния некоторой системы распределенных вычислений (см. например [16]). Система мониторинга и учета ресурсов (СМУР) предназначена для отслеживания текущего состояния ресурсов, заданий и других объектов в грид-системе. Инструментарий СМУР должен предоставлять как статическую, так и динамическую информацию о функционировании грид-системы (примером динамической информации может служить состояние очередей на вычислительном кластере), а также результаты статистического анализа этой информации. Среди основных задач мониторинга отметим следующие:

- Непрерывное наблюдение за состоянием грид-сервисов, как базовых (общих для всей инфраструктуры), так и относящихся к отдельным ресурсным центрам;
- Получение информации о вычислительных ресурсах (количество вычислительных узлов для

выполнения задач, архитектура вычислительной системы, установленное программное обеспечение, доступные специализированные программные пакеты) и о потребленном процессорном времени;

- Данные о доступе виртуальных организаций к ресурсам и использовании ими квот на вычислительные ресурсы;
- Мониторинг выполнения вычислительных заданий и задач (запуск, изменение состояния, коды завершения и т.п.).

Среди параметров мониторинга, необходимых для последующего моделирования, наиболее существенными являлись следующие:

1) число задач (симуляция, анализ, реконструкция) поступающих в систему;

2) объем используемой оперативной памяти;

3) использованное процессорное время;

4) число обработанных событий;

5) время расчета задачи;

6) объем используемых данных.

Примеры диаграмм, показывающих динамику изменения параметров мониторинга показаны на рисунках 2 и 3.

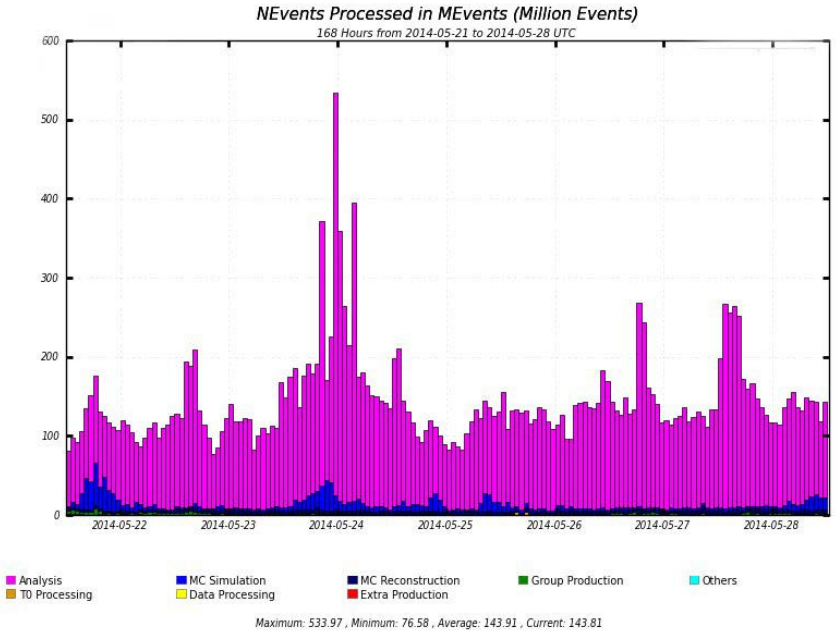



Рис. 2. Динамика распределения числа физических событий (в млн) по группам решаемых задач (анализ, симуляция, обработка)

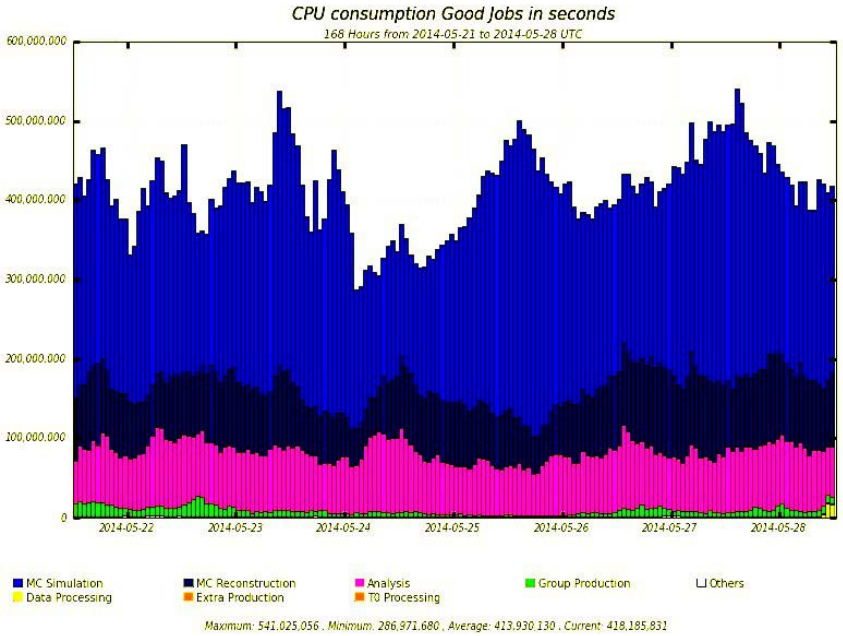



Рис. 3. Динамика распределения загрузки процессоров (число выполняемых заданий в секунду) по разным типам решаемых задач

## 4 Схема программы моделирования SyMSim

В настоящей работе на основе данных мониторинга одной из грид-систем WLCG [7], сохраняемых в специально разработанной базе данных, через веб-интерфейс выполняется их статистический анализ, результаты которого позволяют затем генерировать адекватный поток заданий для изменения параметров моделирования.

Схема программы SyMSim (Synthesis of Monitorung and Simulation – Синтез Мониторинга и Моделирования), реализующей идею синтеза процессов мониторинга и моделирования, представлена на рисунке 4. Данные мониторинга реальной грид-системы поступают в базу данных следующим образом: задание отправляется на сервер (1), далее сайт-пилот запрашивает задание на обработку (2), сервер отправляет задание на исполнение (3), информация о выполнении задания поступает в базу данных StatDB (4). На основе данных мониторинга пользователь задает входные параметры модели (5) и потока заданий (6), далее модель обрабатывает задания (7,8). Результаты работы модели доступны пользователю для дальнейшего анализа.

При разработке программы SyMSim удалось расширить ее сферу применения на ставшие весьма актуальными в последнее время системы облачных вычислений.

Для иллюстрации возможностей разработанной программы SyMSim по имитационному моделированию облачных вычислений ниже приведен пример ее применения для оптимизации простой облачной структуры. Эта структура мыслилась, как предназначенная для обработки данных физического эксперимента, но в принципе в такую схему укладываются и другие структуры связанные с хранением и обновлением больших массивов цифровой информации, в том числе и возможный вариант реализации электронной библиотеки.

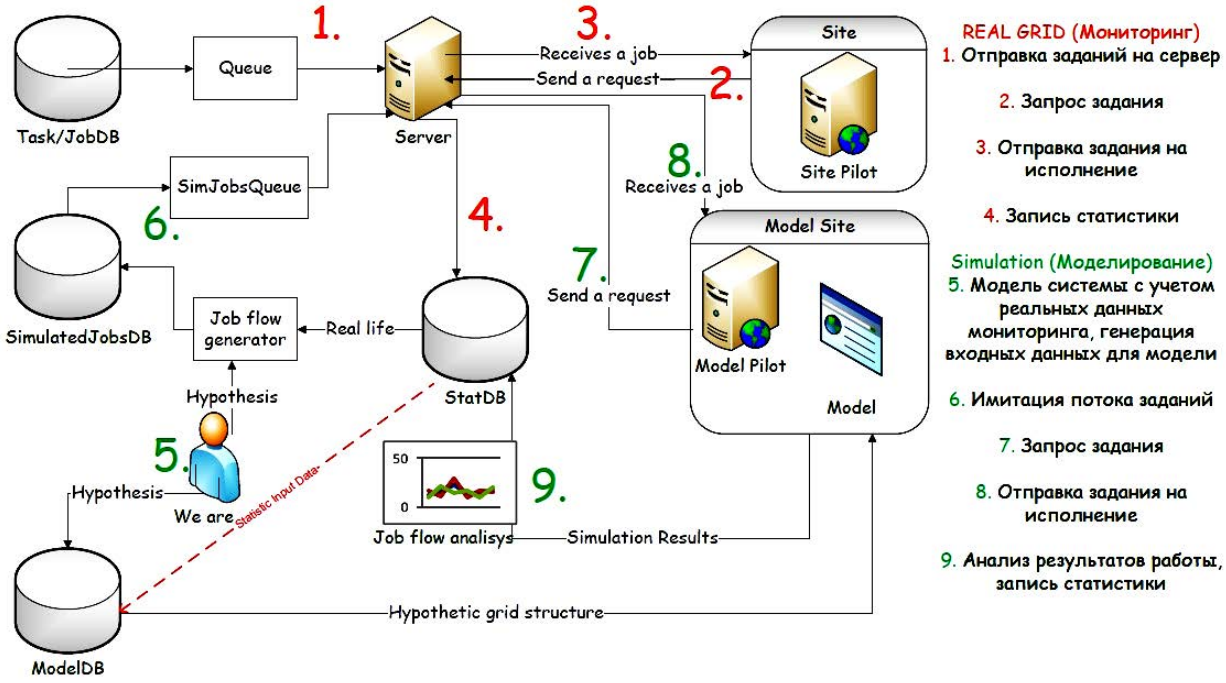



Рис. 4. Схема программы SyMSim моделирования системы распределенных вычислений с учетом данных мониторинга этой системы

## 5 Постановка задачи и результаты моделирования

Рассматривается модель реализации облачной структуры, предназначенной для хранении данных в роботизированной библиотеке с тысячами кассет с магнитными лентами, из загрузчиков-драйвов которых робот автоматически достаёт требуемые ленты и устанавливает в одно или несколько устройств чтения-записи. Достижения современных технологий позволяют записать до 6,25 терабайт на один ленточный картридж, причем современные стандарты записи на магнитные ленты под углом поддерживают деление ее на разделы подобно дискам, что позволяет обращаться к ним, как к традиционному каталогу. Фотография такого хранилища с роботом-загрузчиком представлена на рисунке 5.

Проектируемая структура состоит из ленточного робота, массива ленточных картриджей и кластера процессоров. Стоимость драйва – 5 условных единиц, процессора – 3 единицы.

Рис. 5. Фото хранилища с роботом-загрузчиком

В качестве критерия оценки выбирается время прохождения тестового потока из 99 заданий. Надо найти оптимальное соотношение количества процессоров и количества драйвов при ограниченном бюджете в 100 условных единиц.

**Результаты моделирования**

1. *Определение степени загрузки кластера*

Загрузка кластера $W = T_{100}/Ta$, где $T_{100}$ – процессорное время выполнения пакета $Ta$ – астрономическое время.

На рисунке 6 показана степень загрузки кластера в зависимости от количества процессоров и драйвов. Мы видим, что при большом количестве процессоров загрузка кластера падает, поскольку процессоры простаивают в ожидании монтирования кассет с данными на драйвы. Следовательно, надо выбирать оптимальное соотношение.

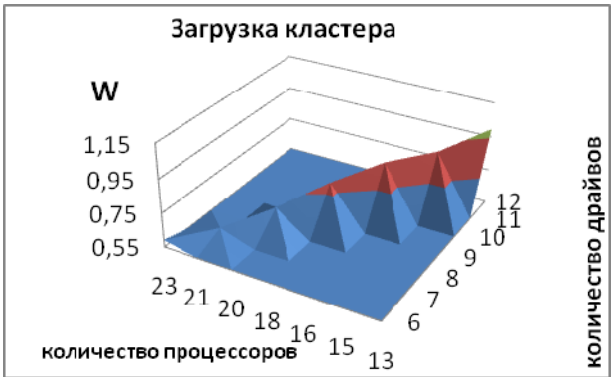



Рис. 6. Степень загрузки кластера в зависимости от количества процессоров и драйвов

2. *Время выполнения пакета заданий в зависимости от количества процессоров и драйвов*

На рисунке 7 стрелкой показан оптимум по числу вычислительных процессоров в кластере и дорогих драйвов. Таким образом, конфигурация, обеспечивающая минимальное время исполнения должна состоять из 18 вычислительных процессоров и 9 драйвов-загрузчиков.

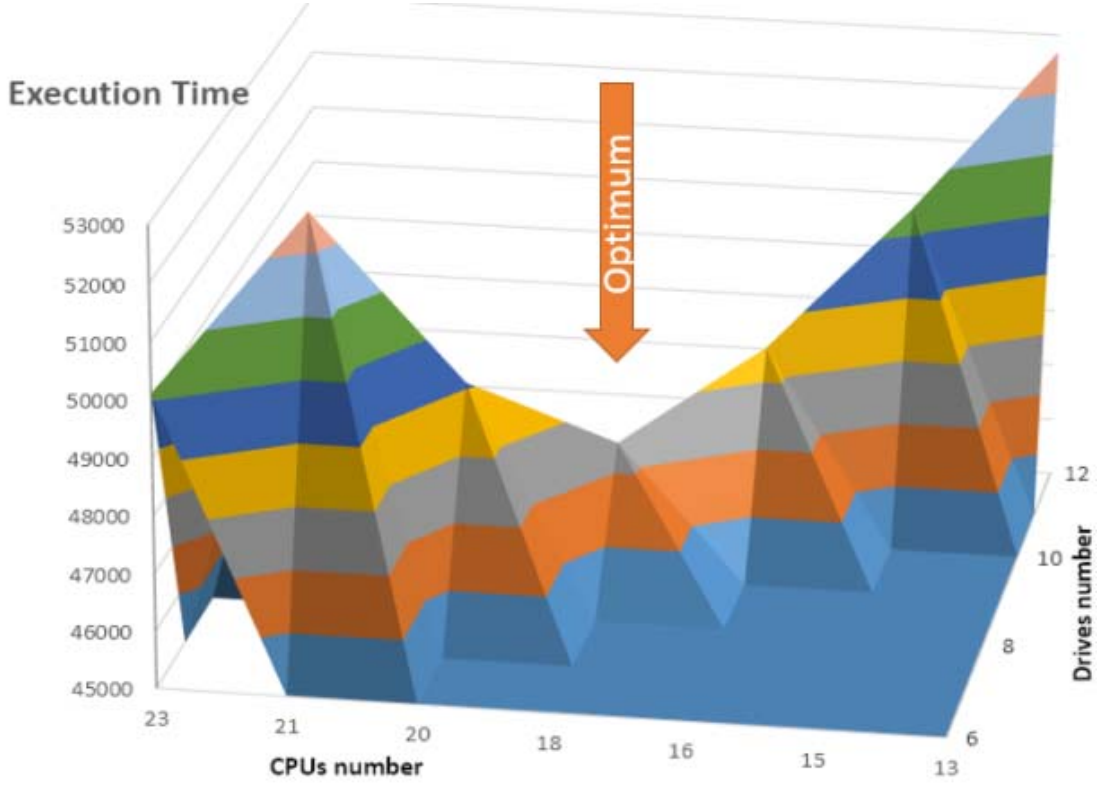



Рис. 7. Время выполнения пакета заданий в зависимости от количества процессоров и драйвов

## Заключение

Предложенный подход к моделированию и анализу вычислительных грид-облачных структур в экспериментальной физике высоких энергий основан на учете данных их мониторинга, используемых затем для динамической коррекции параметров моделирования. В силу общности своей реализации разработанная программа моделирования SyMSim может быть также применена для решения более широкого класса задач проектирования виртуальных центров обработки и хранения больших массивов данных. В частности программу можно применять для проектирования и последующего развития хранилищ информации общего доступа, не ограниченных областью физического эксперимента.

## Литература


[1] Akst, D. (2003). The Digital Library: Its Future Has Arrived. Carnegie Reporter, 2(3), 4–8.

[2] Doug Laney Meta Group: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[3] Rob Livingstone Advisory: http://rob-livingstone.com/2013/06/big-data-or-black-hole/

[4] Timo Elliott: 7 Definitions of Big Data You Should Know About http://timoelliott.com/blog/2013/07/7-definitions-of-big-data-you-should-know-about.html

[5] А. Себрант. Что такое Big data: http://www.slideshare.net/yandex/big-data-30799013

[6] В. Майер-Шенбергер, К. Кукьер. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим. Издательство: Манн, Иванов и Фербер, 2014.

[7] Компьютинг в ЦЕРНе: http://home.web.cern.ch/about/computing

[8] José M. Hernández (CIEMAT, Madrid) On behalf of the Spanish LHC Computing community, Perspectives on LHC Computing, Jornadas CPAN 2013, Santiago de Compostela.

[9] Jim Gray et all, Scientific Data Management in the Coming Decade, SIGMOD Record, Vol. 34, No. 4, Dec. 2005.

[10] А. Климентов, А. Ваняшин, В. Кореньков. За большими данными следит ПАНДА // Суперкомпьютеры. – 2013. – №15. – С. 56.

[11] Кутовский Н.А. Развитие методов построения грид-сред и систем облачных вычислений для задач физики высоких энергий: дис. … канд. физ-мат. наук. – Дубна: ОИЯИ, 2014.

[12] А.В. Нечаевский, В.В. Кореньков. Пакеты моделирования DataGrid // Системный анализ в науке и образовании: электронный журн. – 2009. – № 1.

[13] GridSim web-portal: http://www.gridbus.org/gridsim/

[14] В.В Кореньков, А.В. Нечаевский, В.В. Трофимов. Разработка имитационной модели сбора и обработки данных экспериментов на ускорительном комплексе НИКА // Информационные технологии и вычислительные системы. – 2013. – № 4. – С. 37–44.

[15] D. Klusacek, L. Matyska, and H. Rudova. Alea – Grid scheduling simulation environment // 7th International Conference on Parallel Processing and Applied Mathematics (PPAM 2007). – Springer, 2008. – Vol. 4967 of LNCS. – P. 1029–1038.

[16] Грид в ОИЯИ, Мониторинг и аккаунтинг. Веб-портал: http://grid.jinr.ru/?page_id=77


## Simulation of Grid and Cloud Services as the Means of the Efficiency Improvement of Their Development


Vladimir V. Korenkov, Andrey V. Nechaevskiy, Gennadiy A. Ososkov, Dariya I. Pryahina, Vladimir V. Trofimov, Aleksandr V. Uzhinskiy



After the introduction, showing the relevance of the distributed computing networks based on grid/cloud structures at the Big Data era, a new grid and cloud services simulation system are described. This system is developed at the LIT JINR Dubna and focused on improving the efficiency of the grid/cloud systems development by using work quality indicators of some real system to design and predict its evolution. For this purpose the simulation program is combined with real monitoring system of the grid/cloud service through a special database. An example of the program usage to simulate a sufficiently general cloud structure, which can be used, e.g., as a repository for information sharing, such as the digital library, is given.

# Methods for Anomaly Detection: a Survey

© Leonid Kalinichenko © Ivan Shanin © Ilia Taraban
Institute of Informatics Problems of RAS
Moscow
leonidandk@gmail.com ivan_shanin@mail.ru tarabanil@gmail.com

## Abstract

In this article we review different approaches to the anomaly detection problems, their applications and specific features. We classify different methods according to the data specificity and discuss their applicability in different cases.

## 1 Introduction

*Anomalies* (or outliers, deviant objects, exceptions, rare events, peculiar objects) is an important concept of the data analysis. Data object is considered to be an outlier if it has significant deviation from the regular pattern of the common data behaviour in a specific domain. Generally it means that this data object is “dissimilar” to the other observations in the dataset. It is very important to detect these objects during the data analysis to treat them differently from the other data. For instance, the anomaly detection methods are widely used for the following purposes:

- Credit card (and mobile phone) fraud detection [1, 2];
- Suspicious Web site detection [3];
- Whole-genome DNA matching [4, 5];
- ECG-signal filtering [6];
- Suspicious transaction detection [7];
- Analysis of digital sky surveys [8, 9].

The anomaly detection problem has become a recognized rapidly-developing topic of the data analysis. Many surveys and studies are devoted to this problem [1, 3, 4, 5, 10, 11]. The main purpose of this review is to reveal specific features of widely known statistical and machine learning methods that are used to detect anomalies. All considered methods will be categorized by the data form they are applied to.

The paper is organized as follows. In Section 2 we introduce three generic data representations that are most commonly used in anomaly detection problems: Metric Data, Evolving Data and Multistructured Data. In Sections 3, 4 and 5 these data forms are discussed in detail, each form is related to a certain class of problems and appropriate methods that are presented with the application examples. In Section 6 we discuss specific features of the anomaly detection problem that make strong impact on the methods used in this area. Section 7 contains conclusions and results of this review.

**Proceedings of the 16th All-Russian Conference “Digital Libraries: Advanced Methods and Technologies, Digital Collections” — RCDL-2014, Dubna, Russia, October 13–16, 2014.**

## 2 Data forms

The precise definition of the outlier depends on the specific problem and its data representation. In this survey we will establish a correspondence between concrete data representation forms and suitable anomaly detection methods. We assume that the data are usually presented in one of three forms: Metric Data, Evolving Data and Multistructured Data. Metric Data are the most common form of data representation, when every object in a dataset has a certain set of attributes that allows to operate with notions of "distance" and "proximity". Evolving Data are presented as well-studied objects: Discrete Sequences, Time Series and Multidimensional Data Streams. Third form is the Multistructured Data, under this term we understand the data that are presented in unstructured, semi-structured or structured form. This data form may not have a rigid structure, and yet it can contain various data dependencies. The most usual task with this type of data is to extract attributes that would allow using metric data oriented methods of the outlier analysis. In our survey the Multistructured Data are specialized as the Graph Data or Text Data.

## 3 Metric Data Oriented Methods

In this section the methods are considered that use the concept of “metric” data: such as the distance between objects, the correlation between them, and the distribution of data. We assume that the data in this case represents the objects in the space, so-called points. Then the task is to determine regular and irregular points, depending on the specific metric distance between objects in the space, or the correlation, or the spatial distribution of the points. In this case, we consider a structured data type, i.e., objects, which do not depend on time (time series are discussed in Section 4). Metric data form is the most widely-used, usually due to the fact that almost all entities can be represented as a structured object, a set of attributes, and thus as a point in a particular space [12]. Thus, these methods are used in various applications, e.g., in medicine and astronomy. We subdivide methods based

on the notion of distance, based on the correlations, data distributions and finally related to the data with high dimension and categorical attributes. We now turn to a more detailed review of certain types of these methods.

### 3.1 Distance-Based Data

Basic set of methods that use the notion of distance includes *clustering* methods, *K nearest neighbors* and their derivatives. Clustering methods use the distance defined in space to separate the data into homogenous and dense groups (clusters). If we see that the point is not included in large clusters, it is classified as anomaly. So we can assume that small clusters can be clusters of anomalous objects, because anomalies may also have a similar structure, i.e., be clustered. *K*-nearest neighbors method [13] is based on the concept of proximity. We consider *k* nearest points on the basis of certain rules, that decide whether the object is abnormal or not. A simple example of such rule is the distance between objects, i.e., the farthest object from its neighbors the more likely is abnormal. There are various kinds of rules starting from the distance-based rules to the neighbor distribution-based. For example, *LOF* (Local outlier factor) [14] is based on the density of objects in a neighborhood. Examples of clustering methods of anomaly detection in astronomy can be found in [15, 16, 17]. Besides classic clustering methods, many machine learning techniques can be used: e. g. modified methods of neural networks – *SOM* (Self-organizing map) [18, 19].

As an example, consider [20]. Authors propose their own clustering algorithm that also classifies anomalies. The main task in this case is to find erroneous values and interesting events in sensor data. Using Intel Berkeley Research lab dataset (2.3 million readings from 54 sensors) and synthetic dataset their algorithm reached Detection rate = 100%, False alarm rate = 0.10% and 0.09% respectively. These experimental results show that their approach can detect dangerous events (such as forest fire, air pollution, etc.) as well as erroneous or noisy data.

### 3.2 Correlated Dimension Data

The idea of these methods is based on the concept of correlation between data attributes. This situation is often found in real data because different attributes can be generated by the same processes. Thus, this effect allows to use linear models and methods based on them. A simple example of these methods is the linear regression. Using the method of linear regression of the data we are trying to bring some plane, which describes our data, then as the anomalous objects we pick those that are far away from this plane. Also often *PCA* (Principal component analysis) [21] can be used aiming at the reducing of the dimensionality of the data. Due to this the *PCA* is sometimes used in preprocessing data as in [15]. But it can also be directly used to separate anomalies. In this case, the basic idea is that at new dimensions it is easier to distinguish normal objects from abnormal objects [22].

### 3.3 Probabilistically Distributed Data

In probabilistic methods, the main approach is to assume that the data satisfy some distribution law. Thus, anomalous objects can be defined as objects that do not satisfy such basic rule. A classic example of these methods is the EM [23, 24], an iterative algorithm based on the maximum likelihood method. Each iteration is an expectation and maximization. Expectation supposes the calculation of the likelihood function, and maximization step is finding the parameter that maximizes the likelihood function. As well there are methods based on statistics, data distribution. These include the tail analysis of distributions (e.g., normal) and using the Markov, Chebyshev, Chernoff inequality.

An example of finding anomalies in sensors of rotating machinery is considered in [27]. In this task rolling element bearing failures are determined as anomalies. In practice, such frequent errors are one of the foremost causes of failures in the rotating mechanical systems. Comparing with other SVM-based approaches, the authors apply a Gaussian distribution. After choosing threshold and calculating parameters of distribution the anomalies are found. For testing they use vibration data from the NSF I/UCR Center for Intelligent Maintenance Systems (IMS – www.imscenter.net) and reach 97% accuracy.

Another examples of application of these methods can be found in [25, 26].

### 3.4 Categorical Data

The appropriate anomaly detection methods operate with continuous data - thus, one approach is to translate the categorical into continuous attributes. As an example, categorical data can be represented as a set of binary attributes. Certainly this kind of transformation may increase the dimension of the data, but this problem can be solved with methods of dimensionality reduction. Different probabilistic approaches also can be used for processing categorical data. It is clear that these approaches are not the only ones that can work with the categorical data. For example, some methods may be partially modified for using categorical data types: distance and proximity can be extended for categorical data.

### 3.5 High-Dimensional Data

In various applications the problem of the large number of attributes often arises. This problem implies the extra attributes, the incorrectness of the concepts of the distance between the objects and the sophistication of methods. For example, correlated dimension methods will work much worse on a large number of attributes. The main way of solving these problems is the search of subspaces of attributes. Earlier we mentioned the *PCA*, which is most commonly used for this task. But when selecting a small number of attributes other problems will be encountered. By changing the number of attributes, we lose information. Because of the small samples of anomalies, or the emergence of new types of anomalies, previously "abnormal" attributes can be lost.

More subtle approach for this problem is the Sparse Cube Method [28]. This technique is based on analysis of the density distributions of projections from the data, then the grid discretization is performed (data is forming a sparse hypercube at this point) and the evolutionary algorithm is employed to find an appropriate lower-dimensional subspace.

Many applications are confronted with the problem of high dimension. [29] will be taken as an example. Here authors searched for images, characterized by low quality, low illumination intensity or some collisions. They compare the PCA-based approach and the proposed one which is based on the random projections. After projection LOF works with neighborhood that was taken from source space. Both approaches show good results, but the second is much faster at large dimensions than PCA and LOF.

# 4 Evolving Data

It is very common that data is given in a temporal (or just consecutive) representation. Usually it is caused by the origin of the data. The temporal feature can be discrete or continuous, so the data can be presented in sequences or in time series. Methods that we review in this section can be applied to various common problems in medicine, economy, earth science, etc. Also we review methods suitable for "on-line" outlier analysis in data streams.

## 4.1 Discrete Sequences Data

There are many problems that need outlier detection in discrete sequences (web logs analysis, DNA analysis, etc. [3, 4]). There are several ways to determine an outlier in the data presented as a discrete sequence. We can analyze values on specific positions or test the whole sequence to be deviant. Three models are used to measure deviation in these problems: distance-based, frequency-based and *Hidden Markov Model* [10]. In the survey [30] the methods are divided in three groups: sequence-based, contiguous subsequence-based and pattern-based. The first group includes *Kernel Based Techniques, Window Based Techniques, Markovian Techniques*, contiguous subsequence methods include *Window Scoring Techniques* and *Segmentation Based Techniques*. Pattern-based methods include Substring Matching, Subsequence Matching and Permutation Matching Techniques [30].

In the work [34] the classic host-based anomaly intrusion detection problem is solved. The study is devoted to Windows Native API systems (a specific WindowsNT API that is used mostly during system boot), while most of other works consider UNIX-based systems. Authors analyse system calls in order to detect the abnormal behaviour that indicates an attack or intrusion. In order to solve this problem authors use a slide window method to establish a database of "normal patterns". Then the SVM method is used for anomaly detection, and in addition to that several window-based features are used to construct a detection rule. The method was tested on the real data from Win2K and WinXP systems (including logs of the important system processes such as svchost, Lsass, Inetinfo) and showed good results. One of the practical examples is given also in [31].

## 4.2 Time Series Data

If the data strongly depends on time, then we are facing the need to predict the forthcoming data and analyze the current trends. The most common way to determine an outlier is a surprising change of trends. The methods considered are based on well-developed apparatus of time series analysis including *Kalman Filtering*, Autoregressive Modeling, detection of unusual shapes with the Haar transform and various statistic techniques. Historically, the first approach to finding this sort of outliers used an idea from the immunology [33].

# 5 Multistructured Data

Sometimes the data is presented in a more complex form than numerical "attribute / value" table. In this case it is important to understand what an outlier is by using of the appropriate method of analysis. We will review two cases that need specific analysis: textual data (e.g., poll answers) and data presented as graph (e.g., social network data).

## 5.1 Text Data

In connection with the development of communications, world wide web, and especially with the advent of social networks, an interest in the analysis of texts on the Internet greatly increased. Considering the text analytics and anomaly detection, several major tasks can be distinguished: searching for abnormal texts – such as spam detection and searching for non-standard text – novelty detection. When solving these problems, the main problem is to represent texts in metric data. Thus we may use the previously defined methods. A simple way is to use the standard metrics for texts, such as the tf-idf. Extaction of entites from texts also is widespread. Using natural language processing techniques such as *LSA* (Latent semantic analysis) [34] it is possible to group text, integrating it with the standard anomaly detection methods. Due to the large number of texts, often the learning may have supervised character.

In [36] a study is focused on spam detection. Using the tf-idf measure their algorithm is based on computing distances between messages. Then it constructs “normal” area using training set. Afterwards area’s threshold determines whether an email was a spam. LingSpam (2412 ham, 480 spam), SpamAssassin(4150 ham, 1896 spam) and TREC(7368 ham , 14937 spam) were selected as experimental data sets. The spam detector shows high accuracy and low false positive rate for each dataset.

## 5.2 Graph Data

In this section we review how methods of data analysis depend on the graph structure. The main

difference is that the graph can be large and complex or, in the contrary, can consist of many smaller and simpler graphs. The main problem here is to extract appropriate attributes from nodes, edges and subgraphs that allow to use methods considered in Section 3. In the first case we will review methods that extract numerical attributes from smaller graphs and treat them like data objects using algorithms from Section 3. In case of a large and complex graph we may be interested in node outliers, linkage outliers and subgraph outlier. Methods that analyze node outliers usually extract attributes from the given node and its neighborhood, but in case of a linkage outlier detection the concept of an outlier itself becomes very complex [10, 3]. We will consider that edge is an outlier if it connects nodes from different dense clusters of nodes. The most popular methods are based on the random graph theory, matrix factorization and spectral analysis techniques [10]. Another problem in this section is to detect subgraphs with a deviant behavior and to determine its structure and attribute extraction [37].

Concrete definition of the outlier node or edge can differ according to a specific problem. For example, in [38] several types of anomaly are considered: near-star, near-clique, heavy-vicinity and dominant edge. Anomalous subgraphs are often detected using the *Minimal Description Length principle* [39, 40, 41]. One of the most important application today is Social Network Data – many popular modern techniques are used in this area: *Bayesian Models* [42], *Markov Random Field, Ising Model* [43], *EM algorithm* [44] as well as *LOF* [45].

In [44] authors perform anomaly detection methods for social networks. Social network contains information about its members and their meetings. The problem statement is to find abnormal meeting and to measure its degree of abnormality. The problem specificity is that the number of meetings is very small compared to the number of members, that makes challenging to use common statistical methods. In order to solve the problem authors use the notion of hypergraph. The vertices of the hypergraph are considered as members of the social network and the edges are considered as meetings of the members (each edge of a hypergraph connects some set of vertices together). The anomalies are detected through density estimation of p-dimensional hypercube (the EM algorithm tunes a two-component mixture). The method is tested on a synthetic data and shows relatively low estimation error. It is also considered to be a scalable method, which makes it very valuable to use on large social networks.

## 6 Specific features of the anomaly detection methods comparing to the general machine learning and statistics methods

In this article we show the application for the anomaly detection of various data mining methods that can re-use of the general machine learning and statistical algorithms. The anomaly detection problem has its own specific features making possible to tune the appropriate general algorithms properly turning them into the more efficient ones.

Let us consider one of the basic concept of machine learning – the classification problem. The anomaly detection problem can be considered as a classification problem, in that case the data is assumed to have the class of anomalies. Most of the methods that solve classification problems assume that data classes have some sort of inner predictable structure. But the only prediction that can be made about anomalies is that these objects do not resemble non-outlier "normal" data. In this case, in order to solve the anomaly detection problem, the outlier class modeling can be senseless and unproductive. Instead of this, one should pay attention to the structure of the normal data, its laws of distribution.

The machine learning methods can be divided in three groups: supervised, semi-supervised and unsupervised methods. The first group is the most learned. It requires the labeled "training" dataset, and this is exactly the situation described above: the information about the outlier class is used to tune a model of it in order to predict it's structure, which has often very complex or random nature. The semi-supervised methods use information only about the "normal" class, so these methods have better specifications for anomaly detection problem as well as unsupervised methods, which do not use any information besides the structure and configuration of the unlabeled data.

Another important specific feature of the anomaly detection problem is that usually abnormal objects are significantly rare (compared to the non-outlier objects). This effect makes hard to construct a reliable training dataset for supervised methods. Also, if this effect is not presented in the data, most of known methods will suffer from high alarm rates [47, 48].

## 7 Conclusion

In this paper we introduced an approach to classify different anomaly detection problems according to the way the data are presented. We reviewed different applications of the outlier analysis in various cases. At the end we summarized specific features of the methods suitable for the outlier analysis problem. Our future plans include preparing of a university master level course focused on the anomaly detection as well as working on the anomaly detection in various fields (e.g. finding peculiar objects in massive digital sky astronomy surveys).

## References


[1] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A Survey. ACM Computing Surveys, 41(3), 1–58. Doi:10.1145/1541880.1541882

[2] Kou, Y., Lu, C., & Sinvongwattana, S. (2004). Survey of Fraud Detection Techniques Yo-Ping Huang, 749–754.

[3] Pan Y., Ding X. Anomaly based web phishing page detection // Computer Security Applications Conference, 2006. ACSAC'06. 22nd Annual. – IEEE, 2006. – С. 381–392.

[4] Tzeng, J.-Y., Byerley, W., Devlin, B., Roeder, K., & Wasserman, L. (2003). Outlier Detection and False Discovery Rates for Whole-Genome DNA Matching. Journal of the American Statistical Association, 98(461), 236–246. doi:10.1198/016214503388619256

[5] Wu, B. (2007). Cancer outlier differential gene expression detection. Biostatistics (Oxford, England), 8(3), 566–75. doi:10.1093/biostatistics/kxl029

[6] Lourenço A. et al. Outlier detection in non-intrusive ECG biometric system // Image Analysis and Recognition. – Springer Berlin Heidelberg, 2013. – С. 43–52.

[7] Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-Based Anomaly Detection. ACM Transactions on Knowledge Discovery from Data, 6(1), 1–39. doi:10.1145/2133360.2133363

[8] Djorgovski, S. G., Brunner, R. J., Mahabal, A. A., & Odewahn, S. C. (2001). Exploration of Large Digital Sky Surveys. Observatory, 1–18.

[9] Djorgovski, S. G., Mahabal, A. A., Brunner, R. J., Gal, R. R., Castro, S., Observatory, P., Carvalho, R. R. De, et al. (2001a). Searches for Rare and New Types of Objects, 225, 52–63.

[10] Aggarwal, C. C. (2013). Outlier Analysis (introduction). doi:10.1007/978-1-4614-6396-2

[11] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A Survey. ACM Computing Surveys, 41(3), 1–58. doi:10.1145/1541880.1541882

[12] Berti-équille, L. (2009). Data Quality Mining : New Research Directions. Current.

[13] Stevens, K. N., Cover, T. M., & Hart, P. E. (1967). Nearest Neighbor Pattern Classification. EEE Transactions on Information Theory 13, I, 21–27.

[14] Breunig, M. M., Kriegel, H., Ng, R. T., & Sander, J. (2000). LOF?: Identifying Density-Based Local Outliers, 1–12.

[15] Borne, K., &Vedachalam, A. (2010). EFFECTIVE OUTLIER DETECTION IN SCIENCE DATA STREAMS. ReCALL, 1–15.

[16] Borne, K. (n.d.). Surprise Detection in Multivariate Astronomical Data.

[17] Henrion, M., Hand, D. J., Gandy, A., &Mortlock, D. J. (2013). CASOS: a Subspace Method for Anomaly Detection in High Dimentional Astronomical Databases. Statistical Analysis and Data Mining, 6(1), 1–89.

[18] Networks, K. (n.d.). Data Mining Self – Organizing Maps, 1–20.

[19] Manikantan Ramadas, Shawn Ostermann, Brett TjadenDetecting Anomalous Network Traffic with Self-organizing Maps.(2003) Recent Advances in Intrusion Detection Lecture Notes in Computer Science. Vol. 2820, 36–54.

[20] Purarjomandlangrudi A., Ghapanchi A. H., Esmalifalak M. A Data Mining Approach for Fault Diagnosis: An Application of Anomaly Detection Algorithm // Measurement. – 2014.

[21] Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433–459. doi:10.1002/wics.101

[22] Dutta H. et al. Distributed Top-K Outlier Detection from Astronomy Catalogs using the DEMAC System // SDM. – 2007.

[23] Cansado, A., & Soto, A. (2008). Unsupervised Anomaly Detection in Large Databases Using Bayesian Networks. Network, 1–37.

[24] Zhu, X. (2007). CS838-1 Advanced NLP : The EM Algorithm K-means Clustering, (6), 1–6.

[25] Spence, C., Parra, L., & Sajda, P. (2001). Detection, Synthesis and Compression in Mammographic Image Analysis with a Hierarchical Image Probability Model, 3–10.

[26] Pelleg, D., & Moore, A. (n.d.). Active Learning for Anomaly and Rare-Category Detection.

[27] Fawzy A., Mokhtar H. M. O., Hegazy O. Outliers detection and classification in wireless sensor networks // Egyptian Informatics Journal. – 2013. – Т. 14, № 2. – С. 157–164.

[28] Aggarwal C. C., Philip S. Y. An effective and efficient algorithm for high-dimensional outlier detection // The VLDB journal. – 2005. – Т. 14, № 2. – С. 211–221.

[29] De Vries, T., Chawla, S., &Houle, M. E. (2010). Finding Local Anomalies in Very High Dimensional Space. 2010 IEEE InternationalConferenceonDataMining, 128–137. doi:10.1109/ICDM.2010.151

[30] Chandola V., Banerjee A., Kumar V. Anomaly detection for discrete sequences: A survey // Knowledge and Data Engineering, IEEE Transactions on. – 2012. – Т. 24, № 5. – С. 823–839.

[31] Budalakoti S., Srivastava A. N., Otey M. E. Anomaly detection and diagnosis algorithms for discrete symbol sequences with applications to airline safety // Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on. – 2009. – Т. 39, №. 1. – С. 101–113.

[32] Wang M., Zhang C., Yu J. Native API based windows anomaly intrusion detection method using SVM // Sensor Networks, Ubiquitous, and Trustworthy Computing, 2006. IEEE International Conference on. – IEEE, 2006. – Т. 1. – С. 6.

[33] Dasgupta D., Forrest S. Novelty detection in time series data using ideas from immunology // Proceedings of the international conference on intelligent systems. – 1996. – С. 82–87.

[34] Susan T. Dumais (2005). "Latent Semantic Analysis". Annual Review of Information Science and Technology 38: 188. doi:10.1002/aris.1440380105

[35] Allan, J., Papka, R., & Lavrenko, V. (1998). On-line New Event Detection and Tracking.

[36] Laorden C. et al. Study on the effectiveness of anomaly detection for spam filtering // Information Sciences. – 2014. – T. 277. – C. 421–444.

[37] Kil, H., Oh, S.-C., Elmacioglu, E., Nam, W., & Lee, D. (2009). Graph Theoretic Topological Analysis of Web Service Networks. WorldWideWeb, 12(3), 321–343. doi:10.1007/s11280-009-0064-6

[38] Akoglu L., McGlohon M., Faloutsos C. Oddball: Spotting anomalies in weighted graphs // Advances in Knowledge Discovery and Data Mining. – Springer Berlin Heidelberg, 2010. – C. 410–421.

[39] Noble C. C., Cook D. J. Graph-based anomaly detection // Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2003. – C. 631–636.

[40] Eberle W., Holder L. Discovering structural anomalies in graph-based data // Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on. – IEEE, 2007. – C. 393–398.

[41] Chakrabarti D. Autopart: Parameter-free graph partitioning and outlier detection // Knowledge Discovery in Databases: PKDD 2004. – Springer Berlin Heidelberg, 2004. – C. 112–124.

[42] Heard N. A. et al. Bayesian anomaly detection methods for social networks //The Annals of Applied Statistics. – 2010. – T. 4, № 2. – C. 645–662.

[43] Horn C., Willett R. Online anomaly detection with expert system feedback in social networks // Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. – IEEE, 2011. – C. 1936–1939.

[44] Silva J., Willett R. Detection of anomalous meetings in a social network //Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on. – IEEE, 2008. – C. 636–641.

[45] Bhuyan M., Bhattacharyya D., Kalita J. Network anomaly detection: methods, systems and tools. – 2013.

[46] Portnoy L., Eskin E., Stolfo S. Intrusion Detection with Unlabeled Data Using Clustering (2001) // ACM Workshop on Data Mining Applied to Security (DMSA 01).

[47] Laorden C. et al. Study on the effectiveness of anomaly detection for spam filtering // Information Sciences. – 2014. – T. 277. – C. 421–444.

[48] Fawzy A., Mokhtar H. M. O., Hegazy O. Outliers detection and classification in wireless sensor networks // Egyptian Informatics Journal. – 2013. – T. 14, № 2. – C. 157–164.

[49] Yu M. A nonparametric adaptive CUSUM method and its application in network anomaly detection // International Journal of Advancements in Computing Technology. – 2012. – T. 4, № 1. – C. 280–288.

[50] Muniyandi A.P., Rajeswari R., Rajaram R. Network anomaly detection by cascading k-Means clustering and C4. 5 decision tree algorithm // Procedia Engineering. – 2012. – T. 30. – C. 174–182.

[51] Muda Z. et al. A K-Means and Naive Bayes learning approach for better intrusion detection // Information technology journal. – 2011. – T. 10, №. 3. – C. 648–655.

[52] Kavuri V. C., Liu H. Hierarchical clustering method to improve transrectal ultrasound-guided diffuse optical tomography for prostate cancer imaging // Academic radiology. – 2014. – T. 21, № 2. – C. 250–262.

[53] Li S., Tung W. L., Ng W. K. A novelty detection machine and its application to bank failure prediction // Neurocomputing. – 2014. – T. 130. – C. 63–72.

[54] Cogranne R., Retraint F. Statistical detection of defects in radiographic images using an adaptive parametric model // Signal Processing. – 2014. – T. 96. – C. 173–189.

[55] Daneshpazouh A., Sami A. Entropy-Based Outlier Detection Using Semi-Supervised Approach with Few Positive Examples // Pattern Recognition Letters. – 2014.

[56] Rahmani A. et al. Graph-based approach for outlier detection in sequential data and its application on stock market and weather data // Knowledge-Based Systems. – 2014. – T. 61. – C. 89–97.

# Программирование методов разрешения сущностей и слияния данных при реализации ETL в среде Hadoop*

© Вовченко А.Е. © Калиниченко Л.А. © Ковалев Д.Ю.
Институт Проблем Информатики РАН (ИПИ РАН)
Москва
alexey.vovchenko@gmail.com leonidk@synth.ipi.ac.ru dm.kovalev@gmail.com

**Аннотация**

В статье обсуждаются вопросы разрешения сущностей (Entity Resolution) и слияния данных (Data Fusion) в контексте интеграции больших данных в среде Hadoop. Проблема разрешения сущностей ориентирована на решение таких задач как выявление дубликатов (Duplicate Detection), удаление дубликатов (Deduplication), связывание записей (Record Linkage), идентификация объектов (Object Identification), сопоставление связей (Reference Matching) и др. Проблема слияния данных является заключительным этапом интеграции данных. В работе дан краткий обзор методов разрешения сущностей и методов слияния данных. Затем в работе рассматриваются вопросы адаптации таких методов к их применению в ETL процессе при интеграции больших данных в Hadoop. Рассмотрены способы программирования методов разрешения сущностей и слияния данных как части ETL процесса.

## 1 Введение

В различных областях науки наблюдается экспоненциальный рост объема получаемых экспериментальных (наблюдательных) данных. Например, в астрономии текущий и ожидаемый темп роста данных от наземных и космических телескопов удваивается в течение периода от шести месяцев до одного года. Сложность использования таких данных увеличивается еще и вследствие их естественной разнородности. Разнообразие (информационная несогласованность) получаемой информации вызывается, в частности, не только большим числом организаций, производящих наблюдения, и их независимостью, но и разнообразием объектов наблюдения, непрерывным и быстрым совершенствованием техники наблюдений, вызывающим адекватные изменения структуры и содержания накапливаемой информации. Это приводит к необходимости использования неоднородной, распределенной информации, накопленной в течение значительного периода наблюдений технологически различными инструментами.

Для анализа больших объемов накапливаемых данных используются современные распределенные инфраструктуры обработки массивных данных (например, Hadoop [41, 45]). Основной особенностью подобных инфраструктур является почти линейная горизонтальная масштабируемость (производительность системы растет линейно относительно числа узлов кластера).

Главным достоинством подобных инфраструктур является возможность анализировать и обрабатывать разно-структурированные данные, например, реляционные, XML, JSON, тексты и другие. При этом возникает проблема интеграции информации, извлекаемой из разно-структурированных данных.

Традиционно процесс интеграции данных можно представить состоящим из следующих этапов:

- сопоставление схем (Schema Matching),
- интеграция схем (Schema Integration),
- трансформация данных (Data Transformation),
- разрешение сущностей (Entity Resolution [17, 22, 34]),
- слияние данных (Data Fusion [10]).

В разделах 2 и 3 дан краткий обзор традиционных методов разрешения сущностей и методов слияния данных. В разделе 4 показано как можно адаптировать стандартные методы разрешения сущностей при интеграции массивных данных в среде Hadoop. Наконец, в разделе 5 показаны способы программирования методов разрешения сущностей и слияния данных как части ETL процесса в Hadoop.

* *Работа выполнена при поддержке РФФИ (гранты 13-07-00579, 14-07-00548) и Президиума РАН (Программа фундаментальных исследований Президиума РАН № 16 «Фундаментальные проблемы системного программирования»).*

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

## 2 Краткий обзор методов разрешения сущностей

В общем случае под термином разрешения сущностей (entity resolution [17, 22, 25–26, 31–32, 34]) понимается извлечение информации об одной и той же сущности реального мира из разнообразных структурированных коллекций данных, приведение извлеченных данных к унифицированному представлению. При этом применяются методы извлечения, сопоставления (matching), группирования, связывания (linking), устранения дублирования (deduplication) различных представлений информации.

В общем случае процесс разрешения сущностей включает следующие этапы [26]:

- Подготовку данных (Data preparation);
- Выбор методов сопоставления данных (Match Feature);
- Определение методов разрешения пар сущностей (Pairwise ER);
- Определение ограничений (ER Constraints);
- Реализацию алгоритма.

Важным этапом для успешного разрешения сущностей является подготовка данных, которая включает нормализацию схем и нормализацию данных. Нормализация схем включает, например:

- сопоставление атрибутов схем (например, «контактный телефон» и «мобильный телефон»);
- слияние атрибутов (например, «полный адрес» получается из атрибутов «город», «индекс» «улица», …);
- слияние множественных значений и списков (например, «контактные телефоны» и «основной номер телефона», «дополнительный номер телефона») и др.

Нормализация данных может включать приведение к строчному или заглавному регистру; удаление разделителей; поиск и исправления опечаток; поиск сокращений и аббревиатур и замена их на полные стандартные формы; использование словарей для нормализации строк, и много другое.

Сопоставление сущностей может осуществляться разнообразными способами оценки сходства (similarity) сущностей. Мера сходства может быть как булева, так и вещественная. Применяют следующие методы оценки сходства:

- эквивалентность булевых предикатов;
- вычисление функции сходства значений (Levenstein [52], Smith-Waterman [52]);
- вычисление функции сходства множеств (Jaccard [52], Dice [52]);
- вычисление функции сходства векторов (Cosine similarity [49], TFIDF [50]);
- сходство на основе выравнивания (Jaro – Winkler [52], Soft – TFIDF [8], Monge – Elkan [51]);
- сходство фонетических данных: Soundex [52];
- сходство, основанное на переводе (может использоваться для нормализации аббревиатур);
- сходство, основанное на знаниях о предметной области, и др. [52].

Рассматривают также сходство отношений. Меры, используемые для отношений, обычно основаны на сходстве множеств, и предполагают использование аналогичных функций:

- Common Neighbors,
- Jaccard’s Coefficient,
- Adar Coefficient [1].

При сравнении пар сущностей они рассматриваются как вектора, для которых нужно вычислить их сходство. Традиционным подходом является подсчет сходства некоторым методом для каждого из атрибутов независимо. А затем реализуется подсчет взвешенной суммы. Например:

```
0.5*1st – author – match – score +
0.2*venue – match – score +
0.3*paper – match – score
```

Недостатком этого подхода является сложность выбора весов для каждого из атрибутов и сложность выбора порога сходства сущностей. Другим подходом является задание правил для каждого атрибута независимо. Например:

```
(1st – author – match – score > 0.7 AND
venue – match – score > 0.8)
OR (paper – match – score > 0.9 AND
venue – match – score > 0.9)
```

Недостатком этого подхода является сложность формулирования подобных правил вручную. Применяются также методы, основанные на модели Fellegi & Sunter [24].

Для сопоставления пар сущностей применяют также специальные методы машинного обучения, которые позволяют автоматизировать процесс формулирования критериев для сопоставления сущностей:

- Decision trees [18],
- Support vector machines [9, 16],
- Ensembles of classifiers [15],
- Conditional Random Fields (CRF) [27].

Недостатком этих подходов является: несбалансированность результирующих классифицированных множеств (так, в результате образуется значительно больше несхожих объектов, чем схожих), а также высока вероятность того, что объект не будет причислен ни к какому классу (схожих, несхожих). Но оба эти недостатка могут решаться путем тонкой настройки алгоритмов. Ключевой проблемой при использовании методов машинного обучения при сравнении пар сущностей является выбор обучающего множества.

Выделяют следующие методы, не требующие построения обучающей выборки для классификации сущностей:

- Обучение без учителя или с частичным привлечением учителя [29, 36, 42];

- Методы с активным обучением
  – Ансамбли классификаторов [38, 39];
  – Доказуемая оптимизация точности/полноты (precision/recall) [3, 4];
  – Краудсорсинг [33, 40].

Таким образом, при выборе методов сопоставления сущностей выделяют: множество алгоритмов сходства, методы, основанные на машинном обучении, и методы, основанные на активном обучении и краудсорсинге. Последняя группа методов сейчас считается наиболее перспективной, но требует проведения дополнительных исследований.

Примеры правил, используемых для установления сходства сущностей:

- Транзитивность: если M1 и M2 схожи, и M2 и M3 схожи, тогда и M1 и M2 схожи;
- Эксклюзивность: если M1 и M2 схожи, тогда M3 не может быть схож с M2;
- Функциональные зависимости: если M1 и M2 схожи, тогда M3 и M4 должны быть схожи.

Транзитивность часто используется для методов удаления дубликатов (Deduplication), а эксклюзивность используется в методах установления связей (Record Linkage).

В заключение можно отметить, что разрешение сущностей является быстро развиваемой областью. Исследуются новые меры сходства [52], ведутся работы по применению перспективных методов машинного обучения [3, 4, 33, 38–40]. Развивается применение функциональных зависимостей при очистке данных (data cleaning) [2, 13, 23]. Ведутся работы по построению сущностей с наиболее представительными данными (включающими данные из разнообразных дубликатов – Canonicalization [5]). Также ведутся работы по методам, когда решения по сходству двух сущностей принимается на основе анализа совокупности сущностей, применения вероятностных логик сходства, латентной модели Дирихле [6, 7, 14].

## 3 Краткий обзор методов слияния данных

Под слиянием данных (Data Fusion [10, 12, 21]) понимается образование интегрированного представления информации об одной же сущности реального мира, полученной из различных источников данных. Задачами процесса слияния данных является: слияние записей о сущностях, разрешение возможных конфликтов, обнаружение и удаления ошибочных данных. Методы слияния данных, кратко рассмотренные в данном разделе, исследованы в Потсдамском университете в диссертации [12]. Различные аспекты проблемы слияния данных представлены на рис. 1.

### 3.1 Типы конфликтов при слиянии данных

Различают два типа конфликтов: конфликты, вызванные неопределенными значениями и конфликты, вызванные противоречивыми значениями. Неопределенность означает, что в одном источнике данных содержатся неизвестные значения (null), а в другом известные. Проблема заключается в том, что семантика неопределенных значений (null) может сильно отличаться. Различают три варианта: неизвестные значения, несуществующие значения (например, атрибут «имя супруга» всегда будет null для неженатых), скрытые значения (такие данные, которые по каким-то причинам не позволено видеть). Противоречивость значений означает появление двух различных не нулевых (not null) значений. Возможны различные стратегии обработки подобных конфликтов.

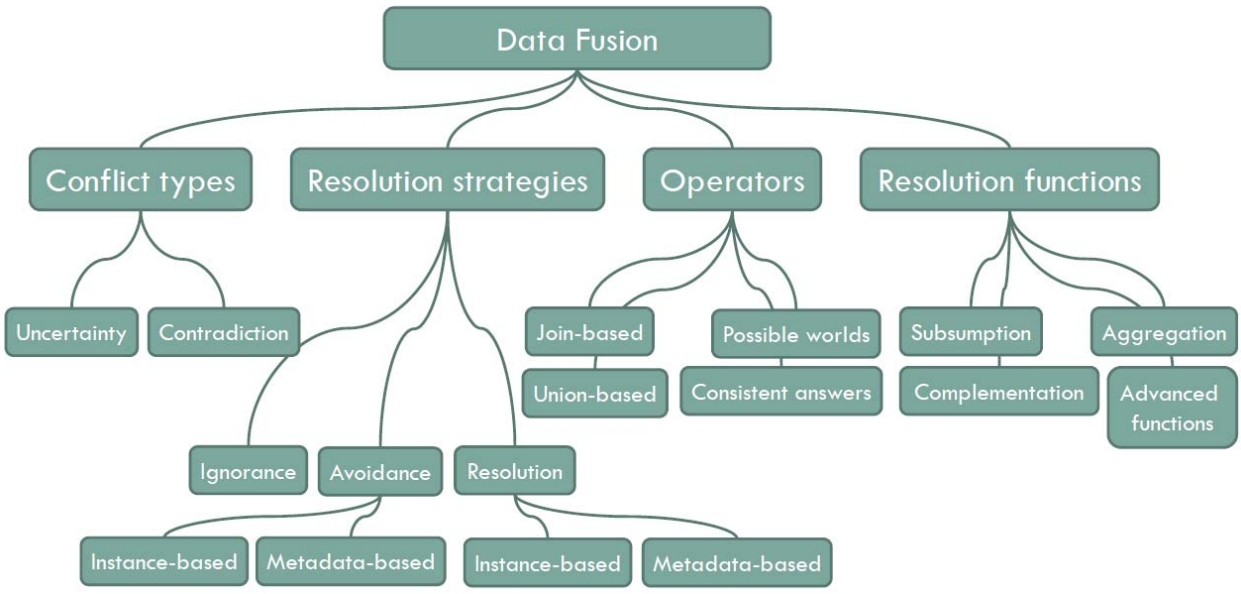



Рис. 1. Различные аспекты проблемы слияния данных

### 3.2 Стратегии разрешения конфликтов

Различают следующие виды подходов к разрешению конфликтов:

- игнорирование конфликтов;
- избегание конфликтов;
- разрешение конфликтов.

Стратегия игнорирования конфликтов предполагает извлечение всей доступной информации. Например, для строк это может быть обычная конкатенация строк, а пользователь уже сам решает, какие данные верны.

Стратегия избегания конфликтов предполагает выбор данных на основе самих данных (по некоторому алгоритму) или на основе метаданных. Примером функции на основе данных может служить функция coalesce (выбор первого не нулевого значения), или функция выбора самого длинного значения. Примером функций на основе метаданных может выступать выбор в зависимости от самого источника (например, известно, что один из источников наиболее достоверный). Другим примером является функция выбирающая значение из того источника, в котором большее число значений было выбрано для других атрибутов.

Стратегии разрешения конфликтов учитывают все значения, и выбирают из них «достоверное». Примером подобной функции могут выступать всевозможные функции голосования, функции выбора случайного значения, функции среднего значения, функции наиболее часто встречающегося значения и др.

### 3.3 Основные функции разрешения конфликтов

Вводится операция ***outer union*** [12], результатом которой является объединение двух отношений. Если схемы не совпадают, то результирующая схема является объединением двух исходных схем. Например, пусть даны два отношения A с набором атрибутов = {a, b, c, d}, и отношение B с набором атрибутов = {c, d, e, f}. Результирующая схема будет содержать набор атрибутов = {a, b, c, d, e, f}. В результирующие кортежи для недостающих атрибутов помещаются нулевые значения. Эта операция не является стандартной и отсутствует в большинстве реляционных СУБД. В реляционной алгебре подобная операция может быть представлена как:

```
(SELECT a, b, c, d, NULL as e, NULL as f
FROM A)
UNION
(SELECT NULL as a, NULL as b, c, d, e, f
FROM B)
```

Вводится функция tuple ***subsumption*** [12]. Говорят, что кортеж t1 поглощает другой кортеж t2 (поглощаемый кортеж), если у них:

- совпадают схемы;
- в t2 больше неизвестных (null) значений чем в t1;
- в t2 все известные значения совпадают со значения в t1.

Например, пусть дан кортеж t1 = (5, ‘text’, null, 7) и t2 (5, null, null, 7). Видно, что каждый атрибут в t2 либо совпадает с аналогичным атрибутом в t1, либо он null. Для этого примера кортеж t1 поглощает кортеж t2.

Вводится функция ***tuple complementation*** [12]**.** Говорят, что кортежи t1 и t2 дополняют друг друга если:

- у них совпадают схемы;
- они не совпадают;
- значения соответствующих атрибутов в t1 и t2 совпадают, либо одно их не определено, либо оба не определены;
- t1 и t2 имеют как минимум один атрибут, значения которого совпадают.

Например, пусть дан кортеж t1 = (5, ‘text’, null, null) и t2 (5, null, null, 7). Видно, что кортежи дополняют друг друга. Результатом операции дополнения для этих двух кортежей будет новый кортеж t = (5, ‘text’, null, 7).

### 3.4 Операторы слияния данных

Различают два основных подхода к слиянию данных. Это подходы основаны на операции объединения (union based) или на операции соединения (join based). Различают следующие основные операции.

***Minimum Union*** [12] (union based). Операция представляет собой выполнение операции ***outer union***, а затем удаления из результата всех поглощаемых (subsumed [12]) кортежей. Пример операции представлен на рис. 2.

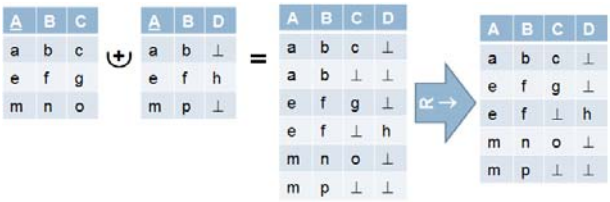


Рис. 2. Пример операции Minimum Union

***Complemetation Union*** [12] (union based). Операция представляет собой выполнение операции outer union, а затем дополнения (complementation) всех возможных кортежей. Пример операции представлен на рис. 3.

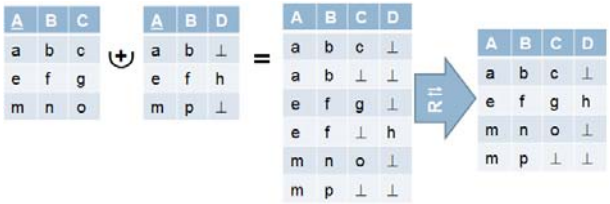


Рис. 3. Пример операции Complementation Union

**Grouping and Aggregation** [12] (union based). Операция предполагает выполнение outer union, а

затем группировки по общему атрибуту и применения функции агрегации к остальным атрибутам. Пример операции на языке SQL представлен ниже.

```
WITH OU AS (
  ( SELECT A, B, C, NULL AS D FROM U1 )
       UNION (ALL)
  ( SELECT A, B, NULL AS C, D FROM U2 )
),

SELECT A, MAX(B),  MIN(C), SUM(D)
FROM   OU
GROUP BY A
```

**Full Disjunction** [37] (join based). Операция представляет собой ***full outer join*** (стандартная реляционная операция), после чего применяется ***subsumption*** к результату. Пример представлен на рис. 4.

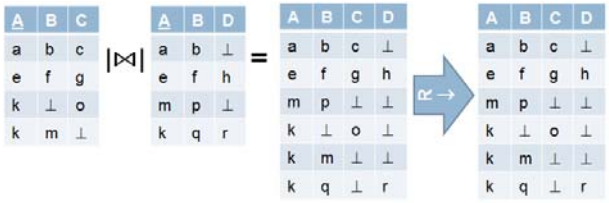


Рис. 4. Пример операции Full Disjunction

**Match Join** [12] (union+join based). В операции выбираются всевозможные комбинации значения атрибутов, после чего выполняется ***full outer join***. Фактически реализуется ***outer union*** двух коллекций. После чего определяется N – 1 вспомогательных отношений, где N – число атрибутов, каждое из которых содержит по два атрибута, один общий, и какой-то другой. После чего происходит ***full outer join*** N – 1-го отношения. Пример реализации операции на языке SQL представлен ниже.

```
WITH
  OU(A,B,C,D) AS (
  ( SELECT A, B, C, NULL AS D FROM U1 )
    UNION
  ( SELECT  A, B, NULL AS C, D FROM U2 )
),// ← Outer Union
  B_V (A,B) AS ( SELECT DISTINCT A, B
FROM OU ), // ← 1-е отношение (N = 4)
  C_V (A,C) AS ( SELECT DISTINCT A, C
FROM OU ), // ← 2-е отношение (N = 4)
  D_V (A,D) AS ( SELECT DISTINCT A, D
FROM OU ), // ← 3-е отношение (N = 4)
SELECT A, B, C, D
FROM B_V FULL OUTER JOIN C_V FULL OUTER
JOIN D_V // ← Full Outer Join
```

**Merge** (union+join based). Операция объединяет операции соединения и объединения. Для каждого общего атрибута формируются две версии значений, нулевые значения удаляются функцией COALESCE (выбор первого ненулевого значения). Пусть даны два отношения A с набором атрибутов {a, b, c} и B с набором атрибутов {a, b, d}. a – конфликтующий атрибут, b – атрибут с нулевыми значениями. Пример реализации на SQL представлен ниже, а результат операции представлен на рис. 5.

```
(SELECT A.a, COALESCE(A.b, B.b), A.c, B.d
FROM A LEFT OUTER JOIN B ON A.a = B.a)
UNION
(SELECT B.a, COALESCE(B.b, A.b), A.c, B.d
FROM A RIGHT OUTER JOIN B ON A.a = B.a)
```

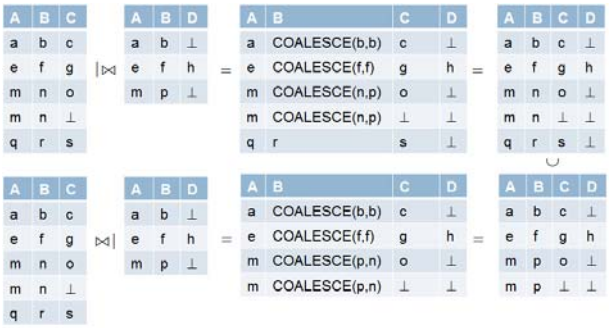


Рис. 5. Пример операции Merge

## 4 Разрешение сущностей для больших данных

Для манипулирования большими разно-структурированными данными служат Hadoop инфраструктуры [41, 45], предоставляющие масштабируемое хранилище и обеспечивающие высокую скорость обработки больших данных за счет распределенной их обработки. Таким образом, для применения методов нужна *адаптация* алгоритмов для их распределенного выполнения на различных узлах Hadoop кластера.

В среде Hadoop реализована парадигма распределенного программирования для анализа данных Map-Reduce [20, 46], по имения основных функций. В начале на всех узлах кластера обрабатываются блоки данных независимо друг от друга (Map). После чего данные группируются по заранее выбранным для алгоритма ключам и поступают на выполнение на один или более узел в зависимости от алгоритма (Reduce).

Таким образом для реализации любого алгоритма в Hadoop инфраструктуре требуется его адаптация к виду Map-Reduce. Другим вариантом является реализация алгоритма на одном из языков высокого уровня, таких как: Pig [47], Hive [48], Jaql [43, 44]. Все эти языки автоматически переписывают программы, реализованные на них, в Map-Reduce приложения для выполнения на Hadoop кластере.

В случае больших данных и распределенных инфраструктур традиционные подходы требуют доработок. Различают два основных метода разрешения сущностей над большими данными: разбиение данных на блоки (blocking [19, 35]) и распределенный метод разрешения сущностей.

Суть разбиения на блоки заключается в следующем. Пусть у нас представлены 1000 компаний в 1000 городах. И нам нужно сравнить компании. Алгоритм полного попарного сравнения

потребует $10^{12}$ сравнений. При этом, если предположить, что компании из разных городов не могут совпадать, то потребуется $10^9$ сравнений. Ключевой проблемой данного подхода является выбор критерия, по которому разбивать данные. Различают два основных метода: основанный на хэш функции [26], и основанный на сходстве соседей [26]. Метод, основанный на хэш функции, предполагает разбиение на блоки по хэш ключу. Основной проблемой алгоритма является выбор хэш функции. Метод, основанный на сходстве соседей, предполагает, что совпадать могут только объекты, схожие по некоторой мере. Все объекты сортируются по какому-то признаку (ключу – простому или составному, уникальность ключа не требуется). После этого выбирается размер окна. И объекты сравниваются только внутри окна. Проблемой данного метода является выбор ключа сортировки.

Распределенный метод разрешения сущностей предполагает реализацию традиционных алгоритмов этого семейства в виде Map-Reduce приложения, что требует зачастую полного пересмотра исходного алгоритма. Другой вариант – реализация алгоритма разрешения сущностей на специализированных языках, чему будет посвящен следующий раздел. Третий вариант – использование специализированных инструментов, направленных на распределенное выполнение методов разрешения сущностей над Hadoop [30].

## 5 Программирование операций разрешения сущностей и слияния данных на языке HIL

Язык HIL (Highlevel Integration Language) [28], новый специализированный язык, разработанный IBM, ориентированный на разрешение и интеграцию сущностей в Hadoop инфраструктуре.

HIL компилируется в язык Jaql [43, 44], который в свою очередь автоматически переписывается в Map-Reduce, если этого требует алгоритм.

### 5.1 Реализация методов разрешения сущностей

Пусть даны структуры данных, включающие три атрибута: id, value, name. Тогда простейшее правило разрешения сущностей на языке HIL будет выглядеть следующим образом:

```
declare Duplicated: ?;
declare Generated: ?;
declare Deduplicated: ?;

create link Deduplicated as
select
[gen: [id: g.id, name: g.name, value:
g.value],
dup: [id: d.id, name: d.name, value:
d.value]]
from Generated g, Duplicated d
match using
  rule_id: g.id = d.id,
  rule_name: g.name = d.name,
  rule_value: g.value = d.value;
```

В этом примере используется простое сопоставление сущностей, по совпадению значений. Если требуется ввести какую-то функцию меры для значений, это можно реализовать внешней функцией Jaql:

```
@jaql{
compareValue =
javaudf("org.ipiran.similarity.ValueSimil
arity");
}
```

После этого такую функцию можно вызывать из языка HIL:

```
declare compareValue: function ? to ?;
declare Duplicated: ?;
declare Generated: ?;
declare Deduplicated: ?;

create link Deduplicated as
select
[gen: [id: g.id, name: g.name, value:
g.value],
dup: [id: d.id, name: d.name, value:
d.value]]
from Generated g, Duplicated d
match using
rule_id:
  compareValue(g.id, d.id) > 0.7,
rule_name:
  compareValue(g.name, d.name) > 0.7,
rule_value:
  compareValue(g.value, d.value) > 0.7;
```

Можно также ввести меру для сравнения не отдельных значений, а для сравнения объектов целиком. Пусть описана функция **compareObject**, которая принимает на вход объекты, тогда правило на языке HIL изменится, т.к. в этом случае используется другой вид правил:

```
insert into Deduplicated
select
[gen: [id: g.id, name: g.name, value:
g.value],
dup: [id: d.id, name: d.name, value:
d.value],
value: compareObject(g,d)]
from Generated g, Duplicated d
where compareObject(g, d) > 0.7;
```

Во всех этих случаях происходит сравнение всех объектов со всеми, сложность подобного сравнения $O(n^2)$. Несмотря на то, что сравнения будут выполняться независимо и распределены на всех узлах кластера (т.к. HIL переписывается в Jaql, а тот в свою очередь в Map-Reduce), время их выполнения может быть достаточно большим. Для уменьшения количества сравнений, как было описано в четвертом разделе, можно разбивать данные на блоки.

Пусть имеется функция **calcHash**, которая вычисляет hash для объектов. В результате функция может выдавать столько уникальных значений, на сколько блоков нам надо разбить данные. Тогда объединив правила, рассмотренные выше, выбрав в начале те объекты что совпадают по хэш функции, а далее вычислив общую меру, можно получить результат за более короткое время:

```
declare calcHash: function ? to ?;
insert into GeneratedHash
select [$.*, hash: calcHash($.*)]
from Generated;
insert into DuplicatedHash
select [$.*,hash: calcHash($.*)]
from Duplicated;

create link Deduplicated as
select [
  gen: [id: g.id, name: g.name, value:
g.value],
  dup: [id: d.id, name: d.name, value:
d.value]]
from GeneratedHash g, DuplicatedHash d
match using
  rule_id: g.hash = d.hash;
insert into Measured
select [gen: dd.gen, dup: dd.dup, value:
compareObject(dd.gen, dd.dup)]
from Deduplicated dd
where compareObject(dd.gen, dd.dup) > 0.8
```

### 5.2 Реализация методов слияния данных

На данном этапе будем считать, что этап разрешения сущностей уже пройден и нам дана некоторая коллекция Deduplicated где уже установлены соответствия одним из выше перечисленных способов. Например, пусть у нас есть две коллекции A (id, a, b, c) и B (id, a, b, d). Атрибуты a, b, c, d могут содержать NULL значения, атрибуты id совпадают. Ниже дан пример подобных данных для коллекции A в формате JSON:

```
[{"a":null,"b":null,"c":"wmqhxfgmac","id"
:919132322},
{"a":null,"b":null,"c":"wmqhxfgmac","id":
919132322}]
```

Тогда коллекция разрешенных сущностей может быть получена следующим образом:

```
create link Deduplicated as
select
[gen: [id: a.id, a:a.a, b:a.b, c:a.c],
dup:  [id: b.id, a:b.a, b:b.b, d:b.d]]
from A a, B b
match using
  rule1: a.id = b.id;
```

Рассмотрим теперь реализацию Minimum Union и Fusion оператор [11] на языке HIL.

Как было описано в третьем разделе, **Minimum Union** - это последовательное применение операций outer union и subsumption [12]. Outer Union фактически реализуется с помощью индекса **FusionIndex**. Использование индекса оправдано, т.к. существует несколько записей, описывающих одну сущность. Ключом является атрибут id. Ниже представлена реализация операции Outer Union.

```
insert into FusionIndex![id: f.gen.id]
select [a: f.gen.a, b: f.gen.b, c:
f.gen.c] from Deduplicated f;

insert into FusionIndex![id: f.dup.id]
select [a: f.dup.a, b: f.dup.b, d:
f.dup.d] from Deduplicated f;
```

Далее для реализации subsumption требуется удалить все ненужные кортежи. Это делается на языке Jaql. Для этого нужна функция, которая бы определяла, поглощается ли один кортеж другим. К сожалению, в языке Jaql нет возможностей написания общих (generic) методов, универсальных для всех коллекций, поэтому функцию сравнения можно реализовать на java и подключить к языку Jaql как демонстрировалось в разделе 5.1 на примере функций вычисления меры. Либо же можно реализовать функцию для сравнения конкретных коллекций на языке Jaql, как показано ниже:

```
is_subsumed = fn(i,j) ((
isnull(j.a) or (i.a == j.a) ) and (
isnull(j.b) or (i.b == j.b) ) and (
isnull(j.c) or (i.c == j.c) ) and (
isnull(j.d) or (i.d == j.d) ) and (
i != j)) ;
```

Функция **is_subsumed(i,j)** проверяет, поглощает ли один кортеж другой кортеж при помощи попарного сравнения атрибутов или проверки на null.

```
removeSubsumed = fn (a) ( b = a,
subs = for (i0 in b) [a->filter
is_subsumed(i0,$)], s = subs -> expand,
a -> filter not $ in s);
```

Функция **removeSubsumed** удаляет все поглощенные записи из кортежа. Здесь реализован наивный алгоритм, который попарно для каждого кортежа находит все поглощенные им, и удаляет их.

```
minUnion = fn(id,a) ( {id:id, minunion :
removeSubsumed(a)});
```

Функция **minUnion** нужна для построения результирующих кортежей при реализации Minimum Union.

Теперь операцию Minumum Union можно описать следующим образом на языке HIL:

```
insert into MinimumUnion
select minUnion(i.dup.id,
            FusionIndex![id : i.dup.id])
from Deduplicated i;
```

Для каждого id достаются все соответствующие записи и удаляются те, которые ими поглощаются.

**Data Fusion оператор** [11] представляет собой особый вид функции, использующий группировку для преодоления конфликтов. Основная идея заключается в группировке различных представлений одной и той же сущности по общему атрибуту, а затем в применении функций разрешения конфликтов для всех остальных атрибутов, сливая данные в одну сущность. Различают два вида стратегии для функций разрешения конфликтов:

- deciding-стратегия – заключается в выборе какого-то одного значения каким-то способом (минимум, максимум, случайное значение);
- mediating-стратегия – заключается в агрегации всех значений (среднее значение, сумма).

Пусть имеются две коллекции A (id, name, age) и B (id, name, info), пример которых дан ниже:

```
A
[{"id":760046903,"name":null,"age":null},
{"id":15009544,"name":
"zvqcsxkzxk","age":938781652}]

B
[{"id":15009544,"name":null,"info":null},
{"id":760046903,"name":"pjltaghyug","info
":null}]
```

Пусть для них пройден этап разрешения сущностей и построена коллекция Deduplicated как описано выше в этом разделе. Пусть также для этих данных построен индекс FusionIndex, как показано выше для операции Minimum Union. Тогда Data Fusion Оператор на языке HIL может быть описан следующим образом:

```
@jaql{
average = fn($a) avg($a[*].age);
any = fn($a) any($a[*].name);
concat = fn ($a) strJoin($a[*].info,"_");
}

insert into Fused
select [
id : i.dup.id,
age:
    average(FusionIndex![id : i.dup.id]),
name:
    any(FusionIndex![id : i.dup.id]),
info:
    concat(FusionIndex![id: i.dup.id])]
from Deduplicated i;
```

Функции вычисления среднего, выбора случайного не-null значения, а также конкатенации реализованы на Jaql. Данное правило образует коллекцию **Fused**, причем для атрибута *age* будет подсчитано среднее значение, для имени *name* выбрано любое ненулевое значение, а для атрибута *info* будет получена конкатенация всех доступных значений. Таким образом, в данном примере показана реализация обеих стратегий для функций разрешения конфликтов в Data Fusion операторе.

## 6 Заключение

Рассмотренные методы и операции извлечения и интеграции информации о сущностях реального мира позволяют программировать интеграционные потоки вида ETL, образующие интегрированные структурированные данные, которые могут быть использованы в приложениях для дальнейшего анализа и обработки. В статье рассмотрены методы разрешения сущностей и слияния данных. В статье показаны способы программирования методов и операций извлечения и интеграции информации о сущностях реального мира, включая методы слияния данных на декларативном языке HIL.

## 7 Литература


[1] LA Adamic, E Adar. Friends and neighbors on the Web. Social networks 25 (3), 211–230, 932, 2003.

[2] Rohit Ananthakrishna et Al., Eliminating fuzzy Duplicates in data warehouses, VLDB 2002.

[3] A. Arasu et al., On active learning of record matching packages, SIGMOD 2010.

[4] K. Bellare et al., Active sampling for entity matching, KDD 2012.

[5] O. Benjelloun et al., Swoosh: A generic approach to Entity Resolution, VLDBJ. 18(1), 2009.

[6] I. Bhattacharya & L. Getoor, Collective Entity Resolution in Relational Data, TKDD 2007.

[7] I. Bhattacharya & L. Getoor, A Latent Dirichlet Model for Unsupervised Entity Resolution, SDM 2007.

[8] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg. Adaptive name matching in information integration. IEEE Intelligent Systems, 18(5):16–23, Sep./Oct. 2003.

[9] M. Bilenko & R. Mooney, Adaptve Duplicate Detecton Using Learnable String Similarity Measures, KDD 2003.

[10] J. Bleiholder, F. Naumann. Data Fusion. ACM Computing Survey 2009.

[11] J. Bleiholder, F. Naumann. F. Declarative data fusion – syntax, semantics, and implementation. In Proceedings of the East European Conference on Advances in Databases and Information Systems (ADBIS), p. 58–73, 2005.

[12] J. Bleiholder. Data Fusion and Conflict Resolution in Integrated Information Systems. Dissertation, Hasso-Plattner-Institut, 2010.

[13] P. Bohannon et al., Conditional Functional Dependencies for Data Cleaning, ICDE 2007.

[14] M. Broecheler & L. Getoor, Probabilistic Similarity Logic, UAI 2010.

[15] Z. Chen et al., Exploiting context analysis for combining multiple entity resolution systems, SIGMOD 2009.

[16] P. Christen, Automatic record linkage using seeded nearest neighbour and support vector machine classificaton., KDD 2008.

[17] P. Christen. Data Matching – Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Data-Centric Systems and Applications, 2012.

[18] M. Cochinwala et al., “Efficient data reconciliaton”, Information Sciences, 2001.

[19] A. Das Sarma et al., “An Automatic Blocking Mechanism for Large-Scale De-duplication Tasks”, CIKM 2012.

[20] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. OSDI'04: Sixth Symposium on Operating System Design and Implementation, 2004.

[21] Xin Luna Dong, Felix Naumann. Data Fusion – Resolving data conflicts in Integration. VLDB 2009.

[22] Wenfei Fan, Floris Geerts. Foundations of Data Quality Management. Synthesis Lectures on Data Management № 29, 2012.

[23] Wenfei Fan, Dependencies revisited for improving data quality, PODS 2008.

[24] Fellegi, Ivan; Sunter, Alan. A Theory for Record Linkage. Journal of the American Statistical Association 64 (328): pp. 1183–1210. 1969.

[25] Venkatesh Ganti, Anish Das Sarma. Data Cleaning, A Practical Perspective. Synthesis Lectures on Data Management № 36, 2013.

[26] Lise Getoor, Ashwin Machanavajjhala. Entity Resolution for Big Data. 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Chicago: ACM SIGKDD, 2013.

[27] R. Gupta & S. Sarawagi, Answering Table Augmentaton Queries from Unstructured Lists on the Web, PVLDB 2(1), 2009.

[28] Mauricio Hernández, Georgia Koutrika, Rajasekar Krishnamurthy, Lucian Popa, Ryan Wisnesky. HIL: a high-level scripting language for entity integration. EDBT'13 Proceedings of the 16th International Conference on Extending Database Technology. P. 549–560, 2013.

[29] T. Herzog et al., Data Quality and Record Linkage Techniques, Springer, 2007.

[30] Kolb, L.; Thor, A.; Rahm, E. Dedoop: Efficient Deduplication with Hadoop Proc. 38th Intl. Conference on Very Large Databases (VLDB) / Proc. of the VLDB Endowment 5(12), 2012.

[31] Hanna Köpcke, Andreas Thor, Erhard Rahm. Evaluation of entity resolution approaches on real-world match problems. Proceedings of the VLDB Endowment, Volume 3, Issue 1–2, September 2010, P. 484–493.

[32] Hanna Köpcke, Erhard Rahm. Frameworks for entity matching: A comparison. Data & Knowledge Engineering, Volume 69, Issue 2, February 2010. P. 197–210.

[33] A. Marcus et al. Human-powered Sorts and Joins. PVLDB 5(1), 2011.

[34] Felix Naumann, Melanie Herschel. An Introduction to Duplicate Detection. Synthesis Lectures on Data Management. № 3, 2010.

[35] G. Papadias et al., Beyond 100 million entities: large-scale blocking-based resoluton for heterogenous data, WSDM 2012.

[36] P. Ravikumar & W. Cohen, A Hierarchical Graphical Model for Record Linkage, UAI 2004.

[37] A. Rajaraman and J. D. Ullman. Integrating information by outerjoins and full disjunctions. PODS1996.

[38] S. Sarawagi et al., Interactive Deduplication using Active Learning, KDD 2000.

[39] S. Tejada et al., Learning Object Identification Rules for Information Integration, IS 2001.

[40] J. Wang et al., CrowdER: Crowdsourcing Entity Resolution, PVLDB 5(11), 2012.

[41] Tom White. Hadoop: The Definitive Guide. O'Reilly Media; Third Edition. 2012.

[42] W. Winkler, Overview of Record Linkage and Current Research Directions, Research Report Series, US Census, 2006.

[43] Jaql Overview: Jaql, a query language for JavaScript Object Notation (JSON), 2011. https://code.google.com/p/jaql/wiki/JaqlOverview

[44] IBM InfoSphere BigInsights Version 3.0, Jaql reference. 2014. http://www-01.ibm.com/support/knowledgecenter/SSPT3X_3.0.0/com.ibm.swg.im.infosphere.biginsights.jaql.doc/doc/c0057749.html

[45] Apache Hadoop 2.4.1, 2014. http://hadoop.apache.org/

[46] MapReduce Tutorial, 2013. http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

[47] Apache Pig Project, 2014 http://pig.apache.org/

[48] The Apache Hive data warehouse, 2014. http://hive.apache.org/

[49] Cosine similarity. http://en.wikipedia.org/wiki/Cosine_similarity

[50] Term frequency–inverse document frequency. http://en.wikipedia.org/wiki/Tf%E2%80%93idf

[51] Monge-Elkan Distance Function. http://www.gabormelli.com/RKB/Monge-Elkan_Distance_Function

[52] String metric. http://en.wikipedia.org/wiki/String_metric

## Programming of the Entity Resolution and Data Fusion Methods while Implementing ETL in the Hadoop Environment

A. Vovchenko, L. Kalinichenko, D. Kovalev

The paper is devoted to the problem of Entity Resolution and Data Fusion implementation in the context of big data integration. Entity resolution cares of Duplicate Detection, Deduplication, Record Linkage, Object Identification, Reference Matching, and other ETL-related tasks. Data fusion is the final step in the data integration process. This paper gives a short overview of methods for entity resolution and data fusion techniques. Then the paper presents the techniques for programming of the entity resolution and data fusion methods for implementing of the ETL process in the Hadoop environment.

# Интеграция библиографических данных в Linked Open Data

© Д. А. Малахов © В. А. Серебряков © К. Б. Теймуразов © О. Н. Шорин
Вычислительный центр им. А. А. Дородницына РАН,
Москва
malahov@iqbuzz.ru seerebr@ccas.ru kbt@ccas.ru shorin@nlr.ru

**Аннотация**

В данной работе рассматривается проблема интеграции библиографических записей. Была поставлена задача объединить данные Российской национальной библиотеки и Британской национальной библиотеки в рамках пространства Linked Open Data. В ходе решения задачи была построена прототипная система, с помощью которой данные формата RUSMARC могут быть опубликованы в Linked Open Data и связаны с другими библиотечными данными. Были проведены опыты и получены оценки работы подсистемы связывания данных разных источников.

## 1 Введение

### 1.1 Семантическая паутина

Интернет с самого начала представлял собой множество разрозненных сайтов, никак не связанных друг с другом по смыслу. С течением времени появилось все больше и больше ресурсов, посвященных одним и тем же проблемам. Поиск нужной информации становился все более затруднительным, в то же время росли требования к качеству поиска.

Появилась потребность в семантическом поиске, поиске не по словам, а по смыслу, а также в связывании данных близких по смыслу, но находящихся в разных ресурсах. Стало ясно, что существующие стандарты не в состоянии удовлетворить потребности людей, необходимо было создавать новые стандарты создания структурированных данных, по которым возможен семантический поиск.

Термин Semantic Web был впервые введён сэром Тимом Бернерсом-ли в журнале «Scientific American», и называется им «следующим шагом в развитии Всемирной паутины». Концепция Semantic Web была принята и продвигается W3C (Консорциумом Всемирной паутины).

Идея этой концепции - создать общепринятый способ совместного использования данных различными приложениями, организациями и сообществами, и предоставление возможности получать данные, как вручную, так и автоматическими средствами [1].

Для поддержки этой концепции W3C создал стандарты, понятия, технологии и форматы. В них входят URI (Uniform Resource Identifier), RDF(Resource Description Framework), OWL(Web Ontology Language), SPARQL (Protocol and RDF Query Language).

URI (Uniform Resource Identifier) – последовательность символов, идентифицирующая абстрактный или физический ресурс.

RDF – модель данных, служащая платформой для представления информации. Структура, лежащая в основе любых выражений в RDF, это коллекция триплетов, каждый из которых состоит из субъекта, предиката и объекта. Набор таких триплетов называется RDF-графом. По своей природе это ориентированный помеченный мульти-граф. Каждый триплет представляет объявление отношения между предметами. Выражение RDF триплета говорит о том, что некоторое отношение, указанное предикатом, связывает предметы, обозначенные как субъект и объект, в триплете. Узлами RDF-графа являются объекты и субъекты. Узлы обычно идентифицируется с помощью URI, однако бывают пустые и литеральные узлы. Дуги (предикаты) всегда идентифицируются с помощью URI [2].

OWL – это язык для определения и представления онтологий. Онтология предназначена для описания семантики данных. Она может включать описания классов, свойств, экземпляров классов, их операций. Формальная семантика OWL описывает, как получать логические следствия, имея такую онтологию, т.е. получить факты, которые не представлены в онтологии буквально, но следуют из ее семантики. При построении логических выводов используется модель открытого мира, т.е. если не может быть доказано, что некое утверждение истинно, из этого не следует, что оно ложно [3].

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

Кроме OWL для описания онтологий также используют язык RDFS (Resource Description Framework Schema). Как правило OWL и RDFS используются совместно.

SPARQL – язык запросов для обращений к RDF-хранилищам, служит тем же целям, что и SQL в области реляционных баз данных. SPARQL точка доступа - сервер обрабатывающий запросы.

### 1.2 Linked Open Data

LOD (Linked Open Data) – проект, целью которого является наполнение сети Интернет данными в стандартных форматах Semantic Web, а также устанавливание связей между данными из различных источников [4].

Тим Бернерс-Ли сформулировал следующие четыре принципа связанных данных [5]:

– Использование URI для идентификации сущностей.

– Использование HTTP URI, чтобы эти сущности могли быть найдены людьми.

– При обращении по URI предоставлять полезную информацию о сущности, используя стандартизованные форматы (RDF, SPARQL).

– Предоставлять также другие, связанные URI, для облегчения поиска.

На данный момент опубликовано более 40 млрд троек в рамках этого проекта. Самым крупным источником является DBPedia, более 3,5 млн сущностей, извлеченных из проекта Wikipedia.

### 1.3 Интеграция данных

В каждой предметной области существует много разрозненных источников. Каждая организация может оперировать только той информацией, которая у нее есть. Задача сбора информации часто бывает нетривиальной. Интеграция с пространством Linked Open Data является одним из универсальных решений данной задачи.

Linked Open Data было создано для того, чтобы в каждой предметной области интегрировать внутри себя как можно больше информации. Таким образом, публикуя данные в этом пространстве, мы с одной стороны получаем доступ ко всей информации, которая нас интересуют через свои данные, а с другой даем доступ к своей информации.

## 2 Постановка задачи

Целью данной работы является интеграция и обогащение библиографических записей, предоставленных РНБ (Российская Национальная Библиотека), с данными БНБ (Британская Национальная Библиотека). Набор данных РНБ насчитывает несколько миллионов библиографических записей. Для интеграции был предоставлен тестовый набор данных (около 17 тыс. единиц). Набор данных БНБ насчитывает 3,5 млн единиц, он опубликован в LOD.

Для достижения цели нужно решить задачи:

– Опубликовать данные РНБ согласно принципам LOD.

– Связать данные РНБ с опубликованными в LOD данными БНБ.

### 2.1 Публикация данных

Для решения этой задачи нужно решить подзадачи:

– Описать предметную область.

– Конвертировать данные РНБ в RDF.

– Настроить семантическое хранилище RDF данных РНБ.

– Предоставить доступ к данным РНБ.

#### 2.1.1 Описание предметной области

Необходимо выбрать термины из существующих онтологий, и на их основании создать собственную онтологию. Если найдутся данные, которые нельзя представить в рамках существующих терминов, создать собственные термины и дополнить онтологию.

#### 2.1.2 Конвертация данных

РНБ предоставила данные в формате RUSMARC. Для того, чтобы опубликовать данные согласно принципам LOD, они должны быть представлены в формате RDF с использованием терминов составленной онтологии. Нужно создать инструмент, переводящий RUSMARC формат в RDF.

#### 2.1.3 Создание семантического хранилища

SPARQL точка доступа – сервер, принимающий запросы на языке SPARQL и выдающий данные в формате RDF.

Публикация данных в LOD подразумевает создание семантического хранилища и SPARQL точки доступа, привязанной к нему. Необходимо выбрать одно из существующих семантических хранилищ, загрузить в него данные и настроить логический вывод.

#### 2.1.4 Предоставление доступа к данным

Нужно создать web-сервер, который выдавал бы информацию об объектах хранилища по HTTP запросу, отправленному на URI объекта.

### 2.2 Связывание

Согласно принципам LOD нужно задать как можно больше связей между данными РНБ и данными БНБ. Кроме того необходимо связать данные РНБ между собой.

Итак, следует разработать и реализовать алгоритм связывания разных библиографических

записей и сохранить полученные связи в семантическом хранилище. Необходимо учитывать связи при обработке запросов пользователя.

# 3 Описание предметной области

## 3.1 Общий обзор

Библиографическая запись – это элемент библиографической информации, фиксирующий сведения о документе – объекте записи, позволяющие его идентифицировать, раскрыть его состав и содержание в целях библиографического поиска [6].

Библиографическая запись включает в себя следующие части:

– заголовок;

– классификационные индексы;

– аннотация;

– язык;

– издательство;

– авторы;

– другая дополнительная информация.

## 3.2 Публикация библиографических записей в LOD

Одна из задач библиотеки – предоставление и обработка информации о всевозможных публикациях, а именно метаданные этих публикаций. К ним относятся: описание публикации, информация об авторе, издательстве и т.д. Поэтому интеграция данных различных библиотек является довольно актуальной проблемой.

Публикация данных в LOD вызывает огромный интерес в библиотечном сообществе, т.к. имеет ряд преимуществ по сравнению с другими способами обмена данными между библиотеками:

– В LOD для идентификации ресурса используют единое глобальное пространство имен, объекты идентифицируются с помощью URI, который является уникальным для всего LOD.

– Высокая способность к масштабированию, т.к. не обязательно хранить все данные об определенном объекте в одном хранилище или в одном источнике.

– Обмен данными можно осуществлять порциями, в виде множества законченных утверждений, необязательно единовременно передавать всю существующую информацию об объекте, т.к. ее можно получить в любой момент, когда она потребуется.

– В семантических хранилищах имеется логический вывод, это упрощает интеграцию данных с разными схемами данных (достаточно определить соответствия между схемами и создать связи между объектами) и позволяет выводить новые знания на основе имеющихся (нет необходимости хранить данные, которые могут быть получены с помощью логического вывода).

Несмотря на все плюсы, не так много библиотек внедряют подобные решения. В докладе Инкубаторной группы W3C по библиотечной модели LOD опубликованы причины, мешающие развитию этой области [7]:

– Опубликованные в интернете библиотечные данные слабо связаны между собой.

– Библиотечные стандарты создавались только для библиотечного сообщества.

– Библиотечные данные слабо структурированы и преимущественно хранятся на естественном языке.

– Библиотечное сообщество и сообщество Semantic Web имеют разные терминологии для аналогичных концепций.

– Библиотеки зависят от развития систем поставщиков, и часто не могут по собственной инициативе публиковать данные в LOD.

## 3.3 Форматы представления библиотечных данных

Библиографические записи, как правило, представляются и хранятся в библиотеках в одном из форматов семейства MARC.

Плюсы семейства форматов MARC:

– Достаточно детальное описание записи.

– Формат внедрен повсеместно и это упрощает обмен записями.

Минусы семейства форматов MARC:

– Запись можно хранить только полностью.

– Существует несколько форматов этого семейства плохо совместимых между собой.

– Для некоторых задач форматы семейства MARC избыточны, что влечет за собой избыточную бюрократию.

– Не может быть использован для представления данных в семантических базах данных.

Существует два основных формата семейства MARC:

– MARC21(используется, как правило, в США);

– UNIMARC (международный стандарт).

RUSMARC является диалектом UNIMARC.

Записи в формате семейства MARC могут быть представлены в виде XML(MARC/XML) или в бинарной форме(MARC/bin).

Кроме MARC существуют другие способы хранения данных, такие как представление данных согласно схемам Dublin Core или MODS.

Схема Dublin Core представляет из себя набор элементов данных для описания документов и других объектов в Интернете. Благодаря своей компактности и простоте схема стала широко распространена. При разработке Dublin Core не предполагалось, что новая схема полностью заменит MARC, т.к. Dublin Core не обеспечивает такую полноту, как MARC. Но для многих задач использование Dublin Core достаточно. Кроме того, существует онтология, описывающая термины

Dublin Core. Согласно этим терминам данные могут быть представлены в виде RDF.

Схема MODS (Metadata Object Description Standard) разработана Библиотекой Конгресса, является упрощенной версией MARC. Вместо трехзначных меток полей, абстрактных идентификаторов подполей используются понятные для пользователя вербальные метки (например, «title» вместо «245»). Часть элементов MARC игнорируется, введены новые элементы. MODS создана на основе MARC21 и более детальна по сравнению с Dublin Core. На основе MODS была создана онтология, используя термины которой, можно представлять библиографическую запись в виде RDF [8].

Таким образом, используя термины Dublin Core или MODS, можно хранить библиографические записи в виде RDF, используя семантическую базу данных.

### 3.4 Проекты интеграции библиотечных данных

На данный момент существует несколько проектов по интеграции данных разных библиотек, одними из крупнейших являются VIAF и Europeana.

VIAF (The Virtual Internation Authority File) – виртуальная система международных стандартов для авторитетной информации, совместный проект Библиотеки Конгресса, Немецкой Национальной Библиотеки и ряда других национальных библиотек и организаций. В состав проекта входит более 20 организаций, в том числе РНБ. В рамках этого проекта планируется интегрировать информацию об авторитетных файлах из крупнейших библиотек в мире.

Europeana – европейская цифровая библиотека, цель которой – обеспечить доступ к отсканированным страницам книг, отражающих различные аспекты европейской культуры. Сейчас доступна информация на французском, немецком и английском языках. В проекте участвуют Франция, Великобритания, Испания и Германия.

### 3.5 Выводы

Из описанного выше можно сделать вывод, что интеграция библиотечных данных достаточно актуальная задача, а использование технологии Semantic Web для этой задачи является перспективным. Существует несколько вариантов представления библиографических записей в виде RDF. Появляется все больше и больше успешных проектов в этой области.

## 4 Исследование и построение решения задачи

### 4.1 Публикация данных

#### 4.1.1 Описание предметной области

Для того чтобы преобразовать данные в RDF нужна схема, которая описывается с помощью RDFS и OWL. Эта схема называется онтологией. По концепции LOD онтология должна быть составлена на основе существующих в LOD онтологиях. Данные по библиографическим записям, которые предоставила РНБ, покрываются терминами Dublin Core, а информация об авторе терминами FOAF. БНБ представляет свои данные в Dublin Core и FOAF. Таким образом, имеет смысл использовать не MODS, а Dublin Core. Оно и было выбрано для дальнейшей работы.

#### 4.1.2 Конвертация данных

РНБ предоставляет свои данные в формате RUSMARC/bin. Общее количество записей около 17 тысяч. Необходимо преобразовать их в RDF. Эта задача состоит из двух подзадач:

– Преобразовать данные из бинарного формата RUSMARC/bin в RUSMARC/xml.

– Преобразовать данные из RUSMARC/xml в RDF с помощью онтологии.

Для решения первой подзадачи РНБ предоставила программу на C#, которую следует доработать.

Для решения второй подзадачи следует написать XSLT шаблон на основании онтологии.

#### 4.1.3 Создание семантического хранилища

Семантическое хранилище – это набор программных средств позволяющих хранить RDF данные и манипулировать ими с помощью SPARQL запросов.

Существует два вида хранилищ семантических данных:

– Хранилище, основанное на реляционной БД, при этом эффективно используется дисковое пространство и память, но получается низкая производительность.

– Хранилище, основанное на TDB, при этом достигается высокая производительность, но дисковое пространство и память значительно расходуются.

Было выбрано TDB хранилище, так как публикуемые данные имеют большой размер, производительность в нашем случае важнее.

Также существует несколько библиотек для работы с хранилищами.

– Jena;

– Sesame;

– Virtuoso.

Все эти библиотеки похожи друг на друга, но Jena имеет возможность осуществлять логический вывод OWL, а все остальные ограничиваются логическим выводом RDFS. Кроме того для Jena существует более эффективная реализация логического вывода OWL в библиотеке Pellet.

Таким образом, были выбраны Jena + Pellet.

### 4.1.4 Предоставление доступа к данным

Каждая библиографическая запись, представленная в RDF имеет свой URI. Согласно принципам LOD, при HTTP запросе по этому URI пользователь должен получать полную информацию об этой записи. Это касается и авторов. Для этого нужно создать web-сервер, который будет иметь доступ к семантическому хранилищу, используя библиотеку Jena, и доставать из него всю информацию по полученному URI. Этот сервер должен быть размещен по тому адресу, куда ссылаются записи. Для этого перед наполнением семантического хранилища нужно указать в качестве базового URI в файле RDF адрес сервера.

В качестве web-сервера был выбран сервер Jetty, написанный на Java, т.к. Jena написана на Java, а сервер Jetty встраивается в приложение, и для него ничего не надо дополнительно устанавливать.

## 4.2 Связывание

Для того чтобы связать данные РНБ и БНБ нужно получить связи, сохранить их в RDF и создать семантическое хранилище. Кроме того на web-сервере необходимо учитывать связи, иметь доступ к ним и возвращать их пользователю при обращении к сущностям, к которым эти связи относятся.

Самый тривиальный способ создания связей – сравнить каждый элемент с каждым по какому-то правилу и получить набор связей. У этого подхода есть два минуса при большом объеме данных:

– потенциальный набор связей $\frac{n(n-1)}{2}$;

– количество сравнений $\frac{n(n-1)}{2}$.

Количество записей предоставленных БНБ около 3,5 млн. В текущей ситуации решение, описанное выше, не эффективно.

Связи могут быть созданы с помощью кластеризации. Две записи будут считаться связанными, если попадают в один кластер. В этом случае, очевидно, количество связей, которые надо хранить заметно уменьшается.

В классической кластеризации количество сравнений $\frac{n(n-1)}{2}$.

В потоковой кластеризации не нужно сравнивать каждый элемент с каждым достаточно сравнить элемент с каждым из кластеров, в итоге количество сравнений получается $O(n)$, где $n$ – количество элементов [9]. Было решено воспользоваться именно потоковой кластеризацией.

Готовых библиотек найдено не было. В учебных целях было решено разработать и реализовать алгоритм самостоятельно.

Для получения кластеров записей РНБ и БНБ необходимо записи РНБ разбить на кластеры и по полученным кластерам распределить записи БНБ.

Основная идея алгоритма заключается в том, что запись и набор записей можно представить в виде вектора лексем, где для каждой лексемы задано некоторое число. Библиографическая запись – вектор лексем, где каждая лексема характеризуется числом, пропорционально зависимым от веса лексемы и частоты ее представления в записи. Кластер – вектор лексем, представленный суммой векторов, включенных в кластер записей.

При представлении записи в виде лексем не учитываются стоп-слова и окончания слов.

Кроме того, необходима функция кластеризации, сравнивающая вектор кластера и вектор записи, а также векторы записей между собой и векторы кластеров между собой.

### 4.2.1 Кластеризация записей РНБ

Первоначально имеется набор векторов записей. Последовательно проходя по ним, строится набор кластеров:

1) если кластеров нет, то вектор записи становится первым вектором кластера;

2) для всех кластеров определяем близость записи к ним, если запись не попала ни в один кластер, то создается новый кластер на основе записи, иначе запись добавляется во все релевантные кластеры.

Такая кластеризация зависима от порядка обхода записей. Чтобы уменьшить влияние, необходимо последовательно для каждой записи удалить ее из всех кластеров, затем снова проверить на соответствие всем существующим кластерам. Следует добавить запись в каждый релевантный кластер. Такую процедуру можно применять несколько раз, но, как показывает практика, 2-3 раз достаточно.

После описанных процедур может образоваться достаточно много кластеров, векторы которых почти равны векторам других кластеров. От таких дублей следует избавляться. Кластер признается дублем другого кластера, если в нем встречается 90% лексем другого кластера. Оба кластера удаляются из множества, вместо них добавляется кластер, содержащий записи, представленные в обоих кластерах.

В конечном наборе кластеров могут присутствовать пары кластеров, которые схожи относительно функции кластеризации. Такие кластеры удаляются из множества, а вместо них добавляется кластер, содержащий записи обоих кластеров.

### 4.2.2 Кластеризация записей БНБ

После получения кластеров записей РНБ следует распределить по ним записи БНБ. Этот процесс не зависит от порядка обработки записей БНБ. Кластеризация производится за один проход по

записям БНБ. Каждая запись БНБ сравнивается с кластером РНБ через функцию кластеризации. Если для записи существует хотя бы один релевантный кластер, она записывается в кластер с большей релевантностью.

## 5 Описание практической части

### 5.1 Подготовка данных

Была создана онтология, схема онтологии отражена на рисунке 1.

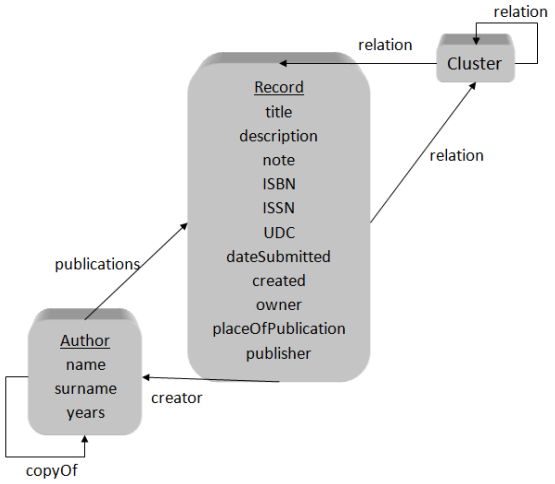



Рис. 1. Сконструированная онтология

Для описания авторов публикаций используется класс Author, который через предикат copyOf может ссылаться на свою копию. Множество объектов, связанных через copyOf, задают полное описание некоторого автора, которого они описывают по отдельности.

Для описания связей используется класс Cluster, если две записи связаны предикатом relation с одним объектом класса Cluster или с разными объектами, но связанными предикатом relation, то они считаются связанным.Для описания записи используется класс Record, имеющий множество предикатов для описания состояний своих объектов.

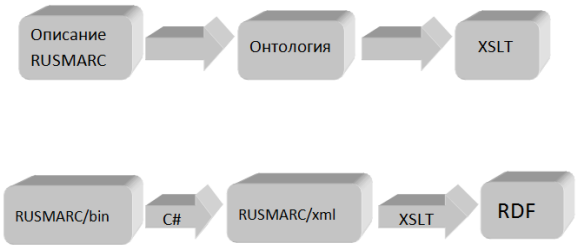



Рис. 2. Процесс получения RDF из данных РНБ

На рисунке 2 приведена схема получения XSLT преобразования из RUSMARC/xml в RDF и получения RDF из RUSMARC/bin.

Полученный RDF был вставлен в хранилище TDB, а поверх него была настроена SPARQL точка доступа fiseki, входящая в состав Jena, с логическим выводом OWL от библиотеки pellet. Логический вывод pellet оказался в десятки раз быстрее стандартного логического вывода OWL, реализованного в Jena, и смог обрабатывать большие объемы данных (300 000 троек). Стандартный вывод зациклился на таком объеме данных.

Был настроен сервер jetty для предоставления информации по HTTP. При получении запроса сервер обращается в хранилище данных РНБ и хранилище связей с БНР с помощью SPARQL запросов, получает нужную информацию и отправляет пользователю.

Согласно алгоритмам кластеризации, описанным выше, заголовки и описания записей РНБ были переведены на английский язык, разбиты на кластеры, и связи с БНБ были получены. Было создано хранилище TDB с SPARQL точкой доступа fiseki, был настроен логический вывод RDFS.

Для алгоритма кластеризации был проведен ряд экспериментов. Были скачаны группы новостей, распознанные системой «Yandex Новости», переведены на английский язык и кластеризованы.

Пусть *Drel* – множество связей распознанных яндексом.

Пусть *Dretr* – множество связей распознанных системой.

Тогда точность определяется формулой (1), а полнота – формулой (2):

$$\frac{|Drel \cap Dretr|}{|Dretr|}, \quad (1)$$

$$\frac{|Drel \cap Dretr|}{|Drel|}. \quad (2)$$

В результате эксперименты была получена точность равная 80% и полнота равная 60%. Эксперимент повторялся несколько раз с разными группами новостей, отклонение составило ±10%.

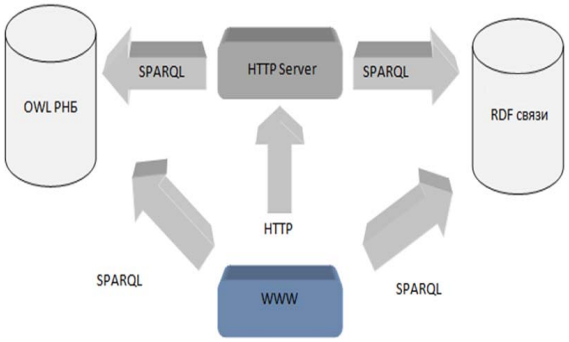



Рис. 3. Схема пользовательского приложения

Общая схема пользовательского приложения, позволяющего просматривать данные о библиографических записях РНБ и их связи между собой, авторами и библиографическими записями БНР, отображена на рисунке 3.

## 6 Заключение

В данной работе были разработаны программные решения, осуществляющие публикацию библиографических записей в пространство LOD и

интеграцию с библиографическими записями других источников. Описана онтология библиографических записей. Разработана процедура преобразования данных из формата RUSMARC/bin в RDF. Создано семантическое хранилище и SPARQL точка доступа. Настроен HTTP сервер для доступа к семантическим данным. Разработаны и реализованы алгоритмы потоковой кластеризации для получения связей записей РНБ и БНБ. Получены оценки качества алгоритма кластеризации.

Дальнейшие работы могут вестись по направлениям:

– полнотекстовый поиск по заголовкам и описаниям;

– создание распределенного хранилища;

– поиск по классификаторам UDC и BDC;

– поиск по issn и isbn.

## Литература


[1] Т. Бернерс-Ли, Д. Хендлер, О. Лассила. Семантическая сеть. http://ezolin.pisem.net/logic/semantic_web_rus.html

[2] Спецификация зыка RDF. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210

[3] Спецификация языка OWL. http://www.w3.org/TR/2012/REC-owl2-syntax-20121211

[4] T. Heath, C. Bizer. Linked Data: Evolving the Web into a Global Data Space. California : Morgan & Claypool, 2011. 136 c.

[5] T. Berners-Lee. Linked Data – Design Issues http://www.w3.org/DesignIssues/LinkedData.html

[6] ГОСТ 7.9–2003. Москва : Изд-во стандартов, 2004. 6 с.

[7] Library Linked Data Incubator Group Final Report http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025

[8] О.Н. Жлобинская. MARC-форматы в современной информационной среде. http://www.rusmarc.ru/publish/MARC_now.pdf

[9] П. Воляк. Проблемы кластеризации новостного потока. http://nlpseminar.ru/lecture50/


## Semantic Integration of Bibliographic Records


D. Malakhov, V. Serebryakov,
K. Teymurazov, O. Shorin



The paper deals with the problems of integration of bibliographic records in frame of the task of integration of data from the Russian National Library and the Britain National Library as a part of the Linked Open Data space. In course of solving the problem a prototype system has been constructed, through which the data format RUSMARC may be published in the Linked Open Data and linked to other library data. Experiments were carried out and quality estimates were obtained for for the subsystem linking data from different sources.

# Персональная цифровая библиотека Libmeta как среда интеграции связанных открытых данных

© О. М. Атаева © В. А. Серебряков
Вычислительный центр им. А.А. Дородницына РАН,
Москва
oli@ultimeta.ru serebr@ultimeta.ru

**Аннотация**

В статье описывается семантическая электронная библиотека Libmeta, ресурсы которой могут быть обогащены за счет использования данных из источников, расположенных в LOD. Связывание происходит посредством онтологии предметной области, которая задается пользователем и определяет его область интереса. Затрагиваются проблемы интеграции ресурсов библиотеки в LOD и создания поисковых запросов по источникам данных, а также обсуждается использование спецификаций и технологий из стека LOD в рамках одной системы.

* Работа выполнена при поддержке РФФИ – проект № 14-07-00058 А.

## 1 Введение

Последнее десятилетие наблюдается бурное развитие технологий Semantic Web и активное развитие сообщества, поддерживающего Linked Open Data (LOD). Основная идея LOD заключается в решении задач интеграции данных, представленных в сети, для чего предлагается представить информацию в формализованном виде, что делает ее доступной для машинной обработки.

Развитие технологий Semantic Web и популярность идеи LOD оказали влияние и на электронные библиотеки, которые трансформируются и превращаются в центры данных, вокруг которых формируется сообщество заинтересованных экспертов и пользователей, принимающих активное участие в их развитии. При консорциуме W3C была создана рабочая группа под названием Linked Library Data, которая выработала рекомендации по связыванию библиографических данных с использованием стандартных семантических технологий RDF, SPARQL, OWL. Появление семантических технологий вызывает необходимость разработки новых подходов к созданию электронных библиотек и расширяет возможности их использования.

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

## 2 Эволюция библиотек

Развитие информационных технологий в XX веке и их использование в библиотеках привело к появлению нового типа библиотек [16].

### 2.1 Электронные библиотеки

Электронные библиотеки возникли достаточно давно и представляют собой набор документоподобных ресурсов и их библиографии, в доступных для компьютеров форматах, а также сопутствующих услуг для их хранения и поиска. При этом в таких библиотеках не выделялись другие виды важных объектов, например, персоналии, организации и т.п. Встретив упоминание персоны в одном месте, невозможно точно установить соответствие с ее упоминанием в другом месте. Даже идентифицировав персону, как правило, нет возможности получить документы, связанные только с ней. Это обусловлено тем, что метаданные рассматривались как нечто, связанное только с документом.

### 2.2 Цифровые библиотеки

Цифровые библиотеки представляют собой информационные системы, которые обеспечивают задачи коллекционирования, хранения и навигации по разнообразным электронным документам, как хранящимся в самой системе, так и доступных по сети. Термин «цифровые библиотеки» часто рассматривается как синоним термина «электронные библиотеки», тогда как цифровые библиотеки являются продуктом следующего этапа развития электронных технологий и исследований в области электронных библиотек, использование результатов которых позволило расширить функциональность электронных библиотек, превратив их в «цифровые».

### 2.3 Семантические цифровые библиотеки

Использование семантических технологий значительно расширяет функциональность библиотек: данные лучше структурированы,

выделены связи между ними, улучшается поиск, появляется возможность интегрировать данные различных типов: персоны, ресурсы, пользователи. Обеспечивается интероперабельность с другими системами, не обязательно являющимися библиотеками, так как основной задачей семантических технологий остается предоставление метаданных в машиночитаемом формате. Онтологии играют основную роль для решения задач, вызванных структурными различиями существующих систем и семантическими различиями стандартов метаданных.

### 2.4 Персональные семантические цифровые библиотеки

Мы выделяем персональные семантические цифровые библиотеки, наполнение которых индивидуально для каждого пользователя системы и выполняется в полуавтоматическом режиме из разнородных источников данных, интегрированных в облако LOD. Будем далее для краткости называть их персональными открытыми цифровыми библиотеками или ПОЦБ. Типы информационных ресурсов и их структура определяются пользователем, исходя из своих интересов, то есть пользователь описывает интересующую его предметную область, определяя тематическое наполнение библиотеки.

Данная статья является продолжением нашей предыдущей работы [1], в которой представлена общая схема системы, выделены ее основные модули и дана характеристика каждого из них. Основное развитие системы произошло в направлении поиска источников связанных данных с использованием технологий из стека LOD. В следующих разделах привидено описание этого модуля и детализация его функций, а также кратко описаны первые практические результаты.

Основная задача системы заключается в предоставлении пользователю унифицированного представления для возможности автоматизированного извлечения интересующей его информации по определенной предметной области.

Представление ресурсов библиотеки в виде связанных данных расширяет функциональность семантических цифровых библиотек, давая возможность:

- включения дополнительных элементов описания данных информационных ресурсов,
- полного или частичного обновления данных из источников,
- использования интерфейсов для создания запросов к интегрированным в LOD источникам данных на основе SPARQL,
- включения в описания ресурсов других типов информации.

Одна из задач, которая решается в ПОЦБ, – это реализация интеграции набора данных в пространство LOD с использованием онтологии предметной области информационных ресурсов, т.е. автоматизированное обнаружение новых наборов данных и, по возможности, установка и поддержка связей с элементами данных из этих наборов данных с уже имеющимися ресурсами в репозитории библиотеки, обеспечивая одновременно рекомендуемую проектом LOD функциональность в рамках одной системы.

## 3 Источники данных

Мы подразделяем источники данных на два типа: внешние и внутренние. Внешними мы называем те источники, которые интегрированы в LOD, и данные которых представлены в RDF и доступны нам с использованием SPARQL. Для своих практических целей мы использовали такие известные источники в LOD, как DBpedia [3], Europeana [4]. Внутренние источники могут представлять собой любой другой тип источника данных, который не интегрирован в LOD. На практике в качестве внутренних источников мы использовали другие библиотеки, которые предоставляли доступ к своим данным по протоколу OAI-PMH.

### 3.1 Внешние источники

Данные из источников LOD хорошо структурированы и обычно доступны через SPARQL точку доступа для поисковых запросов. Так как одним из принципов LOD является использование URI, по которым можно получить по HTTP информацию в стандартном формате, то для доступа к информации определенного ресурса пользователь может использовать только этот URI.

Основной задачей подсистемы подключения внешних источников является создание и поддержка отображения онтологии предметной области на схему источника данных, посредством которого пользователь получит возможность автоматического мониторинга для последующего связывания имеющихся данных в системе с новыми данными по определенным запросам в терминах своей онтологии. При этом в системе при импорте может сохраняться лишь внешний URI ресурса.

### 3.2 Внутренние источники

Несмотря на активное развитие LOD, нельзя игнорировать источники данных, которые в него еще не интегрированы и при этом содержат огромный объем полезных данных. По этой причине в нашей системе реализован блок поддержки протокола OAI-PMH, который широко используется в библиотечной среде для обмена метаданными. Основным его недостатком, с точки зрения извлечения информации опираясь на принципы LOD, является то, что для доступа к информации о ресурсе нужно обладать специальными знаниями о протоколе, при этом знание идентификатора OAI, который используется в таком источнике для представления информации о ресурсе, не сильно облегчает поиск этих данных. Например,

OAI идентификатор ресурса *42041024*, на портале «Научное наследие Росии», для пользователя не обладающего специальными знаниями не позволит найти полезной информации, тогда как идентификатор из источника LOD http://dbpedia.org/page/Mikhail_Lomonosov интуитивно понятен и позволяет получить доступ к полезной информации о ресурсе, а также к связанным с ним ресурсам. Таким образом поддерживая этот протокол, мы внутри нашей системы решаем задачу формального предоставления и интеграции этих данных в соответствии с принципами LOD, при этом сохраняя информацию о первоначальном источнике, одновременно позволяя решать задачи связывания данных с другими источниками из облака LOD в рамках системы.

В работе [2] предлагается улучшенная версия этого протокола, которая является развитием протокола в сторону поддержки связанных данных.

## 4 Функциональность ПОЦБ

К основной функциональности системы, реализующей ПОЦБ относятся:

- функции атрибутного поиска;
- функция выделения неявных связей между ресурсами по их описаниям;
- функция работы с коллекциями;
- создание/просмотр/редактирование/объединение/вложенные коллекции;
- функция отображения онтологии ИД;
- функция детализации, которая обеспечивает преобразование в подзапросы, соответствующих различным ИД;
- функция для выполнения запросов и обработки результатов и предоставления окончательного результата пользователю;
- функция автоматического мониторинга ИД на наличие новых/измененных данных;
- создание словарей, классификаторов, тезаурусов;
- редактирование элементов;
- поддержка («гибкой») классификации ресурсов;
- поддержка настройки уровней доступа к различным ветвям тезауруса.

Исходя из определения источников данных ПОЦБ и перечня функций системы, можно выделить «внутренние» функции, т.е. те, которые оперируют данными в рамках системы и интегрируют данные из «внутренних» источников и фактически определяют обычную семантическую библиотеку. «Внешние» функции обеспечивают подключение и извлечение данных из LOD и позволяют задать тематическое наполнение библиотеки и установить связи, таким образом задавая фактически определение ПОЦБ.

## 5 Онтология ПОЦБ

Онтология ПОЦБ разработана в общем виде без привязки к конкретным методам и способам реализации семантических цифровых библиотек [1], [5].

Фактически общая онтология ПОЦБ состоит из двух онтологий:

1) онтология СЭБ, построенная на основе онтологии информационных систем, включающая в себя основные понятия, необходимые для обеспечения основной функциональности библиотеки, такие как ресурс, пользователь, коллекция, словарь, классификатор, запрос, источник и т.д.

2) онтология и тезаурус предметной области, для которой пользователь определяет ее понятия, их тип, структуру, совокупность словарей и классификаторов, которые представляют тезаурус предметной области, который обеспечивает доступ неквалифицированных пользователей, решающих задачи поиска информации, к знаниям предметной области в разных источниках. Эта онтология позволяет:

- выработать и зафиксировать общее понимание области знания;
- представить знания в удобном для обработки автоматизированными подсистемами виде, обеспечить возможность получения и накопления новых знаний, а также представить возможность многократного использования знаний

Тезаурус же обеспечивает терминологическую поддержку и помогает пользователям сформулировать запрос к системе, в том числе, подобрать правильные ключевые слова для описания искомого результата, имеющихся данных и контекстной информации.

Тезаурус необходим для навигации и для автоматического уточнения и расширения запроса, введенного пользователем, посредством использования зафиксированных в тезаурусе связей между терминами. Например, в частном случае, в качестве предметной области рассматривается онтология из работы [6] со всем набором словарей и классификаторов. Данные, представленные этой онтологией, представляют собой численные значения теплофизических свойств для различных веществ в разных условиях и их библиографии.

Основным классом, поддерживаемым в онтологии СЭБ, является класс *информационный ресурс*, подклассами которого являются такие классы ресурсов как *публикация*, *персона* и т.д. Подключаемые классы предметной онтологии могут являться как подклассами класса *информационный ресурс*, так и расширять структуру подклассов этого класса. Таким образом онтология предметной области одновременно может расширять список информационных ресурсов системы, а также дополнять и расширять структуру информационных

ресурсов. Для поддержки такой интеграции онтологии реализован отдельный модуль поддержки различных типов связей определен минимальный словарь этих связей. Такой подход к созданию онтологии системы позволяет конкретизировать область интересов в рамках конкретной персональной библиотеки.

## 6 Поиск по источникам данных

Поисковые системы, ориентированные на источники, интегрированные в LOD, такие как Sig.ma, Falcons, и SWSE, обеспечивают поиск на основе ключевых слов, ориентированный на использование той же парадигмы, что и существующие лидеры рынка, такие как Google и Yahoo. Пользователю предоставляется окно поиска, в котором он может ввести ключевые слова, связанные с предметом или темой, в которых он заинтересован, и приложение возвращает список результатов, которые могут (или нет) иметь отношение к запросу. Фактически это поиск по вхождению слова в любой элемент описания. Поиск же данных в источниках предполагает, что пользователь знает структуру данных

В работе [8] представлена система поиска LOQUS в репозиториях LOD на основе высокоуровневой онтологии, на которую отображается схема подключаемого источника данных (ИД). Эта онтология составлена на основе высокоуровневой онтологии, которая содержит наиболее общие и самые абстрактные концепты, имеет исчерпывающую иерархию фундаментальных понятий (около 1 тыс.), а также набор аксиом (примерно 4 тыс.), определяющих эти понятия. Каждому концепту определен идентификатор или обобщающее понятие из LOD. Онтология так же, как и в нашем подходе, используется для трансляции SPARQL запросов пользователей в интегрированные ИД. Но недостаточный уровень концептуализации понятий не позволяет в достаточной мере сконцентрироваться на определенной предметной области.

С другой стороны задача автоматизированного поиска релевантных источников данных осложняется тем, что чаще всего информация о связях между ними проставляется в основном на уровне данных с помощью связей sameAs, seeAlso. Даже простой анализ связей sameAs, seeAlso на уровне найденных данных позволит выявить эквивалентные классы, ранее не определенные связи между разными источниками или новые источники. Описание связей на уровне схем затем можно использовать при формировании запросов к источникам данных.

До недавнего времени связи между источниками на уровне схем описывались гораздо реже. В последние несколько лет эта задача решается с введением и активным распространением спецификации VOID [7] для описания источников RDF данных, в которой предоставляется информация о связанных источниках данных. VOID описание содержит информацию об используемых словарях, статистическую информацию о том, сколько ресурсов того или иного типа или значений определенных свойств используются во множестве. При создании словаря VoID была сведена к минимуму необходимость создания новых свойств и классов, путем использования существующих словарей. Например, для описания статистической информации используется словарь SCOVO. На основе этой информации можно делать вывод о релевантности источника тому или иному запросу или предметной области.

В рассматриваемой системе VoID описание набора данных в хранилище генерируется с помощью D2R Server [15]. В сгенерированное описание не попадает информация о подключенных источниках данных и статистика по имеющимся с ними связям. Для включения этой информации были использованы правила, по которым осуществляется поиск связанных данных [12]. Полученное описание в рамках используемой системы позволяет формировать распределенные запросы к подключенным источникам данных в терминах онтологии, используемой в этой системе. Используя VoID описание, запросы из системы транслируются в термины уже источников данных. Также это описание применяется для отображения обобщенного результата поиска.

## 7 Общая схема подключения источников данных

На рисунке 1 представлена общая схема подключения различных источников данных с использованием технологий из стека проекта LOD

Доступ к данным Libmeta осуществляется через ее общую онтологию, которая, как было сказано, состоит из: а) онтологии семантической библиотеки, б) онтологии предметной области, которая задает тематическое направление информационных ресурсов. При этом D2R Server [15] использует онтологию Libmeta для создания SPARQL точки доступа к ее данным. Используются правила, которые задаются для каждого подключаемого источника (правил может быть несколько), с помощью которых осуществляется поиск и сохранение связей между данными Libmeta и источником из LOD. Для задания правил связывания используется фреймворк SILK. Правила описываются в соответсвии с требованиями SILK и хранятся в определенном для каждого источника месте. После описания правила и указания его расположения все действия по запуску и анализу результатов работы SILK выполяются программно, ля этого используется соответсвующая задаче версия фреймворка.

При каждом подключении нового источника или обновлении набора связей уже подключенных нужно обновлять VoID описание множества данных Libmeta, анализируя полученный набор ссылок и

правила, по которым они выполнялись. Это позволит обновить статистическую и структурную части VoID, необходимых для использования при формировании запросов в терминах общей онтологии и их преобразования в запросы к релевантным источникам в соответствующим им терминах.

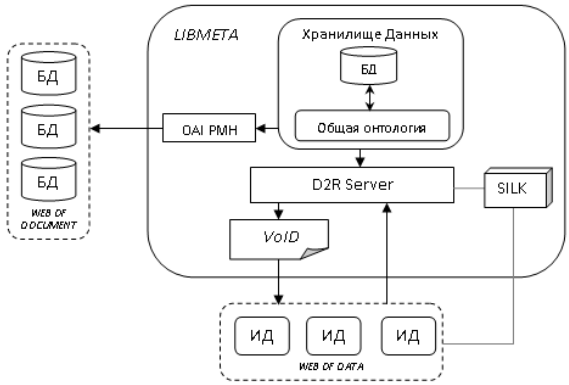



Рисунок 1

Libmeta также исторически поддерживает обмен данными по протоколу OAI-PMH с библиотеками, неинтегрированными в LOD, выступая агрегатором, который интегрирует их данные в LOD.

## 8 Текущее состояние работ

В рамках создания первой версии ПОЦБ был реализован проект по созданию стандартизированной и децентрализованной среды управления информацией электронных фондов Libmeta [10]. В проекте реализованы средства интеграции приложений с разными источниками/каталогами метаданных/данных, сервис директорий метаданных, унифицированный интерфейс поиска данных.

Существенное различие во внутренних моделях данных, используемых в различных музеях, библиотеках и архивах, является главной проблемой на пути решения задачи интеграции данных [9]. Для преодоления этой проблемы в решаемой задаче интеграции данных было предложено участникам экспортировать метаданные из своего внутреннего формата в формат на базе Dublin Core с использованием синтаксиса XML, так как во внутренних используемых форматах удается выделить общую часть, которая ложится в рамки предложенного формата. В системе используется универсальный модуль загрузки метаданных в произвольном XML-формате в соответствии с протоколом OAI-PMH.

Основная коллекция метаданных была получена из библиотеки (тип источника внутренний) «Научное Наследие России» [10]. Для интеграции данных в LOD в качестве внешних источников было проведено связывание с данными DBpedia по авторам, а для связывания музейных экспонатов был проведен эксперимент с данными из Europeana.

Для каждого ресурса Libmeta может быть получено его представление, удовлетворяющее модели Europeana Semantic Elements (ESE) [14], которое определяет ряд обязательных элементов метаданных.

Для мониторинга новых данных и установления связей с внешними источниками данных в рамках системы используется SILK Framework [12]. Для установления связей необходимо указать источник данных, правила доступа к данным и правила связывания. Вся эта информация была написана в виде конфигурационного файла на языке SILK LSL.

Сейчас проводятся работы по связыванию данных с авторитетными файлами VIAF [13]. Это проект, который объединяет все значимые библиотеки, интегрирующие свои данные в LOD.

## 9 Заключение и дальнейшие работы

Разрабатываемая ПОЦБ предполагает поддержку функциональности, рекомендуемую проектом LOD, а именно: средства для представления информации из различных источников как для установления, так и для поддержки связей между RDF-ресурсами, как внутренними, так и внешними, т.е. предполагает осуществление полного цикла интеграции набора данных в пространство LOD.

Основные преимущества реализации принципов LOD в Libmeta:

- Связность. Подключение источников, не обязательно библиотек;
- Машиночитаемость.Представление в RDF, использование общепринятых словарей и онтологий;
- Доступность. Доступные для свободного использования всеми пользователями без каких-либо ограничений в виде авторских прав.

Использование онтологии предметной области позволит не только включать другие типы ресурсов в библиотеку, но и уточнять и включать в библиотеку описания внутренней структуры информационных ресурсов нужной детализации, обращаясь за данными к источникам, которые раньше с трудом могли использоваться в рамках интеграции ресурсов электронных библиотек.

## Литература


[1] О. М. Атаева, В. А. Серебряков, Подход к созданию персональной электронной семантической библиотеки, RCDL, 2013.

[2] Bernhard Haslhofer, Bernhard Schan, The OAI2LOD Server: Exposing OAI-PMH Metadata as Linked Data, 2008. http://eprints.cs.univie.ac.at/284/1/lodws2008.pdf

[3] http://dbpedia.org

[4] http://europeana.eu

[5] R. Weber. Ontological Foundations of Information Systems, Queensland, Australia, Coopers & Lybrand. 1997.

[6] О. М. Атаева, А. О. Еркимбаев, В. Ю. Зицерман, Г. А. Кобзев, К. П. Пушин, В. А. Серебряков, К. Б. Теймуразов. Интеграция данных по теплофизическим свойствам веществ методами онтологического моделирования, RCDL, 2013.

[7] http://www.w3.org/TR/void/

[8] P. Jain, K. Verma, P.Z. Yeh, P. Hitzler, A.P. Sheth. LOQUS: Linked Open Data SPARQL Querying System. Technical report, Tech. rep., Kno. e. sis Center, Wright State University, Dayton, Ohio, 2010. Available from http://www.pascal-hitzler.de/resources/publications/loqus-tr-2010.pdf

[9] А.Б. Антопольский, А.А. Каленкова, Н. Каленов, В.А. Серебряков, А. Сотников. Принципы разработки интегрированной системы для научных библиотек, архивов и музеев // Информационные ресурсы России. – 2012. – № 1. – С. 2–7.

[10] А. Антопольский, О. Атаева, В. Серебряков. Среда интеграции данных научных библиотек, архивов и музеев «LibMeta» // Информационные ресурсы России. – 2012. – № 5. – С. 8–12.

[11] http://e-heritage.ru/index.html

[12] http://lod2.eu/Project/Silk.html

[13] http://viaf.org/

[14] http://pro.europeana.eu/ese-documentation/

[15] http://d2rq.org/d2r-server

[16] Е. Горный. Развитие электронных библиотек: мировой и российский опыт, проблемы, перспективы / Е. Горный, К. Вигурский // Интернет и российское общество / под ред. И. Семенова; Моск. Центр Карнеги. – М. : Гендальф, 2002. – С.158–188.

## Personal Digital Library Libmeta as an Integrating Environment for Linked Open Data

Olga M. Ataeva, Vladimir A. Serebryakov

The article describes semantic digital library Libmeta resources of which can be enriched by means of using data from the sources located in LOD. Binding is due to domain ontology which is user defined and determines his/her field of interest. Problems of integration of library resources in LOD and creation of search queries on data sources are considered as well as use of specifications and technologies from LOD stack within a system considered.

# Модель семантического управления личной информацией

© А. А. Бездушный © А. Н. Бездушный © В. А. Серебряков

МФТИ ВЦ им. А.А. Дородницына РАН
andrey.bezdushny@gmail.com anb@ccas.ru serebr@ccas.ru

**Аннотация**

Целью данной работы является рассмотрение основных подходов к управлению информацией и знаниями, а также прототипирование системы предоставляющей человеку возможность организации личного информационного пространства в соответствии со стандартами Semantic Web и инициативой Linked Open Data. Структурированное представление данных позволяет проводить автоматизированный анализ сведений, с которыми ежедневно сталкивается человек, а использование стандартов Semantic Web предоставляет гибкие возможности для интеграции с репозиториями Linked Open Data. Предлагаемая методика развивает идею подхода Semantic Desktop, введенного Leo Sauermann, способа организации данных на персональном компьютере, в котором любой объект на компьютере – файл, e-mail или событие календаря, рассматривается как RDF ресурс (объект с уникальным идентификатором – URI).

## 1 Введение

Решением вопросов эффективной организации и работы с информацией и знаниями, занимаются системы управления личной информацией (Personal Information Management Systems). Одним из первых свое видение подобной системы, в 1945 г., высказал Вэнивар Буш в эссе «Как мы можем мыслить» [2]. В нем Буш описывает устройство под названием Мемекс (Memex), в котором люди могут хранить всю свою личную информацию – мысли, записи, книги, и которое может выдавать нужную информацию с достаточной скоростью и гибкостью. В основу работы Мемекс Буш закладывал механизмы ассоциативных ссылок и примечаний. Устройство, по его задумке, должно было точно имитировать ассоциативные процессы человеческого мышления, исключая присущие человеку недостатки, такие как забывание информации. Одной из технологий, необходимых для реализации своего устройства, Буш считал возможность организации хранилища, содержащего практически неограниченное количество информации, такое что «даже если бы пользователь вставлял в него по 5000 страниц сведений в день, ему бы потребовались сотни лет чтобы заполнить свое хранилище». Сейчас стоимость жестких дисков мала настолько, что человек может хранить всю изученную им информацию в течение неограниченного количества времени, при необходимости просто увеличивая объем хранилища, путем добавления нового жесткого диска. Таким образом, на пути создания Мемекса, в настоящее время лежит лишь проблема проектирования гибкой системы, способной помогать человеку при выполнении ежедневных задач, дополняя и структурируя его мыслительные процессы.

## 2 Управление информацией

С каждым годом количество информации, с которой ежедневно сталкивается человек, растет, и все больший ее объем переходит в электронный формат – публикуется в сети или сохраняется на персональных компьютерах. Эта информация распределяется между различными источниками, в которых хранится в разнородных форматах. Часть сведений может храниться в виде документов, другая – в виде ссылок или заметок, третья – в контексте не связанных между собой информационных систем. Такая организация данных приводит к *фрагментации информации* – в рамках различных источников, человеку приходится поддерживать различные, зачастую не связанные между собой, но обладающие общей структурой, организационные схемы. Разнородность форматов хранения данных затрудняет процесс задания зависимостей между этими схемами, в результате, сведения о взаимосвязях фиксируются только в памяти человека. Со временем, эти сведения неизбежно забываются, затрудняя процесс воссоздания контекста работы и поиска ресурсов, работа с которыми велась ранее.

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

Системы управления личной информацией автоматизируют процессы ведения и работы с *информационным пространством* – совокупностью всех сведений, с которыми человек работает сейчас или работал ранее. Рассмотрим основные функции, которые, с нашей точки зрения, должны выполнять системы управления личной информацией:

1. Ведение информационного пространства и структуризация *информационных ресурсов*, находящихся в нем. Под информационными ресурсами, будем понимать любые данные, имеющие важность для человека, и выделяемые им в отдельную сущность – это могут быть файлы, заметки, посещенные веб страницы, письма и пр. Системой управления личной информацией, должны обеспечиваться процессы, способствующие формализации информационных ресурсов, а также операций, выполняемых над ними.

2. Поиск по информационному пространству. Часто человек сталкивается с задачей повторного поиска информации, с которой он работал ранее. В таком случае, при поиске, он обычно обладает большим количеством косвенной информации, касающейся искомого ресурса. Эта информация чаще имеет ассоциативный формат, например, с какими еще ресурсами велась работа одновременно с искомым, в какой период времени это выполнялось, какая последовательность действий была совершена при первичном нахождении ресурса. Системы управления личной информацией должны поддерживать ведение такого рода метаданных и предоставлять возможность поиска по ним.

3. Автоматический анализ информационного пространства. Поскольку информационные ресурсы хранятся в структурированном виде, становится возможным проведение их автоматической обработки и анализа. Можно выделить несколько аспектов анализа. Автоматическое пополнение метаданных об информационных ресурсах, сведениями, найденными в сети. Другим аспектом является поиск схожих или связанных ресурсов, как в рамках информационного пространства пользователя, так во внешней сети.

4. Категоризация ресурсов информационного пространства. Часть ресурсов может быть категоризирована системой автоматически, например, научные публикации, музыка, видео и прочие ресурсы, категории для которых определены в сети. Другая часть – с участием человека, в таком случае частичное распределение по категориям производится пользователем, а системой, на основании этого распределения, предлагаются категории для новых ресурсов.

5. Возможность совместной работы. При ведении общего проекта, люди часто сталкиваются с необходимостью совместной работы над частью информации. Для такого случая, системы управления личной информацией должны поддерживать обмен информационными ресурсами, их метаданными или частичное объединение информационных пространств разных пользователей системы, а также предоставлять возможности для комментирования и обсуждения информационных ресурсов.

Таким образом, систему управления личной информацией можно рассматривать, как своеобразного интеллектуального цифрового помощника, сопровождающего и помогающего пользователю вести его информационное пространство. Дополняя сырые данные структурой и семантикой, пользователь получает возможность автоматизации, выполняемых им, интеллектуальных процессов.

## 3 Существующие подходы к управлению информацией

Несмотря на существование большого количество работ в области управления личной информацией, в настоящее время распространены лишь так называемые «персональные органайзеры», рассматривающие самые простые задачи, такие как планирование событий, установка напоминаний, ведение заметок и контактов, работа с электронной почтой. Наиболее популярными примерами таких органайзеров являются Microsoft Outlook, Mozilla Thunderbird и др. Предоставляя возможности для хранения ограниченных типов информации, эти средства, тем не менее, опускают вопросы управления накопленными сведениями – формирования взаимосвязей между данными, организации составных структур, совместной работы.

Рассмотрим основные направления работ, опубликованных в последние годы, по теме управления личной информацией. В работе «The Gnowsis Semantic Desktop for Information Integration» [16], описывается концепция *Semantic Desktop* - подхода к организации данных на персональном компьютере, в соответствии с которым любая информация, используемая пользователем, - файл, e-mail или событие календаря, рассматривается как RDF ресурс с собственным уникальным идентификатором. В этой работе вводится модель личной информации (Personal Information Model - PIMO) – формализующая ментальную модель информационного пространства, составленную пользователем. Основная задача PIMO – предоставить общую модель данных, с которой смогут работать различные приложения, используемые пользователем. Единое представление данных предоставит больше возможностей для организации более гибкой интеграции между приложений. К недостаткам Gnowsis можно отнести то, что основной акцент делается на организации модели данных и ее совместном использовании различными приложениями, в то время как вопросы управления накопленными данными почти не рассматриваются. На этих вопросах концентрируются работы SemEx

[7], IRIS [5], Haystack [13], MyLifeBits [9], DeepaMehta [15]. В SemEx, IRIS и Haystack данные представляются в иерархическом виде, в основе иерархии лежат наиболее распространенные типы данных, такие как email, контакты, проекты. В IRIS и Haystack для каждого типа данных определен набор интерфейсов, предоставляющих базовые операции, такие как возможность ответа или пересылки письма, создания события или напоминания. В Haystack, дополнительно предоставляется возможность настройки стандартных и создания собственных визуальных представлений данных, а также определяются программные интерфейсы для создания дополнительных операций над данными. В работах MyLifeBits и DeepaMehta, рассматриваются альтернативы к иерархическому подходу представления данных. В MyLifeBits, предлагаются интерфейсы отображения ресурсов с использованием временной шкалы. По мнению авторов, введение временной шкалы позволяет более наглядно отобразить ресурсы, а также увеличить количество одновременно выводимых ресурсов. В работе DeepaMehta, данные предоставляются пользователю в форме тематической карты (topic map) – ориентированного графа, узлами которого являются ресурсы, определенные пользователем.

В работе Beagle++ [6] подробно рассматриваются вопросы ранжирования ресурсов, полученных в результате поискового запроса. Ранжирование производится на основании объединения результатов, полученных с помощью алгоритмов ObjectRank и TF/IDF.

Работы iMecho [4] и Fledspar [3], рассматривают вопрос использования ассоциаций при поиске ресурсов. В iMecho, предлагается формировать журнал работы пользователя с ресурсами, который в дальнейшем анализировать для выделения зависимостей между ресурсами. В Fledspar, предоставляется удобный интерфейс для ассоциативной навигации по ресурсам, а также реализуется возможность осуществлять поиск ресурсов на основании информации связанной с ними. В работе Desktop Gateway [12], помимо интеграции между приложениями, также рассматриваются вопросы использования данных полученных из сети.

Перечисленные работы делают больший упор на формирование информационного пространства и работу с ним, в меньшей мере затрагивая вопросы анализа данных и автоматизации действий, выполняемых пользователем. В них слабо освещены вопросы объединения информационных пространств различных пользователей и совместной работы с ними.

Среди коммерческих систем можно выделить Google Now [10] и Dropbox Datastore [8]. Основной задачей Google Now является отображение нужной информации в нужный момент. Основываясь на истории совершенных ранее действий, а также на текущем местоположении и моменте времени, Google Now предоставляет пользователю релевантную информацию, такую как прогноз погоды, пробки и др. Dropbox Datastore обеспечивает возможность хранения структурированной информации в «облаке» Dropbox. Основной структурой в Dropbox Datastore являются таблицы, для чтения и записи в них предоставляется программные интерфейсы.

## 4 Предлагаемое решение

В данном разделе приводится описание архитектуры предлагаемого решения, а также проводится сравнение двух схем, соответствующих разным моделям работы пользователей с данными, одна из которых не использует систему управления информацией (рис. 1), а другая – использует (рис. 2). На схемах, рассматривается работа двух пользователей, чьи информационные пространства частично пересекаются. В процессе работы каждый пользователь пополняет собственное информационное пространство, состоящее из разнородных данных – документов, событий, писем. Работа с ресурсами ведется по средствам различных приложений и информационных систем.

Без использования системы, информационные пространства пользователей нигде формально не определены, только сам пользователь может определить, какие данные связаны между собой и как именно. Вследствие этого, все дальнейшие взаимодействия с данными – поиск, совместная работа, формирование иерархической структуры, могут производиться лишь в рамках того приложения, которое отвечает за конкретный тип ресурсов. Поскольку информация делится между различными приложениями, большое количество метаданных о ресурсах, таких как иерархическая организация, связи и зависимости между ресурсами, могут дублироваться в каждом из них. Системы управления личной информацией вводят дополнительный уровень организации данных, позволяют пользователю явно определить своей информационное пространство и предоставляют интерфейсы для работы с ним. Сведения хранятся в системе в соответствии с форматами, определенными в OWL онтологии, за счет чего, к слабоструктурированным данным добавляется семантика, а также появляется возможность производить их автоматический анализ, категоризацию и индексацию. Также важным моментом является то, что пользователи могут осуществлять работу с данными с помощью привычных для них приложений, т.к. по средствам адаптеров и агентов информация из внешних источников может быть автоматически выгружена в систему. Дополнительно, поскольку система по своей сути является многопользовательским приложением, в рамках нее возможна совместная работа различных пользователей с общим информационным пространством.

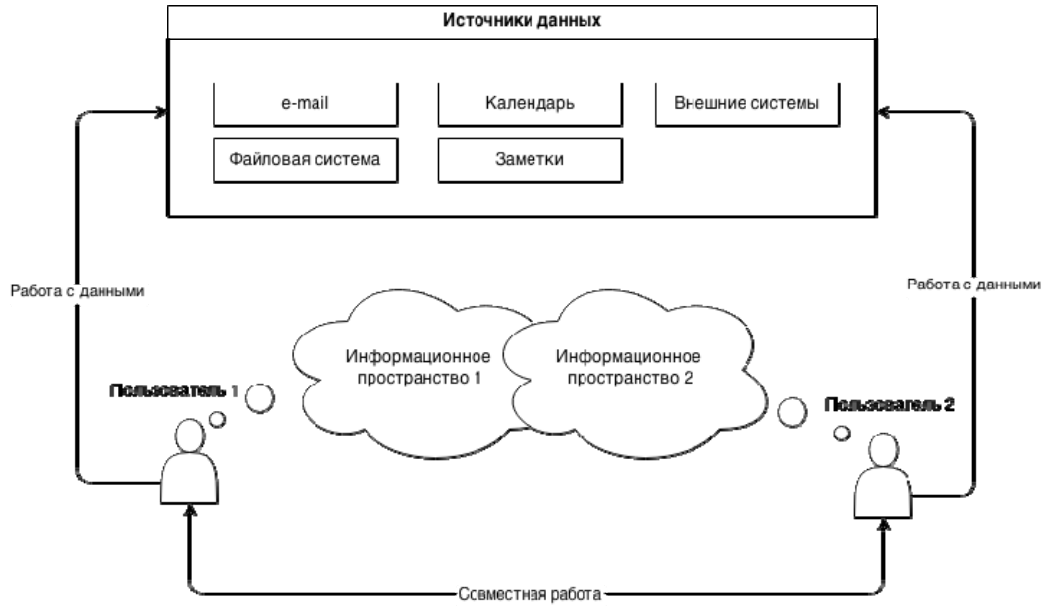



Рис. 1. Работа пользователей без использования системы

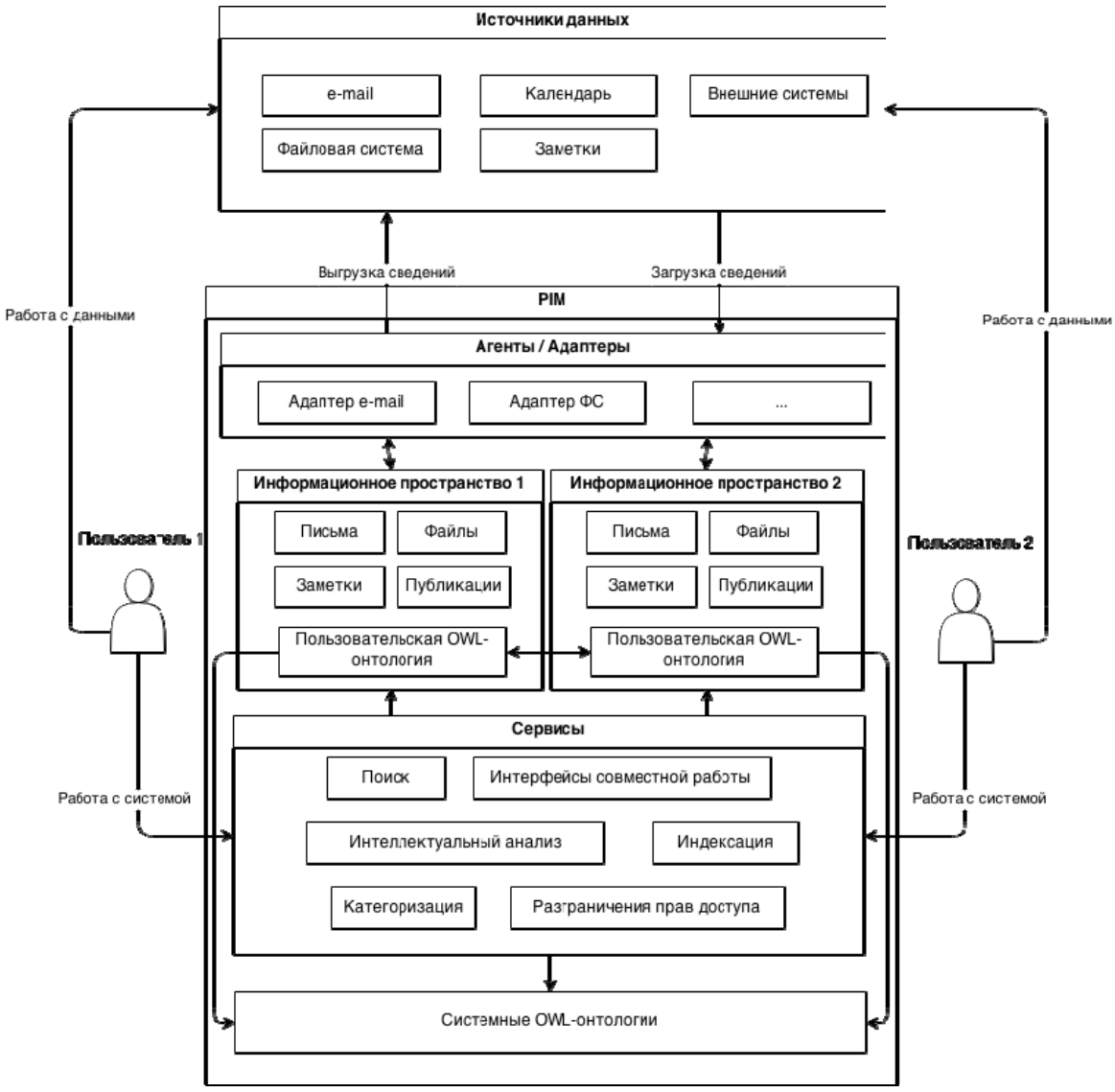



Рис. 2. Работа пользователей с использованием системы

## 5 Информационное пространство пользователя

Информационное пространство – это совокупность всех сведений, с которыми человек работает сейчас или работал ранее. Любые данные, имеющие важность для человека, и выделяемые им в отдельную сущность рассматриваются как элементы информационного пространства – информационные ресурсы. Информационными ресурсами могут быть файлы, заметки, посещенные веб страницы, письма и пр.

В рамках системы, в качестве информационных ресурсов, рассматриваются RDF ресурсы. Таким образом, каждый пользователь в рамках системы ведет работу с собственным RDF репозиторием данных, представляющим его информационное пространство. Структура информационного пространства описывается с помощью OWL онтологии пользователя. По умолчанию, предоставляется системная онтология, которую, при необходимости, пользователь может изменять, добавляя новые классы и свойства, а также расширяя уже имеющиеся. Наполнение репозитория происходит либо по средствам автоматического импорта информации, либо вручную, через пользовательский интерфейс системы.

## 6 Категоризация

Деление файлов на различные иерархические категории, является одним из основополагающих процессов, используемых пользователями при работе с информацией на персональных компьютерах. Тем не менее, исследования [18] показывают, что, несмотря на понимание того, что категоризация в дальнейшем может существенно облегчить поиск, многие пользователи, по различным причинам, игнорируют эту возможность. В качестве объяснения такого поведения пользователи обычно ссылаются на сложности при принятии решения, в какую из категорий отнести файл, проблемы при формировании подкатегорий, таких, чтобы их содержимое не пересекалось, а также на нехватку времени. Поэтому возможность автоматической или полуавтоматической категоризации сведений попадающих в систему, является крайне важной.

Необходимым элементом, для проведения категоризации, является выбор категорий, на которые будут делиться ресурсы. В ряде случаев возможно выбрать их полностью автоматически – это относится к ресурсам, набор категорий для которых, может быть получен из внешних источников. Например, научные статьи делятся на категории на основании тематики работы, музыка и фильмы на основании жанров и направлений.

Поскольку в общем случае, автоматическое выделение категорий не возможно, формировать требуемые классы можно по мере работы пользователя с системой. В таком случае категоризация может проводиться по двум направлениям:

- выделение в общие категории ресурсов, находящихся в общих разделах иерархической структуры, организованной пользователем;
- выделение в общие категории на основании добавленных пользователем метаданных, таких как «теги».

За счет такой категоризации, при добавлении нового ресурса в систему, на основании уже внесенных пользователем сведений, пользователю будет предложено возможное расположение нового ресурса в иерархической структуре, а также метаданные, которые могут быть к нему добавлены.

## 7 Интеллектуальный анализ данных

Одной из наиболее важных функций в системах управления личной информацией, является возможность хранить метаданные о созданных в системе ресурсах. Хорошо известно, что обычно, люди забывают или затрудняются заносить метаданные вручную, поэтому важно, чтобы система могла сформировать максимальное количество метаданных в автоматическом режиме. Как было описано выше, часть метаданных формируется адаптерами к источникам данных, на основании содержимого импортируемого ресурса. Кроме того, в ряде случаев, можно получить дополнительную информацию из глобальной сети, для этого системой предоставляется ряд *адаптеров к внешним репозиторям*. Адаптеры осуществляют поиск «аналогов», для имеющихся в системе ресурсов, в различных репозиториях глобальной сети (в частности, в репозиториях Linked Open Data), и, в случае успеха, переносят соответствующую информацию из найденных ресурсов в репозиторий пользователя. Каждый адаптер отвечает за поиск «аналогов» одного или нескольких классов OWL-онтологии пользователя. Другим аспектом анализа данных является поиск похожих или связанных ресурсов. Для каждого ресурса, находящегося в информационном пространстве пользователя, осуществляется поиск связанных с ним ресурсов, как внутри информационного пространства, так и вне его – в глобальной сети. Алгоритмы поиска схожих ресурсов могут сильно отличаться в зависимости от класса искомого ресурса. Поэтому, за поиск схожих ресурсов для разных классов отвечают различные компоненты.

## 8 Реализация

В рамках данной статьи реализован прототип системы, поддерживающий хранение и анализ научных публикаций. Прототип соответствует описанной выше архитектуре. Уровень адаптеров данных представляет адаптер файловой системы. На уровне сервисов реализован модуль анализа публикаций, осуществляющий пополнение репозитория данными из Академии Google (Google Scholar), а также выполняющий поиск новых публикаций, схожих с загруженными ранее.

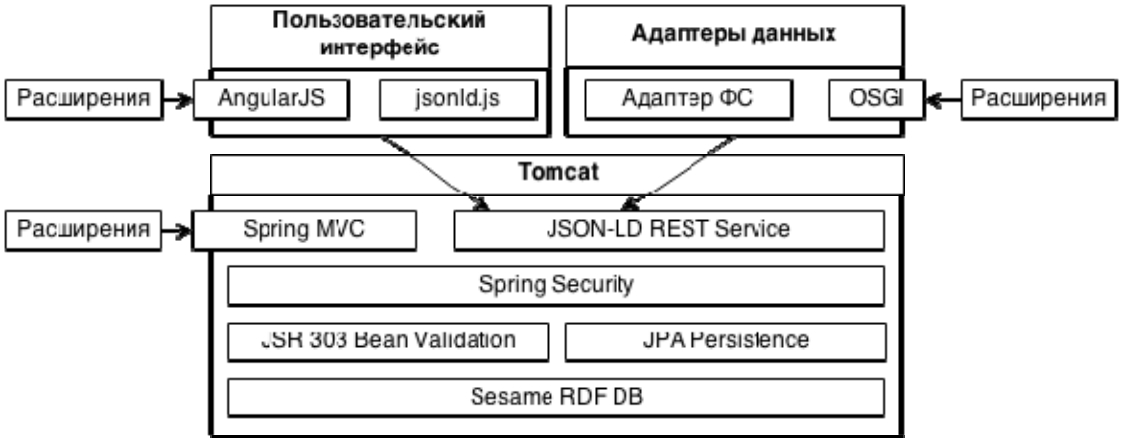



Рис. 3. Архитектура системы

На рис. 3 представлена модель реализованного приложения. Большой упор при реализации делался на расширяемость системы. Серверная часть системы выполнена на языке Java и представляет собой веб приложение, реализованное с использованием библиотеки Spring Framework [17]. В качестве хранилища данных используется RDF-база данных Sesame. Для получения и записи данных в хранилище системой предоставляется REST веб-сервис, использующий формат JSON-LD [11] для представления RDF-данных. Расширение системы возможно как на клиентской, так и на серверной стороне. Использование библиотеки AngularJS [2] предоставляет гибкие возможности для расширения пользовательского интерфейса системы, добавления новых визуальных представлений данных и изменения системных интерфейсов. Для расширения серверной части приложения используется стандарт OSGI [14].

Адаптер файловой системы представляет собой приложение, запущенное на компьютере пользователя, которое отвечает за передачу скачанных пользователем публикаций на удаленный сервер. Приложение работает в фоновом режиме и, при изменениях содержимого папок, передает информацию об этом на сервер. На стороне сервера, на основании текстового содержимого выгруженных файлов, осуществляется поиск статей в Академии Google. В случае успешного поиска, в систему заносятся метаданные, такие как название, год, авторы, а также ссылки на схожие статьи.

Работа пользователя с информационным пространством осуществляется через веб-интерфейс. На рисунке 4 представлен скриншот пользовательского интерфейса системы. Верхнее меню отвечает за навигацию по типам ресурсов, а также предоставляет возможность поиска ресурсов по системе. Рабочая область портала поделена на три блока. В левой части выводится навигационное меню, отображающее папки, синхронизованные с системой. В центральном блоке выводится список публикаций, находящихся в выбранной папке. Публикации, находящиеся в центральном блоке, могут быть отсортированы по средствам интерфейса Drag&Drop. В правой части рабочей области выводится информация, о выбранной статье – схожие статьи и детальная информация. В панели схожих статей, выводятся ссылки на статьи, схожие с выбранной публикацией. На панели детальной информации, выводятся метаданные о выбранной публикации. Все поля, выводимые на панели детальной информации, при необходимости, могут быть отредактированы пользователем.

## 9 Заключение

В данной статье были рассмотрены задачи управления личными знаниями и информацией, описана архитектура системы, автоматизирующая основные процессы, возникающие в ходе выполнения этих задач, представлен прототип, соответствующий описанной архитектуре. Основные направления, по которым проводится автоматизация это: структурирование информации, поиск информации с которой ранее велась работа, категоризация информации, пополнение сведений метаданными, полученными из внешних источников, поддержка совместной работы. В качестве реализации был представлен прототип системы, поддерживающий работу с научными публикациями. В рамках прототипа реализованы модули каждого из описанных уровней архитектуры – адаптер к файловой системе, сервис анализа публикаций, сервис поиска. За хранение ресурсов, загруженных пользователями, отвечает RDF-база данных Sesame.

В дальнейшем, большее внимание, планируется уделить созданию формальной модели системы, явно описывающей основные модули и процессы, выполняемые в рамках системы. Другим направлением работ, является введение элементов логического вывода, в рамках анализа информационного пространства. Поскольку информация хранится в формате RDF и описывается с помощью языка OWL, принципиальных ограничений в этом вопросе нет. Также, отдельным вопросом, заслуживающим изучения, является анализ процесса работы пользователя с информацией. Анализируя выполняемые

пользователем действия, можно выявлять скрытые зависимости между ресурсами, а также формировать метаданные, которые в дальнейшем могут быть использованы пользователем, например при поиске. Исследования по данному направлению обычно исследуются в рамках работ Task Mining.

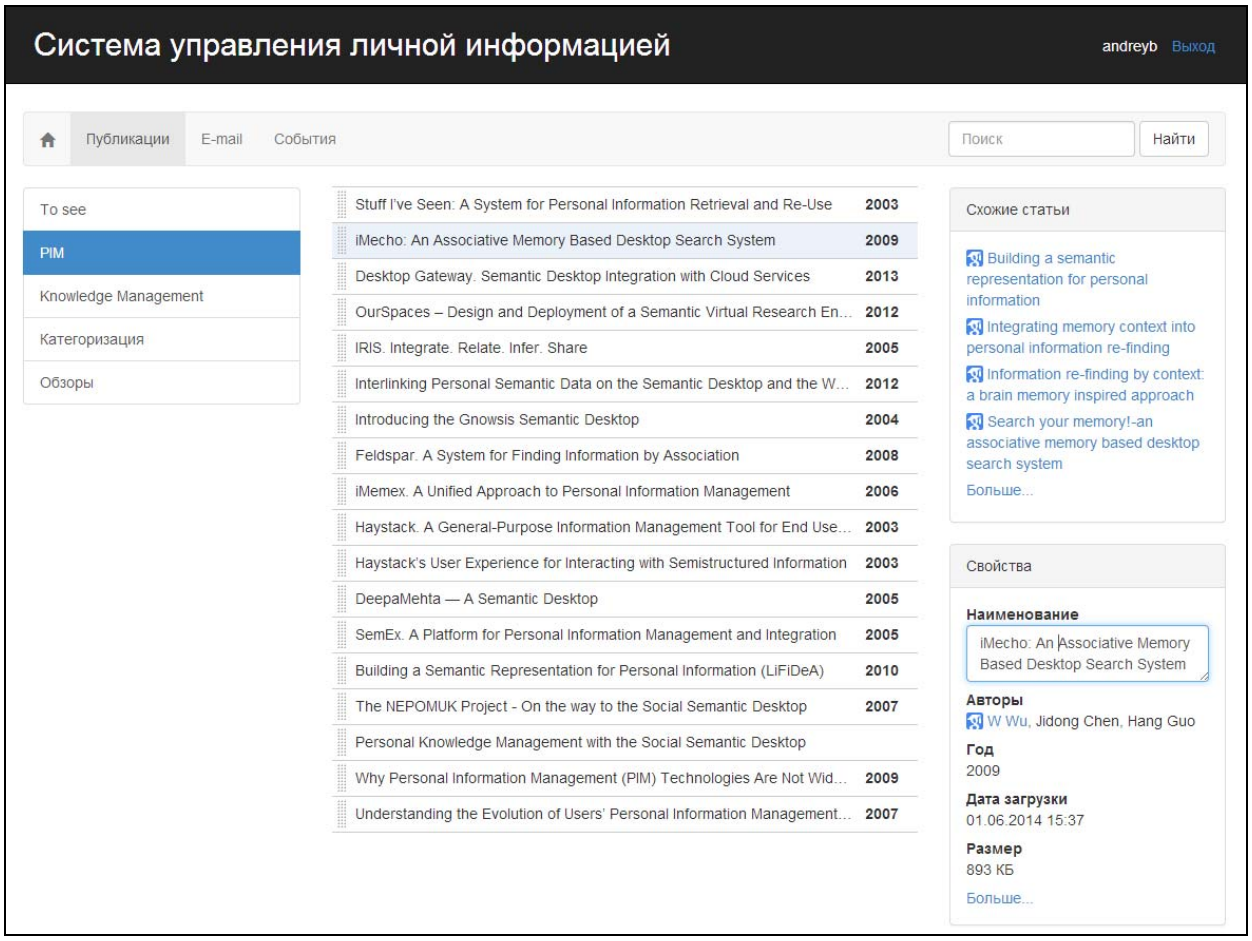



Рис. 4. Пользовательский интерфейс системы

## Литература


[1] AngularJS. https://angularjs.org/

[2] Bush V. As We May Think // The Atlantic Monthly. – Atlantic Media Company, Washington, DC, USA 1945. – V. 176 – P. 101–108.

[3] Chau D. H., Myers B., Faulring A. What to Do when Search Fails: Finding Information by Association // Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Florence, Italy, April 5–10, 2008 – / ACM – New York, NY, USA, 2008. – P. 999–1008.

[4] Chen J., Guo H., Wu W., Wang W. iMecho: an associative memory based desktop search system // CIKM '09: Proceedings of the 18th ACM Conference on Information and Knowledge Management, Hong Kong, China, November 02–06, 2009. – / ACM – New York, NY, USA, 2009. – P. 731–740.

[5] Cheyer A., Park J., Giuli R. IRIS: Integrate, Relate. Infer. Share // Proceedings of the Semantic Desktop Workshop at ISWC Galway, Ireland, November 6, 2005. – 2005. – ISSN 1613-0073. http://ceur-ws.org/Vol-175/17_park_iris_final.pdf

[6] Chirita P. A., Costache S., Nejdl W. et al. Beagle++: Semantically enhanced searching and ranking on the desktop // The Semantic Web: Research and Applications. – Springer Berlin Heidelberg, 2006. – P. 348–362.

[7] Dong X. L., Halevy A. A platform for personal information management and integration // In Proceedings of CIDR 2005, Asilomar, CA, USA, January 4-7, 2005. – P. 119–130. http://www.cidrdb.org/cidr2005/papers/P10.pdf

[8] Dropbox Datastore. https://www.dropbox.com/developers/datastore

[9] Gemmell J., Bell G., Lueder R. et al. MyLifeBits: fulfilling the Memex vision // Proceedings of the tenth ACM international conference on Multimedia, Juan les Pins, France, December 1–6, 2002 /Association for Computing Machinery – New York, NY, USA, 2002. – P. 235–238.

[10] Google Now. http://www.google.com/landing/now/

[11] JSON-LD. A JSON-based Serialization for Linked Data. http://www.w3.org/TR/json-ld/

[12] Kareski A., Jovanovik M., Trajanov D. Desktop Gateway: Semantic Desktop Integration with Cloud Services // BCI'13: Proceedings of the 6th Balkan Conference in Informatics, Thessaloniki, Greece, September 19–21, 2013 / ACM – New York, NY, USA, 2013. – P. 162–168.

[13] Karger D. R., Bakshi K., Huynh D. et al. Haystack: A General-Purpose Information Management Tool for End Users Based on Semistructured Data // Proceedings of CIDR 2005, Asilomar, CA, USA, January 4–7, 2005. – 2005. – P. 13–26. http://www.cidrdb.org/cidr2005/papers/P02.pdf

[14] OSGI. Open Service Gateway Initiative. http://www.osgi.org/Specifications/HomePage

[15] Richter J., Völkel M., Haller H. DeepaMehta – A Semantic Desktop // In Proceedings of the Semantic Desktop Workshop at ISWC Galway, Ireland, November 6, 2005. – 2005. – ISSN 1613-0073. http://ceur-ws.org/Vol-175/30_dm_poster.pdf

[16] Sauermann L., Grimnes G.A., Kiesel M. et al. Semantic desktop 2.0: The gnowsis experience // The Semantic Web – ISWC 2006. – Springer Berlin Heidelberg, 2006. – P. 887–900.

[17] Spring Framework. http://docs.spring.io/spring/docs/4.1.0.RC1/spring-framework-reference/htmlsingle/

[18] Voit K., Andrews K., Slany W. Why personal information management pim technologies are not widespread // ASIS&T 2009 Workshop on Personal Information Management, November 7–8, 2009, Vancouver, BC, Canada – 2009. http://pimworkshop.org/2009/papers/voit-pim2009.pdf

## Model of Semantic Personal Information Management System

Andrey A. Bezdushny, Anatoly N. Bezdushny, Vladimir A. Serebryakov

This paper considers the problem of personal information management by using semantic technologies, proposes the architecture of semantic personal information management systems and presents the prototype of the system implemented in accordance with this architecture. Proposed method develops the idea of the Semantic Desktop – an approach to the personal information space organization in accordance with the Semantic Web and Linked Open Data.

# NLPub: каталог и сообщество русских лингвистических ресурсов

© Дмитрий Усталов
Институт математики и механики им. Н.Н. Красовского УрО РАН
Екатеринбург
dau@imm.uran.ru

## Аннотация

Разрозненность сведений о существующих инструментах и ресурсах для автоматической обработки русского языка является большой проблемой, сильно затрудняющей быстрый старт научных и практических работ, тормозя развитие всего направления. Наличие специализированного каталога лингвистических ресурсов позволит решить эту проблему хотя бы частично. В данной работе представлен каталог и сообщество NLPub, проведено сравнение с аналогичными проектами, описан используемый подход к сбору и представлению данных, продемонстрирована классификация разделов, кратко изложен опыт, полученный с момента основания проекта, и обозначены планы на ближайшее будущее.

## 1 Введение

Словари и тезаурусы, корпусы текстов и банки данных, а также другие информационные ресурсы, имеют огромную ценность в области обработки естественного языка. Это обусловлено спецификой фундаментальных и прикладных задач компьютерной лингвистики, нередко решаемых при помощи разнообразных статистических методов.

За последние годы популярность технологий автоматической обработки естественного языка заметно выросла благодаря таким продуктам, как Apple Siri, Wolfram|Alpha, Google Voice, и др. Возник закономерный общественный интерес, однако разрозненность русскоязычных лингвистических ресурсов затрудняет быстрый старт новых проектов в данной области.

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

Несмотря на ценность и очевидную как научную, так и коммерческую значимость исследований и разработок в области обработки естественного языка, сегодня наблюдаются следующие проблемы:

• отсутствие доступного качественного инструментария и вспомогательных утилит для обработки текста, для распознавания речи, и т.д.;

• нехватка доступных информационных ресурсов: машиночитаемых словарей, тезаурусов, размеченных корпусов текстов, банков данных;

• дефицит сведений об экспертах, тематических мероприятиях и образовательных программах в регионах.

Указанные проблемы делают особенно актуальной задачу сбора, систематизации и распространения сведений о доступных средствах и ресурсах для обработки русского языка.

Цель проекта NLPub[1] заключается в предоставлении на *некоммерческой* основе каталога электронных материалов, направленного на удовлетворение информационных потребностей пользователей, исследователей и разработчиков в области компьютерной лингвистики. Проект NLPub появился и развивается за счет личных средств автора и не имеет аффилированности со сторонними организациями.

## 2 Аналогичные работы

Среди подобных русскоязычных ресурсов можно отметить [1]:

• *Портал знаний о компьютерной лингвистике*[2], созданный в Институте систем информатики им. А.П. Ершова СО РАН, г. Новосибирск;

• *Лингвистика в России: ресурсы для исследователей*[3], созданный в Московском государственном университете им. М.В. Ломоносова, г. Москва;

• *Каталог лингвистических программ и ресурсов в Сети*[4], созданный в Русской виртуальной библиотеке, г. Москва;

• *Математическая и компьютерная лингвистика*[5], созданный в Санкт-Петербургском государственном университете, далее — *mathlingvo*.

### 2.1 «Портал знаний о компьютерной лингвистике»

Портал знаний по компьютерной лингвистике существует с 2006 г. и призван обеспечить систематизацию и интеграцию знаний и информационных ресурсов по компьютерной лингвистике в единое информационное пространство, а также содержательный доступ к интегрированным знаниям и ресурсам.

На портале представлены знания об основных разделах компьютерной лингвистики, о ее предмете и объектах исследования, используемых в ней моделях и методах, разработанных в рамках компьютерной лингвистики технологиях, системах, программных продуктах и лингвистических ресурсах (словарях, корпусах и лингвистических баз данных), а также информация об ученых, сообществах, организациях, включенных в процесс исследования по компьютерной лингвистики и о выполняемых проектах в этой области.

По всей видимости, развитие портала остановилось в 2012 г.

### 2.2 «Лингвистика в России: ресурсы для исследователей»

Научно-образовательный портал «Лингвистика в России: ресурсы для исследователей» создан также в 2006 г. по инициативе Научно-исследовательского вычислительного центра МГУ им. М.В. Ломоносова и Казанского государственного университета им. В.И. Ульянова-Ленина и имеет раздел, посвященный компьютерной лингвистике.

Задачей портала является создание инфраструктуры для поддержки сообществ исследователей и преподавателей для информирования и открытого обсуждения научных и образовательных задач российской лингвистики, интеграция лингвистического сообщества Российской Федерации. На портале собран каталог ссылок на различные российские проекты в области компьютерной лингвистики.

По всей видимости, развитие портала остановилось в 2007 г.

### 2.3 «Каталог лингвистических программ и ресурсов в Сети»

Данный каталог включает в себя описание программ, связанных с анализом текстов и вычислительной лингвистикой, а также соответствующих ресурсов, доступных в Интернете.

Упор при составлении каталога делался на бесплатные программы, доступные для загрузки. Однако также описаны некоторые сетевые и коммерческие версии программ. Тематически каталог разбит на следующие разделы: программы анализа и лингвистической обработки текстов; программы преобразования текстов; психолингвистические программы; генераторы текстов и «говорящие» программы; системы обработки естественного языка; коллекции ресурсов; словари и тезаурусы.

Развитием каталога занимается его единственный составитель, внося достаточно редкие дополнения, правки и изменения. Последнее обновление каталога зафиксировано в 2013 г.

### 2.4 «Математическая и компьютерная лингвистика»

*mathlingvo* — проект кафедры информационных систем в искусстве и гуманитарных науках Санкт-Петербургского государственного университета, созданный в начале 2012 г. и посвященный математической и компьютерной лингвистике в России.

Проект представляет собой коллективный блог под руководством представителей кафедры, в котором уделено внимание перечням тематических конференций, периодических изданий, вакансиям. Также является представительством различных инициатив, таких как OpenCorpora[6].

Лента новостей *mathlingvo* обновляется регулярно и поддерживает добавление новых записей от любого участника на условиях предварительной модерации, однако проект является в большей степени новостным ресурсом и не предоставляет собой каталог как таковой.

## 3 NLPub: каталог и сообщество

**NLPub** — это каталог лингвистических ресурсов для обработки русского языка, основанный на принципах краудсорсинга. День рождения проекта отмечается первого октября 2012 г., когда NLPub был представлен широкой общественности на «Хабрахабре» [2].

**Каталог.** Каталог построен на базе MediaWiki — программного обеспечения, лежащего в основе «Википедии» и «Викисловаря» (рис. 1). Основное отличие NLPub от аналогичных ресурсов, заключается в открытости: любой желающий может внести свои изменения по хорошо известным принципам «Википедии». Благодаря открытости и децентрализованности, материалы NLPub поддерживаются в актуальном, корректном и доступном состоянии с меньшими трудозатратами и большей заинтересованностью участников. Прототипом каталога послужил проект ACLWiki[7], созданный Ассоциацией по компьютерной лингвистике.

**Сообщество.** Важно отметить, что NLPub — это не только краудсорсинговый каталог лингвистических ресурсов, но и сообщество, представленное вокруг этого каталога, вопрос-ответного сервиса NLPub Q&A[8] на базе открытого движка Discourse, и Twitter-аккаунта @nlpub. Также на NLPub расположена и поддерживается документация проекта создания открытого электронного тезауруса русского языка Yet Another RussNet[9].

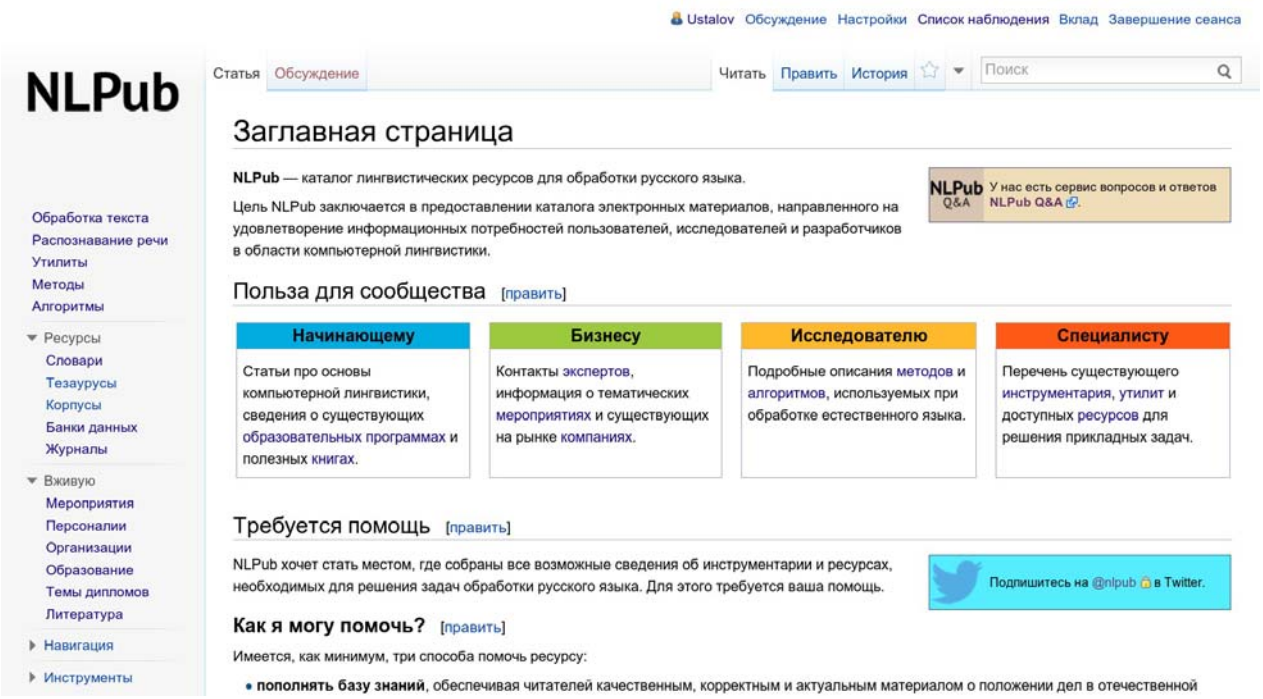



Рис. 1: Главная страница http://nlpub.ru/

## 4 Организация каталога

Каталог организован преимущественно в табличном виде и, в отличие от онтологического подхода [3], представляет собой квазиструктурированные данные в формате более привычной вики-разметки. Это упрощает пополнение и улучшение каталога со стороны человека. Таблицы содержат наиболее ценную информацию об отдельно взятом объекте. Например, для программного продукта в таблице приводится информация о кратком назначении, поддерживаемых языках и условиях использования, а для организации — год основания и ключевые лингвистические продукты.

**Инструменты и утилиты.** Различные инструменты обработки естественного языка (более 140 наименований), распознавания речи (более 20 наименований), утилиты для работы с языковыми моделями и обработки банков данных. Для некоторых инструментов существуют выделенные страницы с подробным описанием и инструкцией по применению. Такими инструментами являются, в частности, Greeb и TreeTagger.

**Ресурсы.** Под *ресурсом* понимаются данные и их производные, используемые в процессе обработки естественного языка: корпусы текстов (более 5 наименований), тезаурусы и словари (более 20 наименований), банки данных. Для некоторых ресурсов существуют выделенные страницы с подробным описанием и перечнем особенностей. Такими ресурсами являются, в частности, словарь Абрамова и YARN.

**Методы и алгоритмы.** Небольшое собрание достаточно важных методов и алгоритмов обработки естественного языка, записанное в виде псевдокода с кратким описанием особенностей и характеристик. Для некоторых алгоритмов существуют выделенные страницы, например про алгоритм удаленной интерполяции и об алгоритме Витерби.

**Образование.** Перечень тематических кафедр, вузов, курсов и программ переподготовки, полезных как начинающим, так и опытным исследователям и разработчикам в области обработки естественного языка.

**Мероприятия.** Список тематических мероприятий и конференций, посвященных обработке естественного языка и компьютерной лингвистике, где можно представить и обсудить результаты своей работы. Существуют выделенные страницы для ряда конференций, например для конференции АИСТ.

**Организации.** Раздел, полезный при поиске работы и при анализе российского рынка решений по обработке естественного языка. Включает в себя достаточно полный список основных игроков на отечественном рынке NLP-продуктов.

**Литература.** Список литературы, полезной для изучения и закрепления знаний об обработке естественного языка и компьютерной лингвистике. Включает ссылки как на учебные пособия, так и на методические указания.

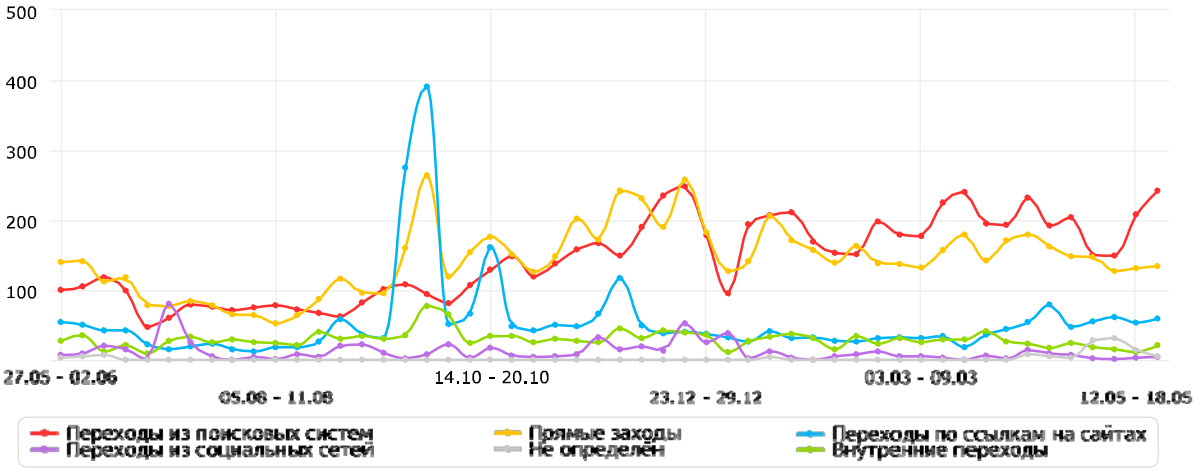



Рис. 2: Данные «Яндекс.Метрики» о посещаемости NLPub с 27 мая 2013 г. по 25 мая 2014 г.

**Эксперты.** Экспериментальный раздел, в котором любой желающий может указать область своей экспертизы и контактную информацию для выполнения какой-либо совместной работы или консультирования.

**Темы дипломов.** Экспериментальный раздел, в котором любой желающий может указать проблемную область, достойную разработки в рамках студенческой или кандидатской работы, и оставить свои координаты для связи.

## 5 Полученный опыт

Первые месяцы существования NLPub были сопряжены с борьбой против активных спам-ботов, специализирующихся на проектах, основанных на MediaWiki. Проблему удалось решить полностью благодаря одновременному принятию трех мер: введению капчи на основе reCAPTCHA при создании учетной записи, подключению черного списка спамерских IP-адресов, а также обязательным подтверждением адреса электронной почты для получения участником возможности вносить правки в статьи.

Данные «Яндекс.Метрики[10]» доступны публично и свидетельствуют о постепенном росте посещаемости NLPub за прошедший год (рис. 2). Это связано с тем, что по мере создания новых страниц и внесения новых сведений страницы становятся более ценными как с точки зрения читателей, так и с точки зрения поисковых машин.

Более высокие позиции в поисковой выдаче способствуют привлечению новых пользователей. Тем не менее, на текущий момент можно считать активность пользователей *эфемерной*, то есть человек попадает на NLPub во время поиска ответа на свой вопрос при помощи поисковых систем. Это свидетельствует о том, что база постоянных читателей и авторов недостаточно велика: упоминание ресурса в популярных блогах или сайтах отражается в статистике как резкий скачок вверх.

В настоящий момент сообщество находится на достаточно ранней стадии своего развития, однако уже сегодня на NLPub Q&A можно получить ответы на достаточно острые и нетривиальные тематические вопросы.

## 6 Заключение

Анализ поисковых запросов и опрос аудитории NLPub показывает заинтересованность в отдельных статьях, посвященных конкретным инструментам, методам и алгоритмам. Эта информация обобщена на специальной странице http://nlpub.ru/TODO. Выделяется три направления предстоящей работы:

- общие статьи об основных разделах автоматической обработки естественного языка: графематический, морфологический, синтаксический анализ, информационный поиск, сходство документов, машинный перевод, извлечение ключевых слов, автоматическое реферирование, анализ тональности, и др.;
- статьи о популярных моделях, методах и алгоритмах: векторные модели (tf-idf, «мешок слов», косинусная мера близости), теоретико-графовые модели, n-граммные модели, общие методы алгоритмического обучения, используемые в лингвистике (перцептрон, наивный Байесовский классификатор, EM-алгоритм), и др.;
- обучающие статьи о важном или слабо документированном программном обеспечении: «Томита-парсер», FreeLing, Stanford NLP, MaltParser, NLTK, и др.

На сегодняшний день можно отметить два основных недостатка ресурса. Во-первых, слабая заполненность некоторых разделов, таких как «Персоналии» и «Литература». Это вызвано достаточно небольшим возрастом NLPub и предполагается, что эта проблема решится путем органического роста проекта. Во-вторых, отсутствие связей между разными разделами каталога усложняет навигацию. Решение этой проблемы состоит в добавлении соответствующих внутренних

ссылок и предоставлении наглядной карты сайта на одной из главных страниц ресурса.

Повышение охвата пользователей и снижение эфемерности их активности можно выполнить путём интеграции с ресурсом *mathlingvo* для автоматической публикации сводок новостей с указанием соответствующих ссылок.

В отдаленной перспективе было бы интересно преобразовать каталог NLPub в семантическую вики для предоставления машиночитаемых данных с одновременным сохранением удобства внесения правок и дополнений в материалы проекта.

**Благодарности.** Автор выражает огромную благодарность всем пользователям NLPub, принявшим участие в работе над материалами проекта.

## Литература

[1] Д. А. Усталов. Каталоги лингвистических ресурсов: состояние и перспективы // Молодой ученый. — 2012. — Т. 1, №12 (47). — С. 148–152.

[2] Д. А. Усталов. NLPub — каталог лингвистических решений. http://habrahabr.ru/post/152429/

[3] Ю. А. Загорулько и др. Подход к построению предметной онтологии для портала знаний по компьютерной лингвистике // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог». — 2006. — С. 148–151.

## Примечания

[1] http://nlpub.ru/
[2] http://uniserv.iis.nsk.su/cl/
[3] http://uisrussia.msu.ru/linguist/_B_comput_ling.jsp
[4] http://www.rvb.ru/soft/catalogue/catalogue.html
[5] http://mathlingvo.ru/
[6] http://opencorpora.org/
[7] http://aclweb.org/aclwiki/
[8] http://qa.nlpub.ru/
[9] http://russianword.net/
[10] https://metrika.yandex.ru/stat/?counter_id=17329045

## NLPub: a Catalogue and a Community for Russian Linguistic Resources

Dmitry Ustalov

The lack of coordination in the information on existing tools and resources for Russian language processing has become a significant problem. Such a problem complicates both research and practical applications thwarting with the progress of the whole field. A specialized catalogue for linguistic resources may assist one in getting this problem solved. In this survey NLPuba catalogue and a community for Russian linguistic resources is presented and compared with its analogs. Its data gathering and representation approaches are also described and the merotomy is demonstrated. The experience obtained since the project start is outlined and future work directions are stated.

# Refinement of Russian Sentiment Lexicons Using RuThes Thesaurus

© N. V. Loukachevitch © I. I. Chetviorkin
Lomonosov Moscow State University
louk_nat@mail.ru ilia2010@yandex.ru

## Abstract

The paper describes a combined approach to extraction of a domain-specific sentiment lexicon. At first, an initial version of a domain-specific lexicon is obtained by application of a supervised model. At the second stage, the ordered list of sentiment words is refined using the thesaurus information. This combined model is applied to several domains and at last the domain-specific sentiment lexicons are united to create an improved version of the Russian sentiment lexicon in the generalized domain of products.

## 1 Introduction

Automatic sentiment analysis of texts is a fast-developing technology in natural language processing. The task of automatic sentiment lexicon construction and improvement is a basic task for sentiment analysis of texts. There are no freely available sentiment lexicons for many languages or the quality of such lexicons is desired to be better. For example, in Russian only one automatically extracted sentiment lexicon has been published [1].

Besides, sentiment analysis of domain-specific texts requires adaptation of machine-learning models or sentiment lexicons to the target domain [6]. So, some sentiment words can loss their polarity in specific domains. For example, such word as *evil* in the movie domain usually refers to the movie plot, but not a user opinion.

Other words can obtain the sentiment polarity in a specific domain. For example, *word киношный* (adjective to Russian word *кино (movie))* can have the negative polarity with the meaning *"far from the real life"*. Another example - word *атмосферный (*adjective to word *атмосфера (atmosphere))* has the positive polarity in art-related domains denoting "creation of a special mood or feeling" (as *atmospheric* in English) – this is a relatively new sense of this word for Russian, not described in Russian dictionaries.

**Proceedings of the 16th All-Russian Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" — RCDL-2014, Dubna, Russia, October 13–16, 2014.**

Automatic extraction of sentiment words can be based on corpus-based or resource-based (dictionary, thesauri) approaches. In this paper we offer a combined approach to extracting sentiment lexicons. At first, an initial version of a domain-specific lexicon is obtained by application of a supervised model on the basis of statistical and linguistic features of sentiment words. This lexicon is presented as a list of words ordered by the decreased probability of their sentiment orientation. At this stage we obtain some sentiment words that are absent in dictionaries or having the domain-specific sentiment polarity. We extract sentiment-oriented words without any positive or negative labels because we consider this process as the first step to further polarity lexicon generation.

At the second stage, the ordered list of sentiment words is refined using the thesaurus information, in our case, newly published thesaurus of Russian language RuThes[1]. We trained a supervised model and tuned a combined model in the movie domain. Then this augmented model was utilized in four other domains. Finally, extracted sentiment lexicons from five domains are united to generate a high quality lexicon in the general product domain for Russian (ProductSentiRus+).

The reminder of this article is organized as follows. In Section 2 we review methods for generating sentiment lexicons. Section 3 briefly presents the structure of RuThes thesaurus, the Russian newly published thesaurus intended for natural language processing. Section 4 presents an approach for extracting sentiment words in various domains. Section 5 describes the refinement of the lexicon in the general product domain. To evaluate the quality of the obtained general resource extrinsically, we conduct the experiments on the tweet subjectivity classification task.

## 2 Related Work

There are two main approaches to sentiment lexicon extraction: corpus-based and dictionary-based methods.

Corpus-based methods utilize co-occurrence of words with each other [5, 9, 10], or appearance them in specific collocations or lexico-syntactic patterns [4]. Contemporary corpus-based approaches exploit a large

[1] labinform.ru/ruthes/index.htm

or hundreds of thousands of user reviews as in [6].

Dictionary-based methods utilize available electronic dictionaries and thesauri and usually begin their work from a set of seed words. In [3] SentiWordNet resource is described. It is the result of the automatic annotation of all the synsets of WordNet where each synset is associated to three numerical scores that indicate how positive, negative, and neutral the terms contained in the synset are. Different senses of the same word may thus have different opinion-related properties.

In [8] authors study semi-supervised approaches to label the polarity of words in a graph of lexical relations such as WordNet. They apply several methods: MinCut, Randomized MinCut, Label Propagation algorithm, described in [11], and show that Label Propagation algorithm achieves the best results. These and similar graph-based algorithms are also utilized in corpus-based approaches to sentiment lexicon extraction [4, 9].

In many studies domain-specific sentiment lexicons are created with corpus-based approaches using various types of propagation from a seed set of words, usually a general sentiment lexicon [6]. An important problem of such approaches is to determine an appropriate seed lexicon, which can depend on the domain.

In our study we create a domain-specific sentiment lexicon from medium-size datasets using multiple features of words and several collections without any co-occurrences between words. Then we improve an initial sentiment lexicon using sentiment labeling of the thesaurus concepts in a specific domain practically without pre-determined seed words. We use only two fixed seed opinionated words (*bad, good*), other potential sentiment words are obtained automatically from a ranked list of a sentiment lexicon (words ordered by the probability of their sentiment orientation) extracted from domain-specific collections.

## 3 RuThes Linguistic Ontology

In our study we use RuThes Thesaurus of Russian language. RuThes is a linguistic ontology for natural language processing, i.e. an ontology, where the majority of concepts are introduced on the basis of actual language expressions. For a long time RuThes has been manually developed within various NLP and information-retrieval projects, and now it is available for public use. The publicly available version of RuThes contains around 100 thousand Russian words and expressions [7].

If compared to WordNet-style resources RuThes is organized as a united semantic net where different parts of speech (nouns, verbs, adjectives) can be text entries of the same concepts. Each concept has a unique unambiguous name. Concepts can be connected with several types of conceptual relations. In addition, RuThes includes a lot of multiword expressions useful for applications and terms of so-called Sociopolitical domain – a broad domain of contemporary social relations, which includes terms from political, economic, military, sports and other fields [7].

Ambiguous words in RuThes are described similar to WordNet-style resources through attachment to several concepts. For example, in the current version of RuThes word *пресный* is attached to three concepts:

- *ПРЕСНАЯ ВОДА (fresh water)*;
- *ПРЕСНЫЙ, БЕЗВКУСНЫЙ (tasteless, bland in taste);*
- *ПРЕСНЫЙ (НЕИНТЕРЕСНЫЙ) (uninteresting).*

The first concept is neutral and not relevant to the movie domain. The second concept is negative but also irrelevant to the domain. Last concept is negative and relevant to the domain.

## 4 Extraction of Sentiment Lexicons

In this section an algorithm for extraction of sentiment words in a specific domain is described. The results of this algorithm are refined using the iterative procedure on the basis of RuThes thesaurus to obtain a high quality domain-specific sentiment lexicon.

Such a method is applied to four other domains without additional manual labeling and the results are combined in a sentiment lexicon in a generalized product domain ProductSentiRus+.

Table 1. Domain-specific collection statistics

| Domain | Reviews | Descriptions |
|---|---|---|
| Movies | 28, 773 | 17, 680 |
| Books | 23, 883 | 22, 321 |
| Games | 7, 928 | 1, 853 |
| Digital Cameras | 10, 208 | 920 |
| Mobile Phones | 30, 620 | 890 |

### 4.1 Extraction of domain-specific sentiment lexicon based on multiple features

At the first stage sentiment words are extracted with a corpus-based method utilizing a trained machine-learning model applied to several domain-specific text collections.

The first domain-specific collection (with high concentration of sentiment words) is a collection of user reviews in the domain (review collection) with numeric scores specified by their authors. In these experiments collections were gathered from the online services *imhonet.ru* and *market.yandex.ru* in five domains: movies, books, computer games, mobile phones and digital cameras. The second domain-specific collection (with low concentration of sentiment words) is a text collection of object descriptions (e.g. plots for movies). The overall collection statistics can be found in Table 1.

Another contrast corpus was a collection of two million news documents. Such a collection is useful for correct classification of general neutral words frequent in news.

Using such collections the feature representation is calculated for each word. The set of features includes the following feature types [1]:

- Frequency-based: collection frequency, document frequency, frequency of capitalized words, frequency of co-occurrence with polarity shifters (*no, not*), TFIDF;
- Score-based: deviation from the average score, word score variance, sentiment category likelihood for each (word, category) pair;
- Linguistic: Four binary features indicating the word part of speech, two binary features reflecting POS ambiguity, predefined list of prefixes of a word.

To train supervised machine learning algorithms, all words with the frequency greater than three in the movie review collection were labeled manually by two assessors. If there was a disagreement about the sentiment of a specific word, the collective judgment after discussion was used as the final ground truth. As a result of the assessment procedure the list of 4079 sentiment words was obtained. The best quality of classification using labeled data was shown by the ensemble of three classifiers: Logistic Regression, LogitBoost and Random Forest from WEKA programming package.

The result of this corpus-based method is a ranked list of domain-specific words ordered by the probability of their sentiment orientation – further *sentiment weights.* The algorithm boosts sentiment words to have high weights (to be closer to the beginning of the list) and neutral words to have low weights.

So in the movie domain in the list of more than 18 thousand words the following words are located in the first positions:

*трогательны (affective), отстой (trash), фигня (crap), отвратительно (disgustingly), посредственный (satisfactory), предсказуемый (predictable), любимый (loved).*

Word *атмосферный (atmospheric)* takes 830th, high-opinionated position in the list.

Evident sentiment adjectives of the movie domain *пресный* and *безвкусный (*both are translated into English as *tasteless*) take even higher opinionated positions: 139th and 193th . But their noun derivations *пресность, безвкусие, безвкусность,* and *безвкусица* are less successful. *Пресность, безвкусие, безвкусность,* are absent from the list because of low frequency; *безвкусица* takes 1515th place in the list. So thesaurus-based improvements may be quite possible.

The obtained model was applied to four other domains (books, games, digital cameras, mobile phones) without any additional manual efforts. The quality of extracted sentiment lexicons was measured using precision measures and presented in the Baseline columns of Table 2.

### 4.2 Refinement of domain-sentiment sentiment lexicons using RuThes thesaurus

To increase the quality of extracted sentiment lexicons we refine them with general thesaurus for Russian language RuThes [7]. The input of the refinement algorithm is a ranked sentiment list obtained with the model described in the previous subsection; however, a similar input can be also generated with other methods.

Table 2. Precision of the domain-specific lexicons at levels 100 and 1000 first words in the sentiment lists

| .Domain | Baseline P@100, % | +RuThes P@100, % | Baseline P@1000, % | +RuThes P@1000, % |
|---|---|---|---|---|
| Movie | 99 | **100** | 81.5 | **85.5** |
| Books | 99 | **100** | 86.0 | **86.2** |
| Games | 97 | **100** | 72.2 | **73.1** |
| Digital Cameras | 85 | **92** | 65.8 | **66.3** |
| Mobile Phones | 85 | **97** | 73.2 | **78.6** |
| General Product Domain | **100** | **100** | 90.5 | **95.2** |

Words from the ranked sentiment list are quite different relative to RuThes descriptions. Some words are not described in RuThes, e.g. three of the most probable sentiment words in the movie domain are absent in RuThes, others are mentioned in text collections exactly in the same senses as described in RuThes, the thirds (e.g. *atmospheric*) are described in RuThes but have an additional (or the other) sentiment polarity. So we should try to correct the word order in the sentiment list carefully applying RuThes descriptions.

The main idea of the lexicon refinement is to label conceptual subgraphs of the thesaurus network as sentiment or neutral and use this labeling to reorder the initial sentiment list. This process in contrast to such a method as Label Propagation [8, 11] should be also regulated with previously obtained sentiment weights of words.

Let us denote a domain-specific lexicon with $W_D$ where all words are ordered by their sentiment weights (*sw*). Initially the algorithm forms two sets of thesaurus concepts using words from the both sides of the list $W_D$: $L_s$ – concepts supposed to be opinionated, $L_n$ – neutral concepts. With this aim the initial average sentiment weights *csw* for all concepts containing words from $W_D$ are calculated. Then the algorithm adds to $L_s$ concepts with the high average weight ($csw_s > 0.85$) and also two pre-defined concepts, corresponding to senses of words *bad* and *good*.

Concepts with the low average weight ($csw_n < 0.05$) are added to the set of neutral concepts $L_n$, which formed without any pre-defined concepts. The thresholds for $csw_s$ and $csw_n$ are obtained from experiments.

Further, every set ($L_s$ and $L_n$) is iteratively augmented with concepts using two conditions: the average sentiment weight threshold and the number of direct thesaurus relations to the existing sets. Formally, $L_s$ and $L_n$ are calculated as shown in Algorithm 1 listing. The algorithm uses also the following additional notation:

– *Adj (L)* is a set of direct-link neighbor concepts to set of concepts *L*;

– *nlink (C, L)* is a function returning the number of direct thesaurus relations between concept *C* and set *L*.

In the last step *sw* weights of all words corresponding to $L_s$ concepts are modified by multiplying them by factor $k_1$ ($k_1 > 1$) and all words corresponding to $L_n$ are multiplied by factor $k_2$ ($0 < k_2 < 1$). The resulting list is reordered by weight.

Low-frequent words (with the frequency less than 3) of the source domain collection are absent in the initial ranked sentiment list and therefore do not have any sentiment weights. The initial sentiment weights of such words are calculated as the average sentiment weights of concepts they related to. The weights of these concepts, in turn, are calculated from other, more frequent synonyms or from average weights of neighbor concepts in the labeling process.

*Algorithm 1. Weights+Relations*

```
Input: concept list with sentiment
weights csw
Output: Ls,  Ln
  Ls = {Cbad, Cgood}∪Chigh, Chigh={Ci:
csw(Ci)>0.85},
  Ln = Ln∪Clow, Clow = {Ci: csw(Ci)<0.05}
  θ = 0.1, Nlink = 3, Ls_iter = Ls, Ln_iter
= Ln
while θ<0.6
    for C ∈ Adj(Ls)
      if nlink(C, Ls) > Nlink &&
csw(C)>0.7-θ
        then Include(C,Ls);
  end

    for C ∈ Adj(Ln)
      if nlink(C, Ln) > Nlink && csw(C)<
θ
        then Include(C,Ln);
    end

    if Ln == Ln_iter && Ls == Ls_iter
        then Nlink = Nlink-1;
  if Nlink == 0
        then θ = θ + 0.05, Nlink = 3;
  Ln = Ln_iter, Ls = Ls_iter
end
```

All parameters of the algorithm are tuned in the movie domain and then applied to four other domains. The quality of domain specific sentiment word lists can be found in Table 2 in RuThes column.

After application of this algorithm in the movie domain our example words *пресный, пресность, безвкусный, пресность, безвкусица* have the following places in the generated sentiment list: *пресный* – 81, *пресность* – 86, *безвкусный* – 115, *безвкусность* – 172, *безвкусие* – 173, *безвкусица* – 943.

The words related to the neutral sense of word *пресный – ПРЕСНАЯ ВОДА (fresh water)* preserved their very low positions in the sentiment list: *вода (water)* – 23059, *айсберг (iceberg)* – 26124.

## 5 Improvement of General Sentiment Lexicon Using RuThes Thesaurus

Integrating sentiment lexicons from various product-oriented domains it is possible to create a general sentiment lexicon in the broad domain of products and services. Such a lexicon for Russian was described in [1], it was called ProductSentiRus[2].

In that paper the lexicons of five domains were summed up using a formula intended to boost words that occur in many different domains and have high weights in each of them.

Thus, for combining multiple weighted word lists the following formula was used:

$$R(w) = \max_{d \in D}(prob_d(w)) \sum_{d \in D} \frac{1}{|D|}\left(1 - \frac{pos_d(w)}{|d|}\right),$$

where $D$ – is the domain set with five domains, $d$ is the sentiment word list for a particular domain and $|d|$ is the total number of words in this list. Functions $prob_d(w)$ and $pos_d(w)$ are the sentiment probability and position of the word in the list $d$. Precision@1000 of ProductSentiRus was reported as 90.5%. Similar combination of improved sentiment lexicons in the new resource (ProductSentiRus+) yields 95.2% in terms of Precision@1000 (Table 2).

We took 5000 of the most probable sentiment words of ProductSentiRus+ lexicon for further work (the same amount as in a previous version) and evaluated it in the tweet subjectivity classification task.

The evaluation is based on TEST data set described in [10], which include two thousand tweets in Russian. We assumed that ProductSentiRus+ comprises sentiment units of Internet language. A tweet was classified as subjective if it contained at least one word from the lexicon. Table 3 demonstrates that such a generalized lexicon can be useful also in tweet subjectivity analysis.

Table 3. Quality of tweet subjective classification

| Lexicon | *P* | *R* | $F_{subj}$ |
|---|---|---|---|
| Twitter-based lexicon from (Volkova, 2013) | – | – | 61.0 |
| ProductSentiRus (data from Volkova, 2013) | – | – | 61.0 |
| ProductSentiRus+ | 58.5 | 84.7 | **69.2** |

## 6 Conclusion

In this paper we described a combined approach to extraction of domain-specific sentiment lexicons. At first, an initial version of a domain-specific lexicon is obtained by application of a supervised model. At the second stage, the ordered list of sentiment words is refined using information described in RuThes

[2] http://www.cir.ru/SentiLexicon/ProductSentiRus.txt

thesaurus of Russian language, which was lately published.

This combined model is applied to several domains and at last domain-specific sentiment lists are united to create a sentiment word list in the generalized domain of products – ProductSentiRus+, which is an improved version of the only published Russian sentiment lexicon and will be also publicly available. The proposed approach can be applied to other languages and can utilize other thesauri.

## Acknowledgments

This work is partially supported by RFBR grant 14-07-00682.

## References

[1] Ilia Chetviorkin and Natalia V Loukachevitch. Extraction of Russian sentiment lexicon for product meta-domain. In COLING, 2012. P. 593–610.

[2] Yejin Choi and Claire Cardie. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Vol. 2, 2009. P. 590–598.

[3] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In Proceedings of LREC, vol. 6, 2006. P. 417–422.

[4] Song Feng, Jun Seok Kang, Polina Kuznetsova, Yejin Choi. Connotation Lexicon: A Dash of Sentiment Beneath the Surface Meaning. In Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics ACL-2013. 2013.

[5] Vasileios Hatzivassiloglou and Kathleen R McKeown. Predicting the semantic orientation of adjectives. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, 1997. P. 174–181.

[6] Raymond Lau, Chun-Lam Lai, Peter Bruza, Kam-Fai Wong. Leveraging web 2.0 data for scalable semi-supervised learning of domain-specific sentiment lexicons. Proceedings of the 20th ACM international conference on Information and knowledge management. ACM. 2011.

[7] Natalia Loukachevitch and Boris Dobrov. RuThes Linguistic Ontology vs. Russian Wordnets. In Proceedings of Global Wordnet Conference. 2013.

[8] Delip Rao and Deepak Ravichandran. Semi-supervised polarity lexicon induction. In Proceedings of the 12th Conference of the European Chapter of the ACL, EACL-2009, 2009. P. 675–682.

[9] Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. The viability of web-derived polarity lexicons. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010. P. 777–785.

[10] Svitlana Volkova, Theresa Wilson, and David Yarowsky. Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual twitter streams. In Proceedings of the 51st Annual Meeting of the Association of Computational Linguistics (ACL13), 2013. P. 505–510.

[11] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University. 2002.

# Сеть естественных иерархий терминов новостных текстов по событиям «Евромайдана»

© Д. В. Ландэ © А. А. Снарский © Е. В. Ягунова
Институт проблем регистрации информации НАН Украины, НТУУ «Киевский политехнический институт», Киев, Украина
dwlande@gmail.com asnarskii@gmail.com
Санкт-Петербургский государственный университет, Санкт-Петербург, Россия
iagounova.elena@gmail.com

**Аннотация**

Описывается методика построения сетей иерархий терминов на основе тематических массивов новостных сообщений. Построены и исследованы такие сети, сформированные на основе автоматической обработки полных текстов сообщений о событиях, связанных с «Евромайданом» в Киеве.

## 1 Постановка проблемы

Построение большой тематической онтологии – сложная и затратная проблема. Определенным этапом разработки общих онтологий является формирование словарных номенклатур, терминологических онтологий. Эффективный автоматический отбор отдельных терминов для таких конструкций на основании неразмеченных текстовых массивов – не решенная окончательно задача [5, 6]. Проблема автоматического установления связей, построения сетей из таких терминов также до сих пор остается открытой.

Другой важной задачей является формальная оценка всплеска новых тем в информационных потоках, и, соответственно, терминов, маркирующих эти темы. Сегодня лингвист, работающий с новостными текстами, не может не заинтересоваться спецификой разных современных сегментов (срезов) по данным СМИ, потоков новостных сообщений [4, 8]. В частном случае, по терминам-маркерам можно понимать соответствие отдельных новостных сюжетов тематикам целых информационных потоков, оценивая используемую в них лексику.

Ниже описаны подходы к формированию терминологической основы цепочки событий, отражаемых в сообщениях электронных СМИ, а также отдельных сюжетов тематических новостей за определенные временные периоды, а также формирование на основании некоторых принципов языковой сети из отобранных терминов. Соответствие терминологии отдельного событийного сюжета и общей тематической терминологии (или терминологии цепочки связанных событий) можно рассматривать как формальный критерий релевантности данного события и рассматриваемой тематики (цепочки событий).

Предварительные этапы формирования языковой сети, связанной с цепочкой взаимосвязанных событий, включают такие шаги:

1. Нахождение релевантных тематике сообщений – формирование корпуса тематических новостных сообщений.

2. Определение динамики тематических сообщений.

3. Определение критических точек (дат) в динамике тематических сообщений.

4. Определение объектов мониторинга (терминов).

Рассмотрим их более подробно.

## 2 Формирование корпуса тематических новостных сообщений

На первом этапе выбирается исходный текстовый корпус, в качестве которого рассматриваются новостные сообщения по тематике противостояний в Киеве в 2013–2014 гг., связанных с так называемым «Евромайданом». Для отбора и последующего анализа тематических сообщений была использована система контент-мониторинга InfoStream [3]. Для нахождения релевантных тематике новостных сообщений был составлен запрос:

***(майдан|евромайдан)&(избиен|разгон|штурм|беркут|молотов|титушк|погиб)&lang.RUS,***

по которому в период с ноября 2013 г. по март 2014 г. было найдено свыше 200 тысяч новостных сообщений на веб-сайтах РуНета (рис. 1).

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

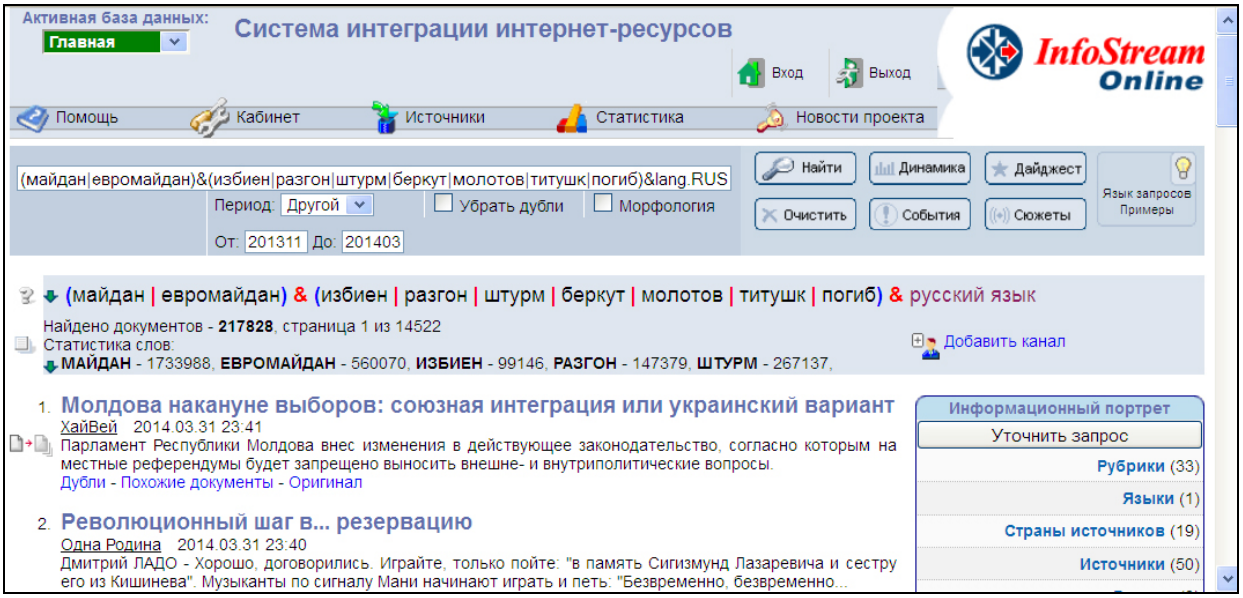



Рис. 1. Поисковый интерфейс системы InfoStream


Рис. 2. Динамика количества публикаций по запросу

## 3 Определение динамики тематических сообщений

Режим «Динамика событий» системы контент-мониторинга позволяет получить данные о количестве публикаций по заданному запросу за указанный промежуток времени. Эти данные отображаются в виде графика (рис. 2).

После этого данные временной динамики за каждые сутки нормируются, т.е. формируется временной ряд, содержащий относительные значения, равные отношению количества тематических сообщений к общему потоку сообщений за сутки (рис. 3). Это, в частности, позволяет избавиться от недельной периодичности в количестве тематических публикаций. Затем происходит переход к процедуре определения критических точек в данном временном ряду.

## 4 Определение критических точек в динамике сообщений

Критические точки как локальные максимумы временного ряда динамики публикаций можно определить, например, визуально по графику, представленному на рис. 3. Вместе с тем, существуют несколько научно-обоснованных методик, одна из которых базируется на вейвлет-анализе [2]. В работе [3] показано, что вейвлет «мексиканская шляпа» наиболее точно отражает динамику информационных операций, результаты применения этого вейвлета приведены на рис. 4, благодаря чему выбраны три даты (2013.11.30, 2014.01.22, 2014.02.19), соответствующие критическим точкам исследуемого процесса.

## 5 Выбор объектов мониторинга

После определения критических точек во временном ряду с помощью системы контент-мониторинга выполняется построение основных сюжетных цепочек из сообщений, соответствующих запросу за выбранные даты, которые определяют основные события за указанные даты (рис. 5).

Для последующего анализа отбирается три массива сообщений, соответствующие трем выбранным датам, особенности лексического состава которых являются объектами мониторинга.

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | **Дата** | **Всего** | **Понятие** | | | |
| 2 | 01.11.2013 | 84687 | 30 | | 0.000354 | |
| 3 | 02.11.2013 | 31369 | 18 | | 0.000574 | |
| 4 | 03.11.2013 | 27291 | 12 | | 0.00044 | |
| 5 | 04.11.2013 | 64836 | 28 | | 0.000432 | |
| 6 | 05.11.2013 | 88322 | 32 | | 0.000362 | |
| 7 | 06.11.2013 | 89133 | 68 | | 0.000763 | |
| 8 | 07.11.2013 | 86815 | 35 | | 0.000403 | |
| 9 | 08.11.2013 | | | | | |
| 10 | 09.11.2013 | | | | | |
| 11 | 10.11.2013 | | | | | |
| 12 | 11.11.2013 | | | | | |
| 13 | 12.11.2013 | | | | | |
| 14 | 13.11.2013 | | | | | |
| 15 | 14.11.2013 | | | | | |
| 16 | 15.11.2013 | | | | | |
| 17 | 16.11.2013 | | | | | |
| 18 | 17.11.2013 | | | | | |
| 19 | 18.11.2013 | | | | | |
| 20 | 19.11.2013 | | | | | |
| 21 | 20.11.2013 | | | | | |
| 22 | 21.11.2013 | | | | | |
| 23 | 22.11.2013 | | | | | |
| 24 | 23.11.2013 | | | | | |
| 25 | 24.11.2013 | 31015 | 733 | | 0.023634 | |
| 26 | 25.11.2013 | 89885 | 2721 | | 0.030272 | |

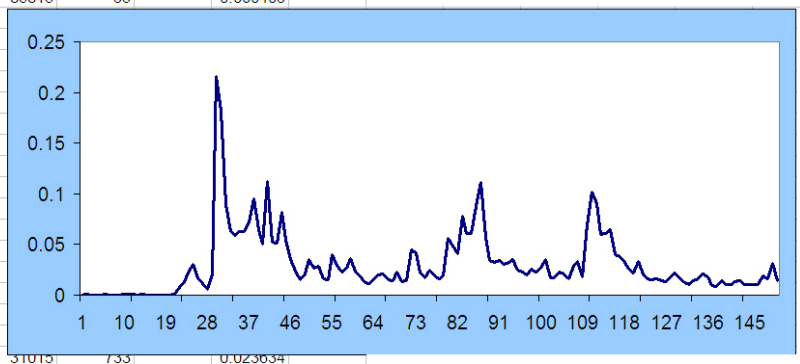



Рис. 3. Нормированная динамика тематических публикаций

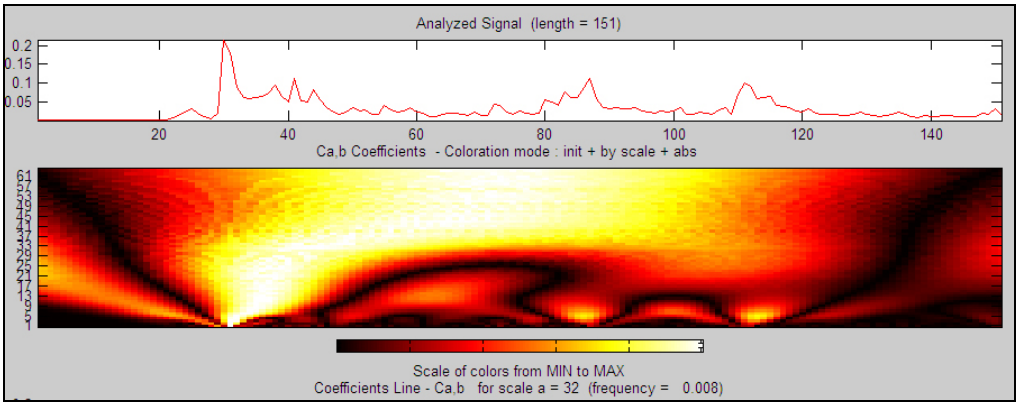



Рис. 4. Вейвлет-спектограммы исследуемого временного ряда

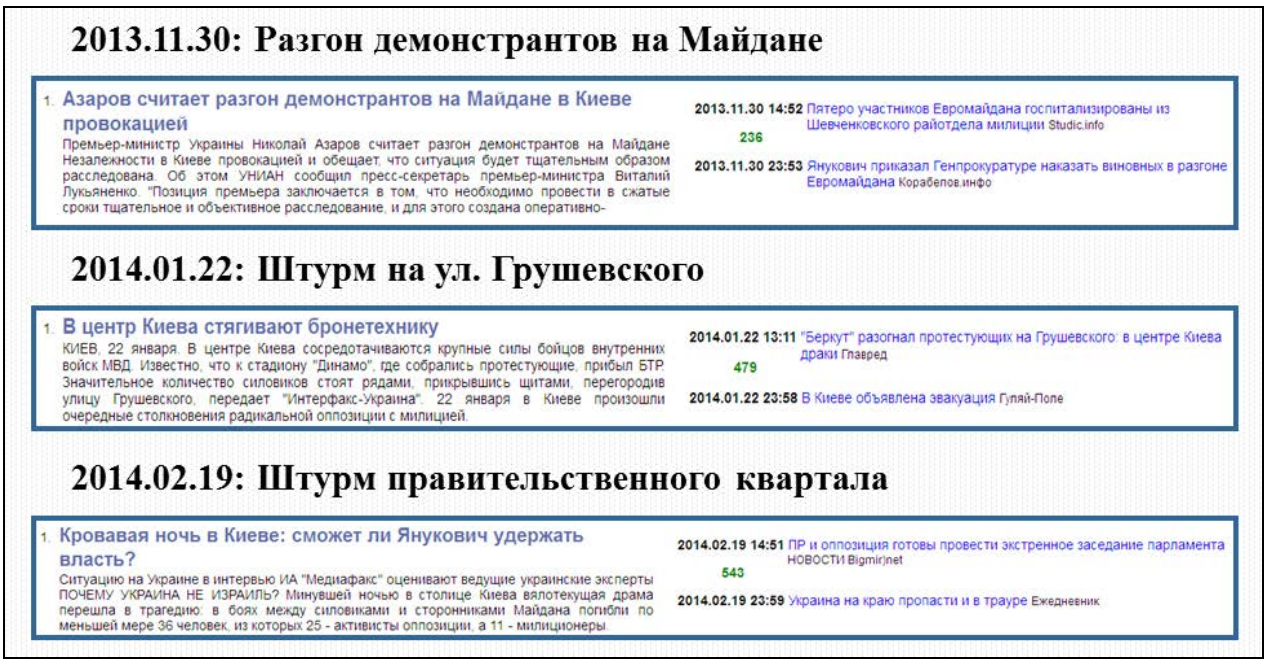



Рис. 5. Основные сюжетные цепочки за выбранные даты

Предварительная обработка отобранных текстовых массивов предусматривает выделение фрагментов текстов (отдельных сообщений, абзацев, предложений, слов, биграмм, триграмм), исключение нетекстовых символов, отсечение флективных окончаний – стемминг.

Далее каждому отдельному терму из текста (слову, биграмме или триграмме) ставится в соответствие оценка его «дискриминантная сила», а именно TFIDF, которая в каноническом виде равна произведению частоты соответствующего термина (Term Frequency) в фрагменте текста на двоичный логарифм от величины, обратной к количеству фрагментов текста, в которых этот терм встретился (Inverse Document Frequency) [14].

## 6 Сеть естественных иерархий терминов

Сеть естественной иерархии терминов (СЕИТ) базируется на разработанной ранее авторами данного доклада методологии выявления информационно-значимых элементов текста, опорных словах и словосочетаний [10, 12]. Использование таких элементов позволяет формировать сетевые информационные портреты, охватывать отдельные области знаний. Опорные слова и словосочетания как правило выбираются с учетом такого их свойства, как дискриминантная сила. Вместе с тем, одного этого свойства часто оказывается недостаточно для построения терминологических онтологий. Иногда слова с низкой дискриминантной силой, в частности, наиболее частотные слова из выбранной предметной области (например, слова «Украина», «Майдан», «Протест» в корпусе новостных сообщений о событиях, связанных с так называемым «Евромайданом» в Киеве) оказываются важнейшими для задач, которые рассматривается ниже.

Формирование сети естественных иерархий терминов базируется на контенте текстовых корпусов выбранной для анализа направленности. «Естественность» в этом случае понимается как отказ при формировании сети от специальных методов смыслового анализа, в том числе, разбора предложений по частям речи. Все связи в такой сети определяются естественным взаимным расположением слов и словосочетаний, которые экстрагируются из текстов статистически значимых объемов. Сеть естественных иерархий терминов, создаваемая полностью автоматически, может рассматриваться как основа для дальнейшего автоматизированного формирования терминологической онтологии с участием экспертов. Методика формирования сети естественных иерархий терминов, которая рассматривается в этой работе, предусматривает формирование компактифицированного графа горизонтальной видимости (CHVG), расчет новых весовых значений слов, биграмм и триграмм, а также непосредственное построение СЕИТ (соединение узлов связями «включения») и ее отображение [11].

Для последовательностей терминов и их весовых значений по TFIDF строятся компактифицированные графы горизонтальной видимости (CHVG) и выполняется повторное определение весовых значений слов уже по этому алгоритму [10]. Данная процедура позволяет учитывать в дальнейшем кроме терминов с большой дискриминантной силой также высокочастотные термины, которые имеют большое значение для общей тематики текстового корпуса. Сеть слов с использованием алгоритма горизонтальной видимости строится в три этапа. На первом на горизонтальной оси отмечается ряд узлов, каждый из которых соответствует словам в порядке появления в тексте, а по вертикальной оси откладываются весовые численные оценки (TFIDF). На втором этапе строится традиционный граф горизонтальной видимости [13]. Для этого между узлами существует связь, если они находятся в «прямой видимости», т.е. если их можно соединить горизонтальной линией, не пересекающей никакую другую вертикальную линию. На третьем, заключительном этапе, сеть компактифицируется. Все узлы с одним и тем же словом объединяются в один узел. Все связи таких узлов также объединяются. Важно отметить, что между любыми двумя узлами при этом остается не более одной связи – кратные связи изымаются. В качестве весовых оценок отдельных слов в дальнейшем используются степени соответствующих им узлов в CHVG. После этого все термины текста сортируются по убыванию рассчитанных весовых значений соответствующих узлов CHVG. Дальнейшему анализу не подлежат термины из так называемого стоп-словаря, являющиеся важными для связности текста, но не несущие информационной нагрузки. Это, как правило, фиксированный набор служебных слов. Используемый в рамках данной работы стоп-словарь был построен на основе различных стоп-словарей, представленных в доступном виде на веб-ресурсах:

https://code.google.com/p/stop-words/downloads/list;

http://www.ranks.nl/stopwords/;

http://www.textfixer.com/resources/common-english-words.txt.

Экспертным методом определяется необходимый размер СЕИТ (число N), после чего выбирается соответствующее количество единичных слов, биграмм и триграмм (всего N+N+N элементов) с наибольшими весовыми значениями по CHVG. Из отобранных терминов строятся сети естественных иерархий терминов, в которых как узлы рассматриваются сами термины, а связи соответствуют вхождением одних терминов в другие. На рис. 6

проиллюстрирован принцип построения связей СЕИТ. Следует отметить, что ранее этот алгоритм применялся к другим видам документов, в частности, докладам тематических конференций и реферативным базам данных [15].

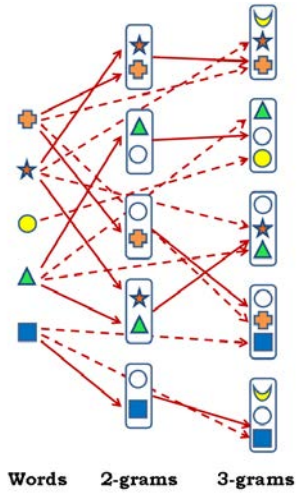



Рис. 6. Формирование связей в трехуровневой сети естественной иерархии терминов

Различные геометрические фигуры на этой иллюстрации соответствуют различным словам. Первой колонке соответствует выбранное множество единичных слов, второй – множество биграмм, а третьей – множество триграмм. Если единичное слово входит в биграмму или триграмму, или биграмма входит в триграмму, образуется связь, которая обозначается стрелкой. Множество узлов, которым соответствуют термины, и связи образуют трехуровневую сеть естественной иерархии терминов [11].

После формирования СЕИТ осуществляется ее отображение программными средствами анализа и визуализации графов. Для загрузки сетей естественных иерархий терминов в базы данным формируется матрица инцидентности общепринятого формата csv.

В таблице 1 приведены списки 20 наиболее весомых терминов (слов, биграмм и триграмм) из новостных сообщений, соответствующих сюжетной цепочке.

На рис. 7 представлена небольшая сеть естественной иерархии терминов размером 20+20+20, которая визуализирована средствами системы Gephi (https://gephi.org/).

На рис. 8 приведен фрагмент более крупной сети естественной иерархии терминов размером 200+200+200.

Для построенных сетей естественных иерархий терминов различных размеров по выбранному тексту было определено распределение исходящих степеней узлов, которое оказалось близким к степенному ( $p(k) = Ck^{\alpha}$ ), т.е. эти сети являются безмасштабными. Оказалось, что коэффициент $\alpha$ для сетей различных размеров (от 20+20+20 до 500+500+500) составляет от 2,1 до 2,3 (рис. 9), что вполне соответствует сетям языка (Language Networks) [1].

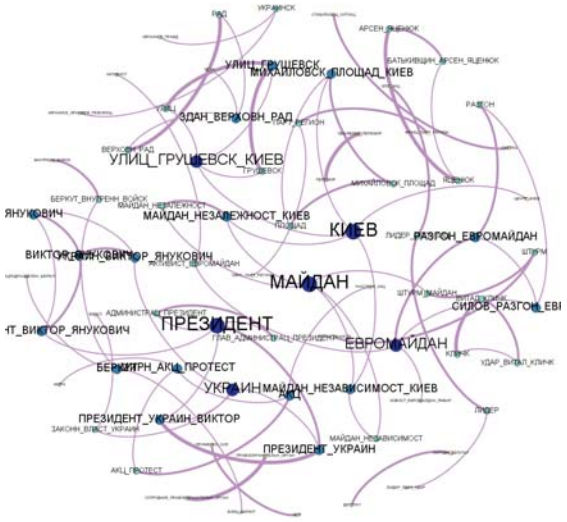



Рис. 7. Вид СЕИТ размером 20+20+20

Очевидно, что в соответствии с предложенным алгоритмом, максимальное количество входных связей для узлов данной сети составляет 5: для узлов из одного слова – 0 входящих связей, для узлов из 2 слов – максимально 2 связи, для узлов из 3 слов – максимально 5 связей – три связи от отдельных слов и две от пар слов. Топ-20 узлов с максимальной входной степенью для СЕИТ размером 200+200+200 приведен в таблице 2.

Наиболее интересными с семантической точки зрения в рассматриваемой СЕИТ оказались узлы с максимальным количеством входных связей, среди которых можно выделить такие словосочетания: «УЧАСТНИКИ АКЦИИ ПРОТЕСТА»; «УЛИЦА ГРУШЕВСКОГО КИЕВ»; «СИЛОВОЙ РАЗГОН ЕВРОМАЙДАНА»; «МИРНАЯ АКЦИЯ ПРОТЕСТА»; «БОЙЦЫ СПЕЦПОДРАЗДЕЛЕНИЯ БЕРКУТ».

По отдельным сюжетам также были рассчитаны значения CHVG для слов, биграмм и триграмм, построены сети естественных иерархий терминов. В качестве примера, отражающего направленность сюжетной цепочки, на рис. 10 приведена визуализация СЕИТ для трех выбранных массивов. Взаимосвязь терминов из новостей, входящих в состав выбранных сюжетов, приведена на рис. 11.

## 7 Релевантность отдельных сюжетов сюжетным цепочкам

На рис. 11 можно видеть, что каждому массиву (узлы, идентифициированные датами) соответствуют термины. При этом в центральной части сети располагаются термины, общие для нескольких дат (О-зона), а «гребешки» на периферии соответствуют специальным терминам, отражающим специфику конкретных сюжетов (С-зоны).

Таблица 1. ТОП-20 по значениям CHVG терминов

| № | Слова | Биграммы | Триграммы |
|---|---|---|---|
| 1 | УКРАИНА | ВИКТОР ЯНУКОВИЧ | ПРЕЗИДЕНТ ВИКТОР ЯНУКОВИЧ |
| 2 | КИЕВ | ЦЕНТР КИЕВА | СОТРУДНИКИ ПРАВООХРАНИТЕЛЬНЫХ ОРГАНОВ |
| 3 | ВЛАСТЬ | ВЕРХОВНАЯ РАДА | ВВЕДЕНИЕ ЧРЕЗВЫЧАЙНОГО ПОЛОЖЕНИЯ |
| 4 | СТРАНА | УЛИЦА ГРУШЕВСКОГО | БАТЬКИВЩИНА АРСЕНИЙ ЯЦЕНЮК |
| 5 | ЯНУКОВИЧ | ПРЕЗИДЕНТ УКРАИНЫ | ОЛИМПИЙСКИЕ ИГРЫ СОЧИ |
| 6 | МАЙДАН | МАЙДАН НЕЗАВИСИМОСТИ | ГЛАВА АДМИНИСТРАЦИИ ПРЕЗИДЕНТА |
| 7 | ЛЮДИ | ПАРТИЯ РЕГИОНОВ | ФРАКЦИЯ ПАРТИИ РЕГИОНОВ |
| 8 | МИЛИЦИЯ | ПРЕСС-СЛУЖБА | ШТАБ НАЦИОНАЛЬНОГО СОПРОТИВЛЕНИЯ |
| 9 | БЕРКУТ | АРСЕНИЙ ЯЦЕНЮК | ДЕЙСТВИЕ БЛАГОДАТИ ПРЕСВЯТОЙ |
| 10 | ОППОЗИЦИЯ | МИХАЙЛОВСКАЯ ПЛОЩАДЬ | МАЙДАН НЕЗАЛЕЖНОСТИ КИЕВ |
| 11 | ПРЕЗИДЕНТ | ЛИДЕРЫ ОППОЗИЦИИ | СТРАНИЦЫ СОЦИАЛЬНЫХ СЕТЕЙ |
| 12 | ЯЦЕНЮК | РАЗГОН ЕВРОМАЙДАНА | УДАР ВИТАЛИЙ КЛИЧКО |
| 13 | УКРАИНСКИЙ | ОБЪЯВЛЕНИЕ ПЕРЕМИРИЯ | ГЕРМАНИЯ ФРАНЦИЯ ВЕЛИКОБРИТАНИЯ |
| 14 | ЕВРОМАЙДАН | ВИТАЛИЙ КЛИЧКО | УЛИЦА ГРУШЕВСКОГО КИЕВ |
| 15 | ШТУРМ | МАЙДАН НЕЗАЛЕЖНОСТИ | ОФИС ПАРТИИ РЕГИОНОВ |
| 16 | АКЦИЯ | АКЦИЯ ПРОТЕСТА | МИХАЙЛОВСКАЯ ПЛОЩАДЬ КИЕВ |
| 14 | ЗДАНИЕ | ПРАВЫЙ СЕКТОР | СИЛОВОЙ РАЗГОН ЕВРОМАЙДАНА |
| 15 | АКТИВИСТ | ОГНЕСТРЕЛЬНОЕ ОРУЖИЕ | БЕРКУТ ВНУТРЕННИЕ ВОЙСКА |
| 16 | МВД | ПРАВООХРАНИТЕЛЬНЫЕ ОРГАНЫ | ПРЕМЬЕР НИКОЛАЙ АЗАРОВ |
| 17 | ПЛОЩАДЬ | ШТУРМ ЗАЧИСТКА | МИРНАЯ АКЦИЯ ПРОТЕСТА |
| 18 | УЛИЦА | ШТУРМ МАЙДАНА | ЗДАНИЕ ВЕРХОВНОЙ РАДЫ |
| 19 | ГРУШЕВСКОГО | ВНУТРЕННИЕ ВОЙСКА | ЗАКОННАЯ ВЛАСТЬ УКРАИНЫ |
| 20 | ЛИДЕР | ПРИМЕНЕНИЕ СИЛЫ | ЛИДЕР ПАРТИИ УДАР |

Таблица 2. Топ-20 узлов с максимальной входной степенью

| № | Выходная степень | Узел |
|---|---|---|
| 1 | 5 | УЧАСТНИКИ АКЦИИ ПРОТЕСТА |
| 2 | 5 | УЛИЦА ГРУШЕВСКОГО КИЕВ |
| 3 | 5 | (ПРЕЗИДЕНТ) УКРАИНЫ ВИКТОР ЯНУКОВИЧ |
| 4 | 5 | СИЛОВОЙ РАЗГОН ЕВРОМАЙДАНА |
| 5 | 5 | МИРНАЯ АКЦИЯ ПРОТЕСТА |
| 6 | 5 | ГЛАВА АДМИНИСТРАЦИИ ПРЕЗИДЕНТА |
| 7 | 5 | ФРАКЦИЯ ПАРТИИ РЕГИОНОВ |
| 8 | 5 | БОЙЦЫ СПЕЦПОДРАЗДЕЛЕНИЯ БЕРКУТ |
| 9 | 5 | БАТЬКИВЩИНА АРСЕНИЙ ЯЦЕНЮК |
| 10 | 4 | АДМИНИСТРАЦИЯ ПРЕЗИДЕНТА УКРАИНЫ |
| 11 | 4 | ЗДАНИЕ ВЕРХОВНОЙ РАДЫ |
| 12 | 4 | ЗДАНИЯ ЦЕНТРА КИЕВА |
| 13 | 4 | ВЕРХОВНАЯ РАДА УКРАИНЫ |
| 14 | 4 | УДАР ВИТАЛИЙ КЛИЧКО |
| 15 | 4 | СОТРУДНИКИ СПЕЦПОДРАЗДЕЛЕНИЯ БЕРКУТ |
| 16 | 4 | СОТРУДНИКИ ПРАВООХРАНИТЕЛЬНЫХ ОРГАНОВ |
| 17 | 4 | СИЛОВОЙ РАЗГОН МИТИНГУЮЩИХ |
| 18 | 4 | ПОЛИТИЧЕСКИЙ КРИЗИС УКРАИНА |
| 19 | 4 | ПРИМЕНЕНИЕ СИЛЫ СТОРОНАМИ |
| 20 | 4 | ПРЕСС-СЛУЖБА МВД |

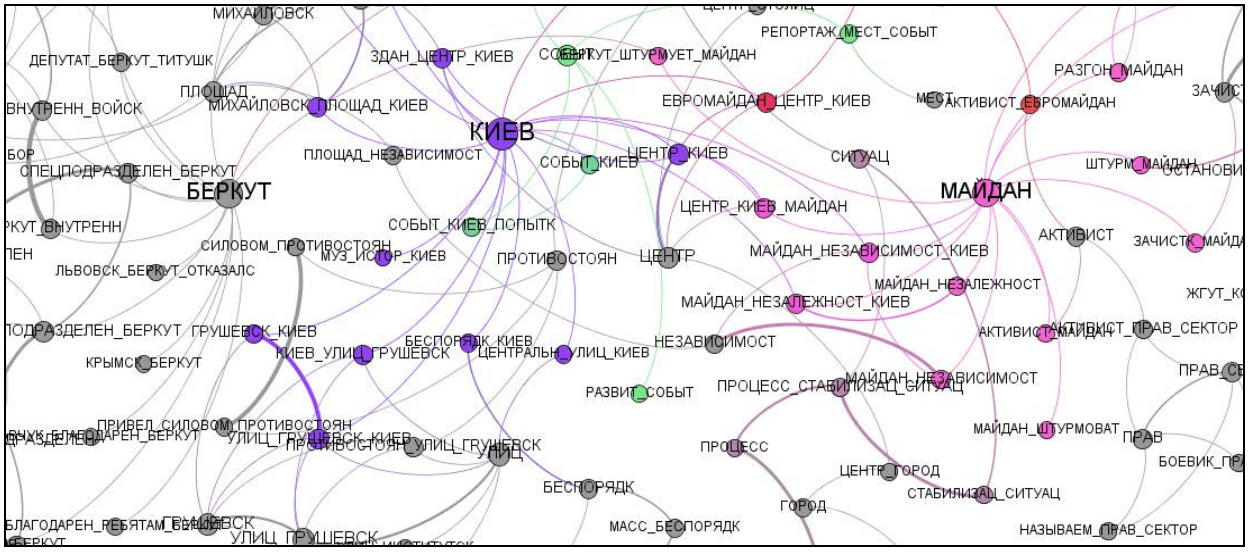



Рис. 8. Фрагмент СЕИТ (визуализация средствами Gephi)

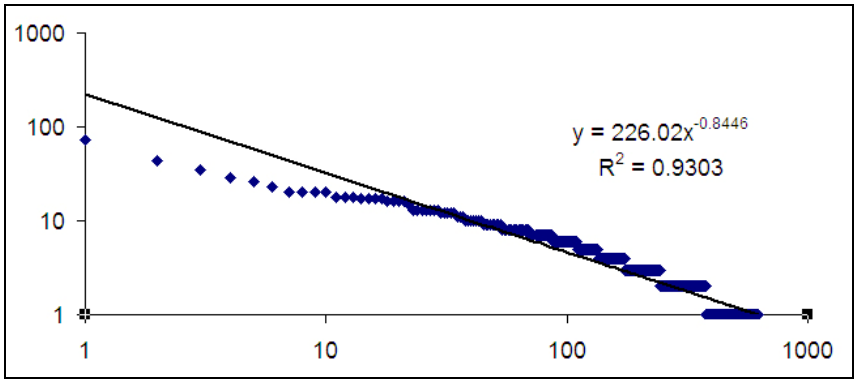
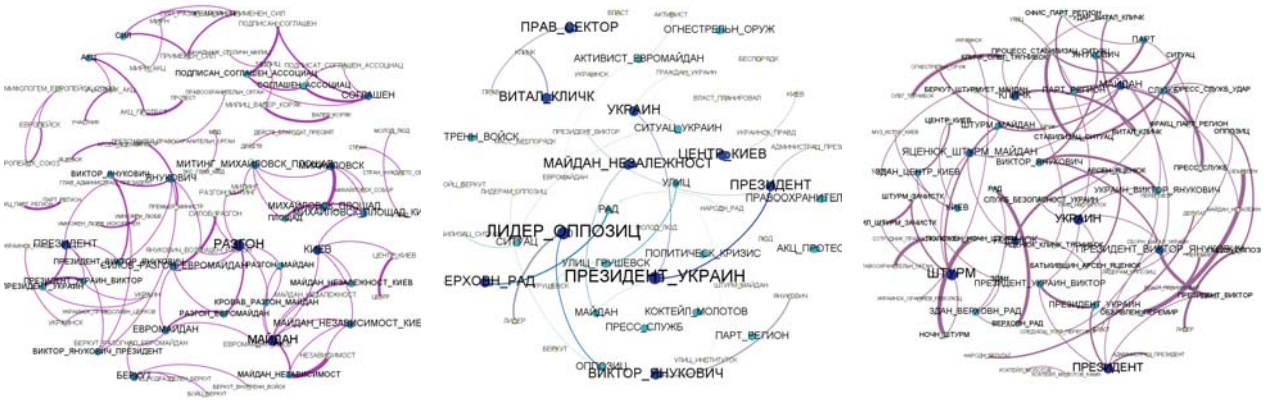



Рис. 9. Ранговое распределение степеней узлов в логарифмической шкале (по оси абсцисс – порядковый номер узла, по оси ординат – степень узла)




а б в

Рис. 10. СЕИТ размером 20+20+20 по массивам (а – 2013.11.30, б – 2014.01.22, в – 2014.02.19)

О-зона не обязательно включает термины из всех сюжетов, достаточно, чтобы термины соответствовали лишь их определенной части (порогу). Чем в сообщениях сюжета больше терминов, попадающих О-зону, тем он лучше вписывается в тематику всей сюжетной цепочки, тем он точнее попадает в ее тренд. В данном случае (рис. 11) именно сюжет 22 января наиболее точно соответствует тематическому направлению всей сюжетной цепочки.

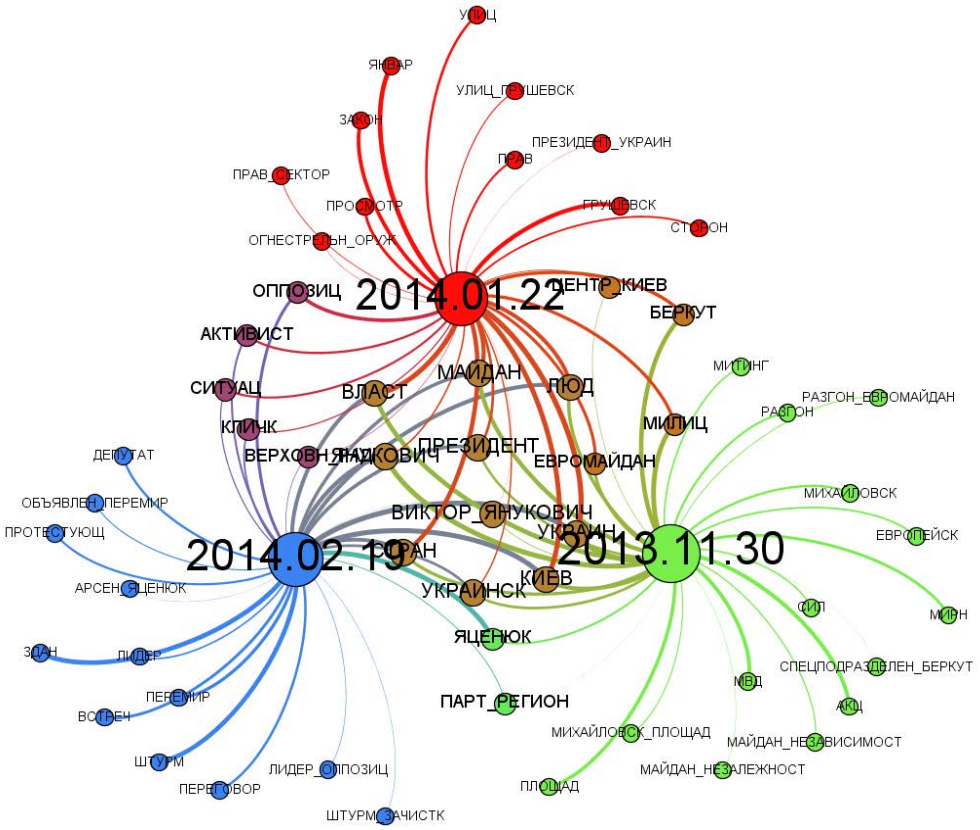



Рис. 11. Сеть связи терминов выбранных событий

Таким образом, можно предложить такой лингвистический критерий релевантности сюжета общей сюжетной цепочке: чем большая часть лексики из него попадает в О-зону, тем он более релевантен. Формально значение этого критерия $k_{i,N}$ для сюжета $i$ сюжетной цепочки $s$ может быть записано:

$$k_{i,N} = \frac{\left|T_{i,N} \cap T_{s,N}\right|}{3N},$$

где $N$ – параметр СЕИТ (количество слов, биграмм и триграмм), $T_{i,N}$ – множество значимых терминов сюжета $i$, $T_{s,N}$ – множество значимых терминов всей сюжетной цепочки.

Представления об информационной значимости наборов терминов для построения СЕИТ, степени их важности для отражения смысла сюжетной цепочки были подтверждены в ходе экспериментов с информантами. Так, для всех сюжетов были проведены эксперименты с вариантами стандартной инструкции «Вспомните сюжет. Подумайте над его содержанием. Выпишите 10-15 слов, наиболее важных для его содержания» (более 20 информантов для каждого сюжета) [7, 8[1]]. При этом количество предложенных экспертами отдельных слов, биграмм и триграмм, характеризующих общий сюжет, совпало с данными табл. 1 на 65, 50 и 45 %, соответственно. Вместе с тем, качество предложенной терминологической сети исследовалось лишь на уровне экспертных оценок, сравнение СЕИТ с другими подобными сетями остается открытым вопросом.

## Выводы

Таким образом, в результате проведенных исследований:

- Описан алгоритм построения СЕИТ на основе анализа текстов новостных сообщений.
- При построении СЕИТ для новостных сообщений был учтен ряд особенностей, связанных с предварительным анализом потока новостей, предложен подход к выбору репрезентативных сюжетов.
- На основании предложенного алгоритма построена СЕИТ (как показала практика, минимальный объем для построения репрезентативной сети составляет около 20 КБ).
- Сеть естественных иерархий терминов оказалась скейл-фри по исходящим связям.
- Выбраны программные средства, позволяющие решать задачу визуализации СЕИТ. При этом задача выбора лучшего алгоритма визуализации не ставилась.
- Предложен критерий релевантности сюжета общей сюжетной цепочке.

[1] Ср. основную инструкцию сходного эксперимента: «Как можно детальнее вспомните события сравнительно недавнего периода, примерно от 4 декабря 2011 г. до 4 марта 2012 г. (от выборов в Государственную Думу до подведения итогов по выборам президента). Напишите 10-15 слов или словосочетаний, относящихся к этим событиям».

Сеть языка, построенную с помощью предложенной методики, можно использовать в качестве базы для построения общей онтологии по выбранной тематике, готового к применению средства навигации в базах данных, а также для организации контекстных подсказок пользователям информационно-поисковых систем.

## Литература


[1] Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Е.И. Большакова, Э.С. Клышинский , Д.В. Ландэ , А.А. Носков, О.В. Пескова, Е.В. Ягунова. – М.: МИЭМ, 2011. – 272 с.

[2] А.А. Давыдов. Системная социология. – М.: Издательство ЛКИ, 2008. – 192 с.

[3] А.Г. Додонов, Д.В. Ландэ. Моделирование и анализ тематических информационных потоков // Информационное противодействие угрозам терроризма, 2013. – № 20. – С. 52–59.

[4] И.В. Крылова, Л.М. Пивоварова, А.В. Савина, Е.В. Ягунова. Исследование новостных сегментов российской «снежной революции»: вычислительный эксперимент и интуиция лингвистов // Понимание в коммуникации: Человек в информационном пространстве: сб. научных трудов: в 3 т. – Яр.-М.: Изд-во ЯГПУ, 2012. – Т. 1. – С. 377–382.

[5] Н.В. Лукашевич, Б.В. Добров, Д.С. Чуйко. Отбор словосочетаний для словаря системы автоматической обработки текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2008». – М., 2008. – С. 339–344.

[6] Ю.Н. Филиппович, А.В. Прохоров. Семантика информационных технологий: Опыты словарно-тезаурусного описания. – М.: МГУП, 2002. – 368 с.

[7] Е.В. Ягунова. Эксперимент и вычисления в анализе ключевых слов художественного текста // Сборник научных трудов кафедры иностранных языков и философии ПНЦ УрО РАН. Философия языка. Лингвистика. Лингводидактика. – Пермь, 2010. – Вып. 1. – С. 85–91.

[8] Е.В. Ягунова, И.В. Крылова, О.Е. Макарова, Л.М. Пивоварова. "Снежная революция в России": значимые номинации, события, оценки (оценка событий информантами и данные СМИ) // "Мы не немы!": творчество протестующей улицы. – М., 2014.

[9] Е.В. Ягунова, А.В. Антонов. Методика работы с коллекциями текстовой информации через анализ информационных портретов // Труды 12-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL’2010, Казань, Россия, 2010.

[10] D.V. Lande, A.A. Snarskii. Compactified HVG for the Language Network // International Conference on Intelligent Information Systems: The Conference is dedicated to the 50th anniversary of the Institute of Mathematics and Computer Science, 20–23 aug. 2013, Chisinau, Moldova: Proceedings IIS / Institute of Mathematics and Computer Science, 2013. – P. 108–113.

[11] D.V. Lande. Building of Networks of Natural Hierarchies of Terms Based on Analysis of Texts Corpora // E-preprint arXiv 1405.6068.

[12] D.V. Lande, A.A. Snarskii, E.V. Yagunova, E.V. Pronoza. The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text // 12th Mexican International Conference on Artificial Intelligence, 2013. – P. 209–215.

[13] B. Luque, L. Lacasa, F. Ballesteros, J. Luque. Horizontal visibility graphs: Exact results for random time series // Phys. Review E, 2009. – P. 046103-1–046103-11.

[14] G. Salton, M.J. McGill. Introduction to Modern Information Retrieval. – New York : McGraw-Hill, 1983. – 448 p.

[15] E. Yagunova, D. Lande. Dynamic Frequency Features as the Basis for the Structural Description of Diverse Linguistic Objects // CEUR Workshop Proceedings. Proceedings of the 14th All-Russian Scientific Conference “Digital libraries: Advanced Methods and Technologies, Digital Collections” – Pereslavl-Zalessky, Russia, 2012. – P. 150–159.


## Network of Natural Hierarchies of Terms of News Messages on the “Euromaydan” Events


Dmitri V. Lande, Andrew A. Snarskii,
Elena V. Jagunova



The technique of building of networks of hierarchies of terms based on the analysis of scientific texts is offered. The technique is based on the methodology of horizontal visibility graphs for the terms – of individual words, bigrams and trigrams, as well as of an inclusion relationships between the terms. The network formed on the basis of news texts on the “Euromaydan” events has been designed and investigated.

# Об автоматической рубрикации терминов тезауруса открытой информационно-аналитической системы

© Бойков В.Н. © Захаров В.Е.
Институт космических исследований РАН
Москва
boykov_bh@bk.ru vezakhar@mx.iki.rssi.ru

© Каряева М.С. © Соколов В.А.
Ярославский государственный университет
Ярославль
mari.s.ka@mail.ru valery-sokolov@yandex.ru

**Аннотация**

В работе рассматривается применение методов лингвистического анализа для автоматического рубрицирования терминов открытого сетевого ресурса «Тезаурус по поэтологии» в составе Информационно-аналитической системы русской поэзии. Приведены основные принципы и процедуры автоматической рубрикации корпуса терминов тезауруса.

Работа поддержана Российским фондом фундаментальных исследований, грант № 13-06-00448.

## 1 Введение

Автоматизация семантического анализа полнотекстовой информации для извлечения релевантных данных является актуальной задачей в инженерии знаний. Это важно и для автоматического построения таких метаописаний и семантических структур предметной области, как тезаурусы и онтологии, где описываются основные понятия и отношения между ними.

Семантическая модель предметно-ориентированного тезауруса предопределяет структуру его данных, тогда как его структурирование осуществляется по мере непосредственного описания его понятий и отношений между ними. Определяющими структуру данных являются иерархические отношения, поскольку задают сложность структуры, как по числу иерархических уровней (рубрик), так и по числу вариаций (подрубрик) на одном уровне. Среди иерархических отношений чаще всего выделяются отношение «рода и видов», задающее классификацию понятий, и отношение «целого и частей», систематизирующее данные.

Задача автоматизации описания понятия, основу которого составляет его определение, решается с помощью известных методов полнотекстового поиска из авторитетных источников (баз знаний, справочников, словарей, энциклопедий).

Задача автоматического выявления отношений между понятиями сводится к извлечению этих отношений из описания понятий и требует специальных лингвистических методов, поскольку чаще всего формально эти отношения в описании не задаются. Для выявления таксономических отношений (синонимии и гиперонимии) между понятиями оказывается эффективным формирование лингвистических шаблонов [1]. Растущий интерес исследователей к лингвистическим методам анализа текста для построения онтологий связан с повышением качества синтаксических анализаторов [2].

Семантическая неопределенность сложных синтаксических конструкций заставляет обратиться к такому фундаментальному семантико-синтаксическому понятию, как синтаксема, которая является и минимальной синтаксической единицей, и носителем элементарного смысла [3]. Синтаксемы могут происходить от различных частей речи, но преимущественно – от имен существительных и представляют собой падежные или предложно-падежные словоформы в синтаксическом контексте. К сожалению, конструкции со сложными предлогами в репертуар синтаксем пока не вошли.

Примером конструктивного использования синтаксемы может служить лингвистический подход, осуществленный в [4] при построении онтологии конкретной предметной области.

## 2 Описание объекта структурирования

Представление о составе Информационно-аналитической системы русской поэзии (ИАСРП), а также о методологических концепциях тезауруса по поэтологии (ТП) дается в работах [5–8].

В исходной комплектации ТП содержится:

– Базовый корпус около 2000 терминов по поэтологии;

– Формуляр терминологической статьи тезауруса (ТСТ), представляющий семантическую модель ТП [8], где заданы 27 полей ТСТ трех типов – поля,

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

относящиеся к термину (основные – «определение» и «рубрика»), и поля иерархических и неиерархических отношений между терминами;

– Базовый рубрикатор терминов, в котором представлена экспертная рубрикация предметной области по 10 подобластям (рубрикам верхнего уровня), соответствующим дисциплинам поэтологии [5], что необходимо и достаточно для автоматической рубрикации всего корпуса терминов:

1. Стиховедение;
2. Стилистика;
3. Поэтика;
4. Риторика;
5. История литературы;
6. Переводоведение и литературная компаративистика;
7. Текстология;
8. Герменевтика;
9. Теоретические школы и направления;
10. Логика и методология науки.

В качестве примера произведена ручная рубрикация 32 терминов 4 верхних уровней подобласти 1. Стиховедение:

1.1. Стих: (кластер);
1.1.1 Метрика: (кластер);
1.1.1.1. Квантитативная метрика: (кластер);
1.1.1.2. Квалитативная метрика: (кластер);
1.1.2. Явления начала и конца стихотворной строки: (кластер);
1.1.2.1. Анакруза, анакруса: (кластер);
1.1.2.2. Каталектика: (кластер);
1.1.3. Ритмика: (кластер);
1.1.3.1. Акцентуация: (кластер);
1.1.3.2. Цезура и Словоразделы: (кластер);
1.1.4. Строфика: (кластер);
1.1.4.1. Строфы: (кластер);
1.1.4.2. Квазистрофические формы и Гиперстрофические формы: (кластер);
1.1.4.3. Твёрдые формы стиха: (кластер);
1.1.5. Рифмика: (кластер);
1.1.5.1. Типы рифмы по количеству слогов: (кластер);
1.1.5.2. Типы рифмы по фонетическому составу: (кластер);
1.1.5.3. Типы рифмы по лексическому составу: (кластер);
1.1.5.4. Рифменные последовательности: (кластер);
1.1.5.5. Квази-рифмические способы организации стиха: (кластер);
1.1.6. Лингвистика стиха: (кластер);
1.1.6.1. Звуковая организация стиха: (кластер);
1.1.6.2. Графическая организация стиха: (кластер);
1.1.6.3. Ритмико-фонетические явления в стихе: (кластер);
1.1.6.4. Морфология стиха: (кластер);
1.1.6.5. Синтаксис стиха: (кластер);
1.1.6.6. Мелодика стиха: (кластер);
1.1.6.7. Поэтическая семантика: (кластер);
1.2. Проза (в отличие от стиха): (кластер);
1.2.1. Формы прозы: (кластер);
1.2.2. Членение прозы: (кластер).

Рубрики 4 верхних уровней всех 10 подобластей представляют всего 115 терминов из двухтысячного корпуса.

ТП разрабатывается с применением Wiki-технологии. Базы знаний с использованием Wiki-технологий имеют ряд преимуществ, так как позволяют энтузиастам-исследователям самим через веб-интерфейс активно включиться в процесс редактирования базы знаний: исправления ошибок, добавления новых материалов и т.д. Коллективное редактирование ТП может ускорить наполнение ТСТ и не должно отразиться на его качестве, поскольку добавление новой информации в ТП отслеживает наряду с администратором сайта модератор системы – квалифицированный специалист в области поэтологии, который принимает или отвергает внесение или изменение контента в ТП.

Вместе с тем, несмотря на возможность получения высокого качества при ручном заполнении ТСТ, трудоемкость и множественность звеньев процесса не обеспечивает его должной скорости, поэтому задача его автоматизации на предварительном этапе представляется достаточно важной. Такая автоматизация позволяет осуществить структурирование предметной области и вследствие чего дает возможность энтузиастам-исследователям завершить описание термина в контексте его места в общей структуре ТП.

ИАСРП в своем составе должен содержать помимо ТП также аналитический блок [6], который предназначен для автоматического решения различных задач стиховедения в отношении поэтических текстов. Для постановки и алгоритмизации этих задач необходим завершенный в достаточной полноте тезаурус, что предполагает, в том числе, и его рубрикацию. В этом контексте очевидна актуальность создания программно-алгоритмического модуля для решения комплекса задач, связанных со структуризацией ТП и рубрикации его терминов.

## 3 Модуль автоматического структурирования ТП

Конечной целью автоматического структурирования ТП является рубрикация его терминов, т.е. отнесение каждого термина к его рубрике в иерархическом дереве, точнее, к цифровому коду его рубрики, идентифицирующему место термина в иерархии. В данном случае каждому термину определяется место в цепочке терминов, привязанной к одной из вершин базового рубрикатора терминов.

В модуле автоматического структурирования ТП выделяются следующие подмодули последовательных автоматических процедур.

Подмодуль 1: автоматическое заполнение поля ТСТ «определение»;

Подмодуль 2: автоматическое заполнение полей ТСТ «родовое понятие» и «видовые понятия»;

Подмодуль 3: автоматическое заполнение полей ТСТ «целое» и «части»;

Подмодуль 4: автоматическое заполнение полей «рубрика» и «дисциплина (рубрика первого уровня)».

**Процедуры подмодуля 1**

Для автоматического заполнения поля ТСТ «определение» используются следующие (в порядке репрезентативности) оцифрованные источники:

– Краткая литературная энциклопедия: В 9 т. – М.: Сов. энцикл., 1962-1978 [9].

– Квятковский А.П. Поэтический словарь. – М.: Советская энциклопедия, 1966 [10].

– Литературная энциклопедия: В 11 т. – М.: Ком. акад., 1929-1939 [11].

Дополнительно полезны также некоторые другие энциклопедии и словари [12–16].

Ключом для извлечения определения понятия из источника служит термин из имеющегося терминологического словника ТП. Все извлеченные определения для данного термина помещаются в поле ТСТ «альтернативные определения».

Затем производится лингвистическая (частеречная и синтаксическая) разметка текстов определений с помощью синтаксического анализатора (парсера), размещенного на электронном ресурсе «Автоматическая обработка текста» [17]. При разметке текста определения термина в нем отмечаются другие термины, включенные в терминологический словник, что важно для результативности процедур последующих подмодулей.

Среди альтернативных определений могут оказаться синонимические определения, а также противоречащие друг другу и нечеткие определения. Решение о помещении того или иного определения в поле ТСТ «определение» и сохранении его в поле «альтернативные определения» принимается модератором системы.

**Процедуры подмодуля 2 и 3**

Хотя для рубрикации достаточно заполнить поле «родовое понятие», но иногда род термина определяется только через его представление в качестве вида другого.

Первой процедурой выявления рода для данного термина является его поиск в полях «видовые понятия» в соответствующих полях других терминов: его нахождение в поле «видовые понятия» некоторого термина означает, что последний и является родом для данного.

Аналогичной процедурой выявления вида для данного термина является его поиск в полях «родовое понятие» ТСТ других терминов: его нахождение в поле «родовое понятие» некоторого термина означает, что последний является видом для данного.

Автоматизация выделения из определения термина его рода и видов с помощью лингвистических методов исходит из выявления синтаксических конструкций, задающих отношения рода и вида. Элементарные единицы русского синтаксиса (синтаксемы) для этих отношений приводятся в [3], хотя они не исчерпывают всех синтаксических конструкций для этих целей. Примеры синтаксем, несущих отношения «род-виды»:

– предмет среди класса предметов – **предлог «среди»** + **род. падеж** (выделяться, находиться среди …);

– отнесение вида к роду – **предлог «к»** + **дат. падеж** (относиться, принадлежать к …).

Для выявления «родо-видовых» отношений служат и другие синтаксические конструкции [18].

Для выявления вида:

– сложное слово, часть которого (производящая основа или словообразующая морфема) задает единство принадлежности к роду, как, например, в «метрике» различают явления «монометрии» и «полиметрии».

Для выявления как рода, так и вида:

– словосочетание, представляющее собой видовой термин, где «в качестве опорного терминоэлемента выступает родовой термин», как, например, в роду «ямб» выделяются двустопный, трехстопный, 4-стопный, 5-стопный и 6-стопный ямбы.

Для выявления «родо-видовых» отношений полезны также конструкции с предметно определенным обобщающим словом или словосочетанием при однородных членах предложения.

После выявления «родо-видовых» отношений данного термина определяющая эти отношения синтаксическая конструкция добавляется в набор шаблонов для синтаксического анализа последующих терминов.

При успешном выявлении рода или вида для данного термина может оказаться так, что соответствующих им терминов в терминологическом словнике нет, и тогда решение о внесении этих терминов в словник принимает модератор системы.

С другой стороны, в самих определениях терминов могут не найтись отсылки к роду и видам (у конечных терминов виды отсутствуют), и, следовательно, не всегда можно вывести родо-видовые цепочки к 4 верхним уровням, имеющим коды рубрик. В этом случае придется использовать открытость системы и компетентность энтузиастов-исследователей предметной области.

Процедуры выявления отношений «целое-части» для данного термина осуществляются по аналогии с предыдущими.

Специфика выделения из определения термина отношений «целое-части» с помощью лингвистических методов отличается более широким набором синтаксем, используемых для выявления этих отношений:

– обозначение частей целого – **предлог «из»** + **род. падеж** (состоять, слагаться, складываться, составляться, собираться или образовываться из …);

– часть, отделенная от целого – **предлог «от»** + **род. падеж**;

– дополнение части к целому – **предлог «к»** + **дат. падеж** (приобщенное к чему-то);

– соединение частей в целое – **предлог «в»** + **вин. падеж** (складываться, собираться в …);

– распадение целого на части – **предлог «в»** + **вин. падеж** (распадаться в …);

– деление целого на части (несколько частей) – **предлог «на»** + **вин. падеж**.

**Процедуры подмодуля 4**

Нахождение данного термина в одном из кластеров рубрики верхнего уровня определяет процедуру заполнения поля «дисциплина» в его ТСТ.

После выявления «родо-видовых» отношений данного термина его «родовое понятие» сверяется с терминами БРТ, имеющими код рубрики, и при его совпадении с одним из таких терминов БРТ, производится рубрикация данного термина и его перевод из кластера в рубрикатор: данному термину присваивается видовой код рубрики найденного термина. Затем код рубрики данного термина вносится в поле «рубрика» его ТСТ.

Далее производится рубрикация «видовых понятий» данного термина и их перевод из кластера в рубрикатор: им присваиваются видовые коды рубрики данного термина. Затем заполняются поля ТСТ видовых понятий данного термина «рубрика» и «родовое понятие», куда вносится данный термин.

## 4 Реализация автоматического заполнения поля ТСТ «определение»

На рисунке 1 для более наглядного представления размещены первые 12 полей ТСТ для конкретного термина. Следует отметить, что заполнение ТП является не только трудоемким процессом, так как необходимо заполнить порядка 50 тысяч полей, но и требует достаточного уровня знаний предметной области.

На рисунке 2 показана схема разметки ТСТ источника, где ее текст открывается термином, который содержится в терминологическом словнике и, следовательно, в БРТ, что позволяет видеть, к какой рубрике верхнего уровня относится данный термин. Соответственно, в ТП автоматически создается новая ТСТ с данным термином и заполняются поля «термин» и «дисциплина».

Далее существует 2 варианта объяснения термина: первый вариант содержит иностранный эквивалент с указанием языка и перевода термина, второй вариант встречается при условии русского происхождения термина или утраты его иностранного происхождения.

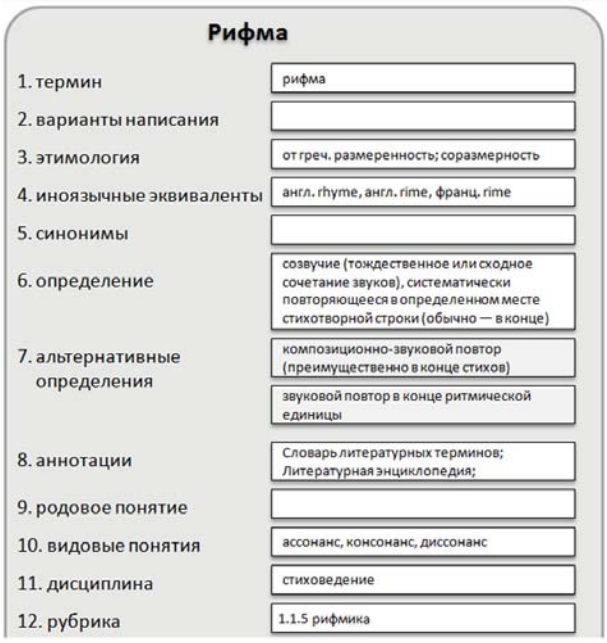



Рис. 1. Пример термина «Рифма» с заполненными полями

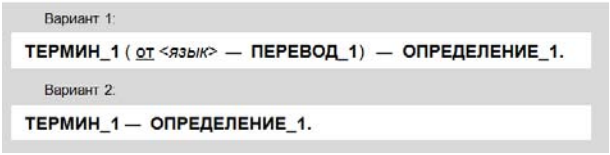



Рис. 2. Разметка «термин-определение».

Так как одно из полей тезауруса содержит поле «иноязычные эквиваленты», то использование конструкции, представленной на рисунке 2, снимает вопрос о заполнении этого поля. Разметка <*язык*> представляется в тезаурусе как дополнительный технических словарь, содержащий набор различных языков в виде «англ.», «нем.», «греч.», «франц.», «араб.» и т.д. Ключевое слово и соответствующая разметка «от» + <*язык*> автоматически указывают на конструкцию строки, которая может быть использована в качестве описания поля «иноязычные эквиваленты».

При использовании варианта 2 разметки терминологической статьи источника поле «иноязычные эквиваленты» остается пустым. Далее, для всех вариантов выделяется оставшаяся часть предложения и оформляется как определение термина.

Пример разметки статьи источника [8] и выделения из нее определения термина «рифма»:

**Рифма** (от греч. — соразмерность) — композиционно-звуковой повтор преимущественно в конце двух или нескольких стихов, чаще — начиная с последнего ударного слога в рифмуемых словах. В русских пиитиках (10—18 вв.) этот повтор назывался «краесогласием».

**ТЕРМИН_1** := рифма

**ПЕРЕВОД_1 :=** соразмерность

**ОПРЕДЕЛЕНИЕ_1 :=** композиционно-звуковой повтор преимущественно в конце двух или нескольких стихов, чаще — начиная с последнего ударного слога в рифмуемых словах.

Каждый из оцифрованных источников, перечисленных выше [9–11], содержит описания и определения не более одной тысячи терминов. Безусловно, литературоведческих справочников недостаточно для полного покрытия предметной области, так как представленные выше источники содержат целый ряд идентичных терминов, что уменьшает размерность общего набора различных терминов. Поэтому неизбежно использование альтернативных и универсальных источников знаний, таких как словари и энциклопедии общей направленности.

## 5 Статусы терминов после автоматического структурирования ТП

После применения процедур автоматической рубрикации все термины ТП условно можно разделить на три группы:

1) Термины с заполненными полями ТСТ, в том числе с полями, определяющими отношения между терминами.

2) Термины с заполненными полями ТСТ, кроме полей, определяющих отношения между терминами.

3) Термины, которые не встретились в литературоведческих и стиховедческих источниках, соответственно, не имеют автоматически заполненных полей ТСТ.

Первая группа терминов является завершенной и имеет статус «Завершено», вторая группа терминов получает статус «В работе» и доступна для дальнейшей автоматической рубрикации, пока не приобретет статус «Завершено». Третья группа терминов имеет статус «Не определено», который показывает, что термин не подвергался автоматической рубрикации ввиду отсутствия соответствующего термина в оцифрованных источниках. Определение этих статусов указывает, что следует прибегнуть к ручному заполнению ТСТ для терминов со статусом «Не определено».

Основные три указанных источника терминов [9–11] содержат порядка 1000 терминов каждый и, если учесть далеко не полное пересечение этих совокупностей терминов, то в целом корпус терминов со статусами «Завершено» и «В работе» может составить около 1500 единиц.

Данный результат был получен с помощью статистических расчетов, проведенных вручную, что гарантирует качество проводимых исследований и может быть достоверной верхней оценкой для автоматического подхода. Экспертная ручная оценка обеспечивает возможность получения близкого результата, поскольку используемые правила могут быть алгоритмизированы. Это дает основание считать, что метод может быть автоматизирован на практике и в дальнейшем улучшен с помощью обучающих методов (в том числе, используя машинное обучение).

## 6 Заключение

В работе представлены методы и процедуры, которые позволяют автоматически структурировать термины такой предметной области, как «поэтология». Благодаря этому для заполнения полей ТСТ не требуется помощь специалиста, необходимо лишь реализовать алгоритмы, использующие оцифрованные литературоведческие и стиховедческие источники. В дальнейшем, при совершенствовании лингвистических методов анализа, детально рассматривающих частные случаи и исключения и использующих наряду со справочными источниками литературоведческие исследования, можно автоматизировать заполнение полей ТСТ и подвергнуть автоматической рубрикации значительную часть группы терминов, имеющих статус «Не определено».

Таким образом, имея достаточный набор неструктурированных терминов, источники знаний и ряд аналитически полученных правил, можно осуществлять автоматическую рубрикацию терминов. Кроме того, такой подход к структуризации предметной области может использоваться в более широком аспекте гуманитарного знания.

## Литература


[1] M.A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora // Proceedings of the 14th International Conference on Computational Linguistics. – 1992. – P 539–545.

[2] J. Makki, A.-M. Alquier, V. Prince. Semi Automatic Ontology Instantiation in the domain of Risk Management // IFIP, Advances in Information and Communication Technology. – 2008. – Vol. 288. – P. 254–265.

[3] Г.А. Золотова. Синтаксический словарь. Репертуар элементарных единиц русского синтаксиса. – М.: Наука, 1988.

[4] Е.А. Оробинская. Метод автоматического построения онтологии предметной области на основе анализа лингвистических характеристик текстового корпуса // Труды XV Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2012). – СПб, 2012.

[5] В.Н. Бойков, В.Е. Захаров, И.А. Пильщиков, Т.М. Сысоев. Тезаурус как инструмент поэтологии // Моделирование и анализ информационных систем. – 2010. – Т. 17, № 1. – С. 5–24.

[6] V.N. Boikov, V.E. Zakharov, M.S. Karyaeva, V.A. Sokolov. Предметно-ориентированный тезаурус в открытой информационно-аналитической системе (Domain-Specific Thesaurus as a Part of an Information-Analytical System) – RCDL-2013.

[7] В.Н. Бойков, В.Е. Захаров, М.С. Каряева, В.А. Соколов. Тезаурус по поэтологии как инструмент для информационного поиска и коллекции знаний // Моделирование и анализ информационных систем. – 2013. – Т. 20, № 4. – С. 125–135.

[8] Бойков В.Н., Пильщиков И.А. Семантическая модель «Тезауруса по поэтологии» в составе информационно-аналитической системы // Интернет и современное общество: сборник научных статей. Труды XVI Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2013). — СПб.: НИУ ИТМО, 2013.

[9] Краткая литературная энциклопедия: в 9 т. – М.: Сов. энцикл., 1962–1978. (http://feb-web.ru/feb/kle/default.asp?/feb/kle/kle.html)

[10] А.П. Квятковский. Поэтический словарь. – М.: Советская энциклопедия, 1966. (wikilivres.ru); (feb-web.ru/feb/kps/kps-abc)

[11] Литературная энциклопедия: в 11 т. – М.: Ком. акад., 1929–1939. (http://feb-web.ru/feb/litenc/encyclop/)

[12] Литературная энциклопедия. Словарь литературных терминов: в 2 т. – М., Л.: Изд-во Л.Д. Френкель, 1925. (enc-dic.com/lit)

[13] Большая советская энциклопедия: в 30 т. – 3-е изд. – М.: Сов. энцикл., 1969–1978. (http://slovari.yandex.ru/dict/bse/)

[14] Лингвистический энциклопедический словарь. М.: Советская энциклопедия, 1990. (www.tapemark.narod.ru/les)

[15] Ахманова О. С. Словарь лингвистических терминов. – М.: Сов. энцикл., 1966.

[16] Розенталь Д. Э., Теленкова М. А. Словарь-справочник лингвистических терминов. – Изд. 2-е. — М.: Просвещение, 1976. (http://www.intruderalarms.sebastopol.ua.); (http://www.gumer.info/bibliotek_Buks/Linguist/DicTermin/index.php)

[17] Автоматическая обработка текста. [Электронный ресурс] // Режим доступа: http://aot.ru/demo/morph.html

[18] Н.А. Гурдаева. Принципы структурной организации лексических терминов как результат родо-видовых отношений системы понятий // Вестник ТГПИ. Специальный выпуск 1. Таганрог, 2011.

## On the Automatic Structuring of the Thesaurus for an Open Information-Analytical System

V.N. Boikov, V.E. Zakharov,
M.S. Karyaeva, V.A. Sokolov

In the work methods of using the linguistic analysis for the automatic structuring of the open network resource “Information-Analytical System of Russian Poetry” are considered. The basic principles that allow to realize a way of the automatic categorization of the thesaurus are given.

# Научные коммуникации на базе электронных библиотек с онлайновой декларацией семантических связей

© М.Р. Когаловский
Институт проблем рынка РАН

© С.И. Паринов
Центральный экономико-математический институт РАН

Москва

kogalov@gmail.com

sparinov@gmail.com

**Аннотация**

Электронная библиотека, обладающая средствами декларации семантических связей между ее информационными объектами в онлайновом режиме, может также служить платформой для новых форм научных коммуникаций в сообществе ее пользователей. Такие информационные системы, созданные для научного сообщества, отличаются от традиционных. Они предоставляют пользователям возможности: явного выражения научных отношений между результатами исследований; навигации по структуре семантических связей информационных ресурсов; новых более содержательных наукометрических исследований, основанных, в частности, на ссылках цитирования, которые несут информацию о его мотивах. В докладе обсуждаются возможности таких систем, позволяющие улучшить традиционные научные коммуникации между авторами и пользователями результатов научных исследований благодаря новым средствам прямого электронного уведомления об актах коммуникаций и последующим ответным реакциям исследователей. Рассматриваются также некоторые особенности реализации предлагаемого подхода в среде действующей системы электронной библиотеки. Поддержка прямых научных коммуникаций между авторами и пользователями результатов исследований на основе электронной библиотеки может существенно улучшить кооперацию и координацию исследовательской деятельности для глобального научного сообщества.

Работа поддержана РГНФ, проект 14-07-12010-в.

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

## 1 Введение

В деятельности научного сообщества ключевую роль играют коммуникации между его представителями. Главная цель научных коммуникаций заключается в обеспечении глобального взаимодействия ученых, включая передачу знаний, полученных в результате исследований, и их практическую проверку. За многовековую историю науки сложился ряд форм научных коммуникаций, как прямых, так и опосредованных, например, через научные издательства. Прямые коммуникации представляют собой личные контакты ученых, включая выступления с докладами и обмен мнениями на научных мероприятиях, в процессе защиты диссертаций и т.д. Традиционными формами опосредованных научных коммуникаций являются: (а) публикация результатов исследований в научных изданиях; (б) восприятие идей опубликованной работы и ее цитирование; (в) рецензирование опубликованных или готовящихся к публикации работ и научных проектов; и др.

Создание Интернет и доступных в глобальной среде информационных сервисов привело к активной модернизации инфраструктуры и технологии научных коммуникаций. Новые информационные технологии позволили осуществлять традиционные виды как прямых, так и опосредованных научных коммуникаций значительно более оперативно и с вовлечением значительно более широкого круга заинтересованных ученых.

Одной из основ для осуществления новых форм научных коммуникаций стали работы по семантическому структурированию контента научных электронных библиотек. К этой области относится и разрабатываемый в последние годы проект авторов данной статьи. Главными отличительными особенностями развиваемой в рамках этого проекта действующей электронной библиотеки Соционет [7] являются: возможность декларации семантических связей на основе поддерживаемой таксономии связей *в онлайновом режиме*; представление семантических связей как полноценных информационных объектов (first-class object) специального типа; наличие в системе специального сервиса уведомления авторов информационных объектов – участников

связей о событиях в системе, касающихся рассматриваемых связей – их создания, изменения их свойств, их удаления. Перечисленные особенности обеспечивают реализацию новых форм прямых и опосредованных виртуальных научных коммуникаций между зарегистрированными пользователями системы и авторами представленных в ней информационных объектов.

В соответствии с развиваемой авторами концепцией электронной библиотеки с семантическими связями [2–4, 6, 14–16], сервис системы после каждого акта создания связей посылает электронные уведомления авторам (создателям) связываемых объектов [17]. В случае, когда исследователь в процессе самоархивирования созданной им научной публикации в электронной библиотеке декларирует семантические связи для регистрации мотиваций, побуждающих его использовать научные результаты уже существующих публикаций (например, причины цитирования определенных материалов), авторы использованных (процитированных) материалов оперативно получают сведения не только об этом факте, но и о характере использования их научных результатов. В ответ на полученные уведомления авторы использованных материалов могут реагировать различным образом. Например, они могут поддержать или опротестовать характер использования их результатов исследований, разъяснить как правильно их использовать, дать публичную позитивную или негативную оценку этому факту и т.п.

Электронная библиотека, в которой авторы опубликованных в ней работ получают информацию о том, каким образом результаты их труда используют другие ученые, поддерживает, в частности, следующие реакции авторов и формы адресных научных коммуникаций между авторами и потребителями результатов научных исследований:

(1) автор использованного материала помогает ученому – потребителю его научных результатов повысить качество работы последнего, например, консультируя его, как полнее и правильнее использовать соответствующий результат;

(2) автор использованного материала дорабатывает свой материал и предоставляет его потребителям более качественный научный продукт;

(3) автор использованного материала выражает несогласие (протест) с неправильным толкованием или применением его результата, чтобы обратить внимание на проблемы с качества материала, использующего его результаты; и т.д.

Важно отметить, что перечисленные сценарии прямых научных коммуникаций в значительной мере решают проблему контроля качества научных результатов, которая традиционно считается самым серьезным препятствием в развитии средств самоархивирования («самиздата») результатов исследований.

На наш взгляд, развитие и распространение на практике адресных научных коммуникаций, подобных описанным выше, являются чрезвычайно важными для научного сообщества и могут радикально изменить в лучшую сторону степень кооперации и уровень координации в научном сообществе, а также и эффективность научной системы в целом.

Данная статья посвящена обсуждению предлагаемого авторами подхода к развитию новых форм научных коммуникаций в онлайновой среде электронной библиотеки, а также представляет первые результаты в его реализации.

Остальная часть статьи организована следующим образом. В разд. 2 дается общая характеристика предлагаемого подхода к созданию среды поддержки новых форм научных коммуникаций. В качестве базовой среды его реализации используется система Соционет [7]. В разд. 3 рассматриваются принципы организации контента этой системы и способы представления семантических связей между информационными объектами. В разд. 4 обсуждаются используемая онтология связей, средства ее представления и поддержки в системе. В разд. 5 рассматриваются наиболее распространенные формы научных коммуникаций в среде электронной библиотеки, осуществление которых возможно с помощью технологии, реализующей обсуждаемый подход. Разд. 6 посвящен описанию механизмов системы Соционет, используемых для обеспечения новых форм виртуальных научных коммуникаций. В заключении подводятся итоги обсуждения.

## 2 Общая характеристика подхода и среда его реализации

В ряде зарубежных работ (например, в [10, 19, 21]) предлагаются технологии, которые позволяют устанавливать связи между информационными объектами библиотеки с явным образом декларированной семантикой (*семантические связи*). Аналогичная технология разработана авторами данной статьи [2-4, 6, 14-16]. Однако, в отличие от других известных проектов, в нашем подходе семантические связи могут декларироваться авторизованными пользователями (зарегистрированными в системе) в онлайновом режиме. Важно при этом, что декларированные семантические связи представляются в системе как обычные информационные объекты. Тем самым они могут сами являться участниками семантических связей.

Благодаря семантическим связям могут декларироваться научные отношения различного рода между представленными в электронной библиотеке информационными объектами, например, между персонами и публикациями, между публикациями, между организациями и персонами и др. Семантика устанавливаемых отношений определяется на основе онтологии связей, поддерживаемой в электронной библиотеке.

Важно здесь отметить что, по существу, создание в онлайновом режиме семантической связи некоторого класса воплощает в среде научной электронной библиотеки некоторый *акт коммуникации* автора создаваемой связи как с сообществом пользователей электронной библиотеки, так и персональной коммуникации с автором целевого, а иногда и исходного объектов связи (*см. разд.* 3). Содержание переда-

ваемого сообщения представляется при этом классом/подклассом связи, а также возможно комментарием автора связи. Коммуникация с автором целевого объекта имеет место, когда создается семантическая связь между профилем автора этой связи и некоторой публикацией. Например, это может быть связь какого-либо оценочного класса, выражающая одобрение целевой публикации связи, негативное к ней отношение и др. Другой случай – коммуникация между автором создаваемой связи и авторами исходной и целевой публикаций этой связи – возникает, например, при создании связи между двумя публикациями. Важно при этом, что авторы публикаций – участников создаваемой связи – в обоих случаях уведомляются о создании связи с их публикациями. Они могут не только ознакомиться с параметрами связи, но и отреагировать на неё. Конечно, персональная коммуникация в виде уведомления авторов публикаций – участников созданной связи – возможна в том случае, если в их профилях, созданных в системе при их регистрации, указан адрес электронной почты. Акт коммуникации с авторами публикаций-участников связи порождается также и в случае, когда эта связь уже существует и изменяются значения некоторых ее атрибутов, прежде всего атрибута, указывающего класс связи.

Реализация рассматриваемого в данной статье подхода осуществлена в среде крупной научной информационной веб-системы Соционет [7], эксплуатируемой в академическом научно сообществе более полутора десятилетий. Хотя авторы системы не называют Соционет электронной библиотекой, фактически она относится также и к этому классу информационных систем.

Соционет поддерживает для авторизованных пользователей возможности своего рода социальной сети [5], позволяющей пользователям декларировать семантические связи между представленными в системе информационными объектами (публикациями различных видов, наборами данных, персонами в различных ролях, организациями и др.). Сформирована таксономия семантических связей, определяющая семантику научных отношений между информационными объектами системы и базирующаяся на разработанной онтологии связей. Таксономия связей поддерживается механизмами системы в виде контролируемых словарей классов связей. На основе динамически развиваемой семантической структуры контента системы Соционет созданы средства генерации наукометрических данных, более информативных по сравнению с традиционно используемыми на практике. В системе обеспечивается также контекстная визуализация установленных связей и семантическая визуальная навигация по структуре контента для пользователей. Для поддержки научных коммуникаций в среде системы предусматривается мониторинг изменений структуры связей - создание новых связей, удаление или изменение свойств существующих связей. Специальный сервис системы уведомляет авторов информационных объектов, которые стали участниками новой связи или связи с изменившимися свойствами. Тем самым стимулируется реакция оповещаемых авторов на эти события – порождение новой коммуникации, результат которой отображается в системе в виде новой связи, одним из участников которой может, в частности, являться связь – стимулятор данного действия, или изменение свойств этой связи.

Конкретные формы возможных коммуникаций и средства их реализации в среде системы Соционет рассматриваются в разд. 5 и 6.

## 3 Контент системы и семантические связи информационных объектов

Рассмотрим, прежде всего, кратко организацию контента системы Соционет – полигона для реализации рассматриваемой здесь технологии новых форм коммуникации в научном сообществе. Более подробные сведения по этому вопросу можно найти в работах [1–3].

Система Соционет построена с использованием технологии Открытых архивов (Open Archives Initiative, OAI) [12-13]. Она обеспечивает доступ к информационным объектам различных типов – статьям, монографиям, научным отчетам, справочникам, классификаторам и др. В соответствии с технологией OAI, эти объекты могут храниться на различных узлах Веба, поддерживаться и использоваться независимо от системы Соционет, имеют собственных владельцев и уникально идентифицируются их URL. В Соционет указанные информационные объекты представлены их описателями – *метаобъектами*, состоящими из наборов метаданных.

Кроме этого, Соционет также имеет дело с информационными объектами другого рода, к которым, относятся в первую очередь организации и персоны. Такие объекты, как и ранее указанные, представляются в Соционет их метаобъектами, называемыми их *профилями*. Однако сами по себе, в отличие, например, от публикаций и монографий, они где-либо в Вебе не представлены. К числу таких виртуальных информационных объектов Соционет относятся также автономные семантические связи (*см. далее*), новости, научные артефакты и др.

Каждый метаобъект имеет в системе *уникальный идентификатор*. Состав метаданных в метаобъекте зависит от типа описываемого информационного объекта.

Однотипные информационные объекты Соционет группируются в *коллекции* по критерию, которым руководствуется администратор открытого архива. Открытый архив может включать произвольное количество коллекций.

Основным информационным ресурсом открытого архива является его *репозиторий метаданных*. Он включает некоторые общие сведения об архиве в целом, а также описания коллекций информационных объектов архива. Каждая коллекция информационных объектов системы представлена в репозитории метаданных архива соответствующей коллекцией метаобъектов. В репозитории метаданных содержится также описание каждой коллекции метаобъектов в целом, а также доступный пользователям

системы каталог этой коллекции метаобъектов, который предоставляется пользователям системы как каталог соответствующей коллекции информационных объектов.

Между информационными объектами открытого архива могут существовать *бинарные ориентированные связи*. Связываемые объекты могут принадлежать разным коллекциям. Информационный объект, из которого исходит связь, далее называется *исходным объектом связи*, а объект, на который связь направлена, – ее *целевым объектом*. Связи между информационными объектами представляются в системе Соционет в виде связей между соответствующими им метаобъектами.

Некоторые связи имеют предопределенную семантику. Она не требует специального явного описания и может в дальнейшем при необходимости уточняться. Приведем несколько примеров таких связей:

- Связь между организацией и персоной – ее сотрудником, обозначающая, что данная организация является местом работы сотрудника. При уточнении семантики этой связи может быть, указана, например, должность персоны в организации.
- Связь между организацией и коллекцией информационных объектов, обозначающая, что организация является владельцем ресурсов этой коллекции.
- Связь между персоной и публикацией или другим информационным объектом, указывающая авторство данной персоны по отношению к этому информационному объекту. Семантика этой связи может быть уточнена, например, указанием вклада персоны в подготовку данной публикации.

Такие связи в репозитории метаданных представляются с помощью значений метаданных - атрибутов соответствующих метаобъектов. Так, для задания связи между персоной и организаций следует в профиле персоны указать нужное значение в атрибуте *Идентификатор профиля организации* (уникальный идентификатор, присваиваемый этому метаобъекту при его порождении в системе). Для задания обратной связи, имеющей в таком случае вид *одна-ко-многим*, нужно задать список *уникальных идентификаторов профилей персон*, являющихся ее сотрудниками. Связь авторства между публикацией и персоной в репозитории метаданных представляется указанием *идентификатора профиля автора* в качестве значения атрибута *Автор* в метаобъекте данной публикации. Связь между персоной-автором и ее публикациями представляется списком идентификаторов метаобъектов этих публикаций в профиле персоны-автора.

Учитывая способ представления связей с предопределенной семантикой в репозитории метаданных системы Соционет, мы называем такие связи *встроенными*. Помимо указанных выше классов встроенных связей в репозитории метаданных поддерживаются также некоторые системные связи – между открытым архивом и его коллекциями, между коллекцией и составляющими ее метаобъектами. Нужно отметить, что встроенные связи могут создаваться только создателями исходных метаобъектов связей, поскольку кроме них никто не имеет доступа к этим метаобъектам с целью их изменения.

Наряду со встроенными связями, в Соционет поддерживаются автономные (внешние по отношению к связываемым метаобъектам) бинарные ориентированные связи с явным образом определяемой семантикой. Мы называем такие связи *семантическими*. Заметим, что семантика может быть определена и для некоторых встроенных связей, которые при этом также становятся семантическими. Важная особенность автономной семантической связи состоит в том, что она рассматривается как обычный *самостоятельный информационный объект* специального типа. Как и для объектов других типов, из автономных связей могут строиться коллекции информационных объектов-связей. Такие связи в отличие от встроенных сами могут быть участниками других связей.

Автономные связи аналогично информационным объектам – организациям и персонам, представляемым только их профилями в репозитории метаданных Соционет, также представляются в виртуальной среде только их метаобъектами в системе. Метаобъект для объекта-связи включает следующие метаданные: уникальный идентификатор этого метаобъекта, уникальные идентификаторы исходного и целевого объектов связи, класс связи, уникальный идентификатор профиля ее автора, отметка времени момента ее создания и комментарий. Такие связи могут создавать в системе зарегистрированные пользователи в онлайновом режиме, порождая при этом их метаобъекты в репозитории метаданных.

Именно такая возможность свободного создания связей информационных объектов в системе, обладающих различной семантикой и отображающих разнообразные научные отношения между объектами научной сферы деятельности, позволяет использовать электронную библиотеку, которая предоставляет пользователям такие возможности, как полигон для новых форм научных коммуникаций представителей научного сообщества.

Как уже отмечалось, система Соционет поддерживает стандарты технологии открытых архивов OAI [12, 13]. Поэтому наряду с указанными возможностями она обладает всеми необходимыми механизмами интероперабельности для того, чтобы не только импортировать в свою среду другие открытые архивы (включая созданные в ее среде), но и предоставлять накопленные в ней метаданные для харвестинга внешними системами, поддерживающими указанные стандарты.

## 4 Описание семантики связей информационных объектов

В системе Соционет семантика создаваемых пользователями связей определяется при их создании на основе *онтологии связей*. За основу ее разработки были приняты результаты ряда известных проектов последних лет. Наиболее продвинутые ра-

боты в этой области были выполнены специалистами в области биомедицины.

Использована, в частности, одна из ранних фундаментальных разработок в этой области - модульный комплекс онтологий *SPAR* (*The Semantic Publishing and Referencing Ontologies*) [20, 22], созданный сотрудниками Оксфордского и Болонского университетов. Этот комплекс включает восемь независимых онтологий, позволяющих описывать семантику библиографических объектов, а также их отношений, в частности библиографических объектов, библиографических записей и источников в пристатейных списках, связи цитирования, контексты цитирования и их связи с релевантными разделами цитируемых публикаций. Кроме того, онтологии комплекса SPAR могут использоваться для описания семантики компонентов документов, ролей и состояний публикаций, потоков работ в издательских процессах. Каждая из онтологий комплекса описана на языках OWL2 DL и RDF консорциума W3C.

В онтологии связей системы Соционет использован также фрагмент другого комплекса онтологий - *SWAN* (*Semantic Web Applications in Neuromedicine*) [18], созданный специалистами в области нейромедицины в Главном госпитале Массачусетса и Медицинской школе Гарварда. Он предназначен для обеспечения в Семантическом Вебе комфортной среды – *социально-технической экосистемы*, которая позволяет создавать и сохранять семантический контекст научных коммуникаций, обеспечивает доступ к нему, его интеграцию, а также обмен неструктурированной и слабоструктурированной цифровой научной информацией. Классы связей в онтологиях SWAN более агрегированы по сравнению с SPAR. Онтологии комплекса описаны средствами языка описания онтологий уровня OWL DL.

Некоторые классы связей в онтологии связей системы Соционет заимствованы из рекомендации *SKOS* (*Simple Knowledge Organization System*) [23] консорциума W3C. Разработчики этой рекомендации создавали ее для использования в системах организации знаний - тезаурусах, схемах классификации, таксономиях и рубрикаторах (*Subject Heading Systems*) в среде Семантического Веба. Рекомендация SKOS определяет концептуальную схему, называемую *общей моделью данных*, которая позволяет совместно использовать и связывать различные системы организации знаний в среде Веба. При использовании такой унифицированной концептуальной схемы упрощается интеграция существующих систем организации знаний в среде Семантического Веба.

Наконец, важным источником при разработке онтологии связей для системы Соционет была модель научных данных *CERIF*, – развиваемая euroCRIS (http://www.eurocris.org/). Один из главных результатов развития CERIF – унифицированная концептуальная схема научно-исследовательской среды, называемая ее авторами *полной моделью данных (Full Data Model*) [8]. Эта модель рассматривается как единая основа разработок информационных систем (*Current Research Information Systems*, CRIS) для поддержки научно-организационной деятельности в разных странах и различных научных организациях. Использование стандартизованной концептуальной схемы обеспечивает интероперабельность таких систем. В последнее время в развитии CERIF уделяется серьезное внимание проблеме спецификации семантики полной модели данных. В частности, предложена онтология [9], определяющая систему терминов для обозначения сущностей этой модели и отношения между ними.

В онтологии связей, разработанной для системы Соционет, используются фрагменты указанных модульных комплексов и полной модели данных CERIF, включен ряд дополнений. В частности, включен класс связей с его подклассами, определяющих *вклад авторов в подготовку публикации* [11]. На основе разработанной онтологии была сформирована *двухуровневая таксономия связей*, которая и используется в системе. Таксономия связей реализована в виде набора *контролируемых словарей* имен классов (подклассов) связей. Каждому классу верхнего уровня таксономии соответствует отдельный контролируемый словарь, а его подклассам – элементы этого словаря. Таксономия включает словари оценочных связей, связей научного использования, связей научного вывода, вклада авторов в создание научного произведения и др.

Важно отметить, что связи каждого класса (подкласса) таксономии могут быть определены только на определенном множестве пар типов исходного и целевого объекта связи. При создании конкретной связи после указания исходного и целевого объектов система использует матрицу допустимости классов связей и позволяет пользователю делать выбор возможного класса создаваемой связи только из некоторого допустимого набора контролируемых словарей, соответствующего паре типов связываемых информационных объектов.

В Соционет имеются средства создания, модификации и использования таксономии связей для декларации семантических связей в системе и изменения их свойств. Более подробно таксономия семантических связей информационных объектов, используемая в системе Соционет, обсуждается в работах [2, 6, 16].

## 5 Онлайновые научные коммуникации в системе Соционет

Как уже отмечалось, в подходе, разработанном авторами и реализованном в системе Соционет, зарегистрированные в системе пользователи могут в онлайновом режиме создавать семантические связи между информационными объектами. Создание конкретных связей фактически реализует акты коммуникации между пользователями системы: между авторами создаваемых связей и авторами связываемых информационных объектов. Содержание передаваемого при этом сообщения от автора связи авторам связанных информационных объектов определяется классом/подклассом установленной связи.

Рассмотрим наиболее часто встречающиеся разновидности форм научных коммуникаций и способы их реализации в системе Соционет.

*5.1. Публикация результатов исследований и предоставление их научному сообществу без посредничества издателей*. Развитие веб-технологий и технологий издательского дела открыло новые возможности для публикации результатов выполненных исследований в научных электронных библиотеках с помощью специалистов или самостоятельно (*самоархивирование*) без традиционного посредничества издателей. Такая форма коммуникации автора публикации с научным сообществом радикальным образом сокращает время поступления сообщения ее автора адресатам в виде его публикации и многократно расширяет круг потенциальных адресатов этого сообщения – специалистов, знакомящихся с этой публикацией. Стимулирующее влияние на развитие такой формы коммуникации оказывают активно поддерживаемые в мировом научном сообществе инициативы открытого доступа к результатам научных исследований. В системе Соционет имеется механизм, называемый персональным роботом, который отслеживает появление в ее контенте публикаций по тематике, интересующей данного пользователя. В случае появления публикаций такого рода такой пользователь получает уведомление от системы.

*5.2. Сообщение пользователя системы автору некоторой публикации оценочного мнения о ней*. Если пользователь системы создает связь одного из оценочных классов, которые предусмотрены в таксономии связей, между своим профилем, хранимым в системе, и некоторой публикацией (ее метаобъектом), то он тем самым осуществляет коммуникацию с ее автором. Посредством этой коммуникации ее инициатор – создатель связи – сообщает автору публикации свое мнение о его работе. Такая коммуникация близка по смыслу традиционному рецензированию. Смысл такой рецензии может выражаться не только в структурированном виде – именем класса связи (позитивная оценка, негативная оценка и т.п.), но и неструктурированными данными – факультативным текстовым комментарием в определении метаобъекта – описателя связи. При создании такой связи в системе активизируется сервис уведомления, который направляет автору публикации сообщение по электронной почте о появлении новой связи, участником которой является его публикация, с указанием класса связи и ссылки на метаобъект связи. Конечно, работа подобного механизма коммуникаций возможна только если все их участники имеют в Соционет персональные профили с указанными в них адресами их электронной почты. В метаобъекте связи автор публикации может узнать о классе связи, ознакомиться с комментарием, авторством и временем создания связи.

*5.3. Сообщение автору публикации о возможном изменении оценочного мнения некоторого пользователя о его работе*. Зарегистрированный пользователь системы обладает полномочиями на изменение описания созданной им семантической связи. Изменение, в частности, может быть вызвано изменением мнения автора связи о ее целевой публикации. Для этого производится замена ранее указанного класса/подкласса связи или изменение комментария в метаобъекте данной связи. После сохранения модифицированного метаобъекта связи сервис уведомления, как и в предыдущем случае, отправляет автору ее целевой публикации сообщение по электронной почте.

*5.4. Сообщение автору мнения пользователя системы о научном отношении между его публикацией и некоторой другой*. Зарегистрированный пользователь может создать связь между двумя публикациями из контента системы. Например, по содержанию исходной публикации он может утверждать, что в ней высказано некоторое оценочное мнение о целевой публикации. Между такими публикациями может быть создана связь независимо от того, имеется ли в исходной публикации связи явное цитирование целевой публикации. В таком случае сообщение с предполагаемой оценкой направляется авторам обеих публикаций с указанием, как и ранее, уникального идентификатора созданной связи. Рассматриваемый случай связан не только с ситуацией, когда работа данного автора не цитируется в исходной работе связи, но и когда она цитируется. Но семантика этой связи цитирования в ней явно не определяется или определяется в контексте ссылки цитирования. Пользователь системы, создавая связь, таким образом, делает ее семантику явно определенной.

Установив связь соответствующего класса таксономии, пользователь системы тем самым указывает, что по его мнению в исходной работе связи явным или неявным образом выражено оценочное суждение о целевой работе связи. Аналогичным образом, специфицируя другие классы для создаваемой связи, можно указать, например, что в этой работе используется метод, предложенный в работе данного автора, либо из нее заимствуются данные в исходной публикации связи. Пользователь может также сообщить создаваемой связью, что в исходной работе связи имеет место плагиат из работы автора целевой публикации.

*5.5. Сообщение автору публикации о возможном изменении мнения пользователя системы о научном соотношении этой и некоторой другой публикации*. Такое сообщение порождается в случае, когда пользователь системы, например, осознал ошибочность своего мнения и вносит какие-либо изменения в метаобъект созданной им связи. После запоминания модифицированного метаобъекта, как и в предыдущем случае, авторам обеих связанных публикаций-участниц обновленной связи направляется сообщение по электронной почте.

*5.6. Проведение научных дискуссий*. Благодаря представлению автономных семантических связей в системе Соционет как обычных информационных объектов, они могут сами становиться участниками других связей. Используя эту возможность, можно поддерживать в системе дискуссионные форумы.

Пусть, например, пользователь системы выразил мнение о некоторой представленной в ней публикации, создав для этого связь какого-либо оценочного класса между своим профилем и метаобъектом указанной публикации (*см. разд.* 5.2). Уведомленный системным сервисом автор этой публикации или какой-либо иной пользователь системы может, в свою очередь, выразить мнение как о рассматриваемой публикации, так и о мнении, высказанным первым пользователем. В последнем случае он создает связь между своим профилем и метаобъектом связи, созданной первым пользователем, и т.д. Во всех случаях создания новых связей при этом сервис уведомления оповещает авторов затрагиваемых публикаций и связей. Таким образом, в такой дискуссионный процесс могут вовлекаться и авторы оцениваемых публикаций, и создатели связей, разумеется, не обязательно оценочных классов.

Ограничимся обсуждением перечисленных выше разновидностей новых форм виртуальных научных коммуникаций. Возможны различные сценарии их использования. Некоторые такие сценарии были рассмотрены выше (*см. разд.* 1). Заметим, что ключевыми механизмами обеспечения научных коммуникаций в среде электронной библиотеки с возможностью установления семантических связей между ее информационными объектами в онлайновом режиме являются механизм создания семантических связей в онлайновом режиме и сервис уведомления персон, затрагиваемых фактом создания новых и/или изменения свойств существующих семантических связей в системе.

## 6 Реализация новой технологии научных коммуникаций

Рассмотрим теперь, какие функциональные компоненты в системе Соционет обеспечивают возможности осуществления описанных форм научных коммуникаций в ее среде. Разумеется, некоторые из этих компонентов используются и для выполнения других функций системы. К числу указанных компонентов Соционет, помимо средств самоархивирования результатов научных исследований, относятся модуль создания персональных профилей пользователей системы, модуль создания контролируемых словарей таксономии семантических связей, модуль создания и обновления связей, сервис уведомления.

*Модуль создания персональных профилей пользователей системы позволяет создавать в репозитории метаданных системы* профили пользователей, в том числе и пользователей-авторов информационных объектов системы. Для поддержки научных коммуникаций необходимыми атрибутами профилей пользователей являются их (профилей) уникальные идентификаторы, присваиваемые им системой, а также их адреса электронной почты. Эти адреса необходимы для поддержки коммуникаций между пользователями системы (в том числе, пользователями-авторами представленных в системе информационных объектов) при направлении им системных сообщений-уведомлений.

*Модуль создания контролируемых словарей* позволяет системному администратору в интерактивном режиме создавать контролируемые словари, представляющие поддерживаемую в системе таксономию семантических связей. Возможно расширение поддерживаемой в системе таксономии семантических связей пользователями путем дополнения подклассов в существующие контролируемые словари классов связей и/или создания новых контролируемых словарей. Состав словарей и их содержание могут безболезненно пополняться, поскольку при этом не затрагивается состояние уже созданных коллекций связей. Включение в таксономию связей новых контролируемых словарей требует внесения соответствующего дополнения матрицы допустимости классов связей (*см. разд. 4*). Теоретически в системе может одновременно поддерживаться несколько конкурирующих таксономий связей, каждая из которых представлена своим набором контролируемых словарей классов связей.

*Модуль создания семантических связей* позволяет пользователям системы создавать, пополнять и изменять коллекции автономных семантических связей, а также удалять отдельные связи из коллекций или полные коллекции связей. Все эти операции осуществляются пользователями в онлайновом режиме и модерируются системным администратором. К коллекциям автономно представляемых связей применимы все имеющиеся в системе функциональные возможности управления коллекциями любого типа данных. Заметим, что информационные объекты-участники связей могут быть *внутренними* для системы (содержащимися в ее контенте) или *внешними*. Внешние объекты не представлены в системе, для них нет соответствующих им метаобъектов. Они должны быть доступны в Вебе по их унифицированному идентификатору ресурса (URI). Допустимость внешних информационных объектов, а также публикаций, представленных в системе только их библиографическими описаниями, в качестве участников связей позволяет охватить поддерживаемыми в системе семантическими связями более широкое цифровое научное информационное пространство.

Для заданной пары информационных объектов может быть создано несколько связей. Один и тот же автор связей не может создать несколько связей одного класса с противоречивой семантикой для заданной пары объектов, но имеет возможность создать несколько связей разных классов. Для одной и той же пары объектов разными авторами связей может быть создано несколько связей одного класса, в том числе, и с противоречивой семантикой.

*Сервис уведомления пользователей и авторов информационных объектов* создан в системе специально для поддержки коммуникаций пользователей, включая пользователей-авторов информационных объектов, представленных в системе. В процессе функционирования Соционет осуществляется *мониторинг* состояния семантических связей, а также семантической структуры контента в целом. При обнаружении изменений сервис уведомления, как

уже указывалось, информирует о них заинтересованных пользователей – авторов связанных информационных объектов по электронной почте. Например, уведомление направляется пользователю-автору какой-либо публикации, представленной в системе, если какой-либо другой пользователь создает связь с участием этой публикации. В направляемом системой сообщении указываются: идентификатор и название публикации данного автора, ставшей участницей новой связи или связи с измененными характеристиками; идентификатор этой связи и ее атрибуты, в частности, ее уникальный идентификатор и класс связи.

Система может уведомлять пользователей и в ряде иных случаев. Например, при появлении семантически противоречивых связей между некоторой парой информационных объектов, созданных разными авторами, их авторы могут оповещаться об этой ситуации сообщениями по электронной почте (пока не реализовано). Уведомления призваны стимулировать ответную реакцию их получателей, которую они могут выразить, создавая соответствующие семантические связи.

Сервис уведомления не может функционировать по принципу триггера в системах баз данных – действий в системе, описанных в схеме базы данных и активизируемых системой непосредственно после выполнения тех операторов в пользовательском запросе, которые указаны в описании триггера. Проблема в том, что всякое обновление контента системы Соционет модерируется, в соответствии с установленным регламентом. С периодичностью один раз в сутки администратор системы дает свою санкцию на фиксацию заданных обновлений в контенте системы или отвергает их. Поэтому приниматься во внимание сервисом уведомления должны лишь те предлагаемые изменения в контенте, которые прошли через этот фильтр и санкционированы администратором системы.

Автор публикаций может блокировать системные уведомления для случаев, когда какая-либо его публикация становится участницей новой семантической связи либо изменились свойства уже существующей связи, в которой она участвует. Для этого он должен отключить сервис уведомления в своих персональных настройках в Личной зоне Соционет. При необходимости этот режим может быть снова включен.

Сервис уведомления стимулирует ответную реакцию авторов публикаций – участников связей, а также пользователей системы – создателей связей, ставших участниками других связей. Тем самым этот системный механизм является *движителем коммуникационного процесса*.

## 7 Заключение

Предусмотренная в онлайновой электронной библиотеке деятельность пользователей по созданию семантических связей в режиме социальной сети представляет собой новые формы научных коммуникаций. Рассылка уведомлений авторам информационных объектов и создателям связей стимулирует новые коммуникации указанных персон, которые могут способствовать повышению качества научных результатов участников этих коммуникаций, а также существенно ускорять процессы создания и тестирования нового научного знания. Новые формы научных коммуникаций в среде электронных библиотек с рассмотренными функциональными возможностями характеризуются открытостью сообщений, передаваемых при их осуществлении. Это обеспечивает более высокий уровень ответственности участников коммуникаций перед научным сообществом. Для усиления «общественного контроля» за активностью подобного рода результаты всех действий пользователя по созданию семантических связей, а также реакции на них, фиксируются в его статистическом портрете [17]. Эта статистика является публично доступной и в определенной степени формирует научную репутацию соответствующего пользователя.

## Литература


[1] Когаловский М.Р., Паринов С.И. Метрики онлайновых информационных пространств // Экономика и математические методы. 2008. Т. 44. Вып. 2. С. 108–120.

[2] Когаловский М.Р., Паринов С.И. Семантическое структурирование контента научных электронных библиотек на основе онтологий // Современные технологии интеграции информационных ресурсов: сборник научных трудов. – Санкт-Петербург: Президентская библиотека им. Б.Н. Ельцина, 2011. С. 26–45.

[3] Когаловский М.Р., Паринов С.И. Классификация и использование семантических связей между информационными объектами в научных электронных библиотеках // Информатика и ее применения. 2012. Т. 6, вып. 3. С. 32–42.

[4] Когаловский М.Р., Паринов С.И. Новый источник данных для наукометрических исследований // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. XV Всероссийская научная конференция RCDL’2013. Ярославль, Россия, 14–17 октября 2013 г.: труды конференции. – Ярославль, 2013. С. 107–117.

[5] Когаловский М.Р., Паринов С.И. Технологии социальной сети для создания семантических связей информационных объектов в научной электронной библиотеке // Программирование. МАИК/Наука «Интерпериодика». 2014. Т. 40, № 5 (в печати).

[6] Паринов С.И., Когаловский М.Р. Технология семантического структурирования контента научных электронных библиотек // Труды XIII Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, элек-

тронные коллекции» – RCDL-2011. Воронеж, 19–22 октября 2011 г. – Воронеж: Воронежский государственный университет, 2011. С. 94–103.
[7] Паринов С.И., Ляпунов В.М., Пузырев Р.Л. Система Соционет как платформа для разработки научных информационных ресурсов и онлайновых сервисов // Российский научный электронный журнал «Электронные библиотеки». 2003. Том 6, вып. 1. http://www.elbib.ru/index. phtml?page=elbib/rus/ journal/2003/part1/PLP
[8] CERIF 1.3 Full Data Model (FDM): Introduction and Specification. euroCRIS, 2012. http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/ Specifications/ CERIF1.3_FDM.pdf
[9] CERIF 1.3 Semantics: Research Vocabulary. CERIF Task Group, euroCRIS, 2012. http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/Specifications/ CERIF1.3_Semantics.pdf
[10] Dix A., Levialdi S. & Malizia A. Semantic halo for collaboration tagging systems. In the Social Navigation and Community-Based Adaptation Technologies Workshop-June 20th, 2006.
[11] Liz Allen, Amy Brand, Jo Scott, Micah Altman and Marjorie Hlava. Credit where credit is due. http://www.nature.com/polopoly_fs/1.15033!/menu/main/topColumns/topLeftColumn/pdf/508312a.pdf
[12] Open Archives Initiative. http://www.openarchives.org/
[13] The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0 of 2002-06-14. Document Version 2008-12-07T20:42:00Z. http://www.openarchives.org/OAI/2.0/openarchi vesprotocol.htm
[14] Parinov S. Open Repository of Semantic Linkages // Proceedings of 11th International Conference on Current Research Information Systems e-Infrastructure for Research and Innovations (CRIS 2012), Prague, 2012. http://socionet.ru/publication.xml?h=repec:rus:mqijxk:29
[15] Parinov S. Towards a Semantic Segment of a Research e-Infrastructure: necessary information objects, tools and services. Metadata and Semantics Research // Communications in Computer and Information Science / J. M. Dodero, M. Palomo-Duarte, P. Karampiperis, Eds. – Springer, 2012. Vol. 343. P. 133–145. http://socionet.ru/pub.xml?h=RePEc:rus:mqijxk:30
[16] Parinov S., Kogalovsky M. Semantic Linkages in Research Information Systems as a New Data Source for Scientometric Studies // Scientometrics. 2014. Vol. 98, Iss. 2. P. 927–943. DOI 10.1007/s11192-013-1108-3
[17] Parinov S., Kogalovsky M., Lyapunov V. A Challenge of Research Outputs in GL Circuit: From Open Access to Open Use // The Gray Journal. An International Journal on Grey Literature. 2014. Vol. 10, No. 2.
[18] Semantic Web Applications in Neuromedicine (SWAN) Ontology. W3C Interest Group Note, 20 October 2009. http://www.w3.org/TR/2009/NOTE-hcls-swan-20091020/
[19] Shotton D. Open citations // Nature. 2013. Vol. 502, October 17. P. 295–297. http://www.nature.com/polopoly_fs/1.13937!/menu/main/topColumns/topLeftColumn/pdf/502295a.pdf
[20] Shotton D. Introduction the Semantic Publishing and Referencing (SPAR) Ontologies. 2010. October 14. http://opencitations.wordpress.com/2010/10/14/introducing-the-semantic-publishing-and-refere ncing-spar-ontologies/
[21] Shotton D. Use of CiTO in CiteULike. 2010. http://opencitations.wordpress.com/2010/10/21/use-of-cito-in-citeulike/
[22] Shotton D., Peroni S. Semantic annotation of publication entities using the SPAR (Semantic Publishing and Referencing) Ontologies / Beyond the PDF Workshop, La Jolla. 2011. January 19. http://imageweb.zoo.ox.ac.uk/pub/2010/Publications/Shotton&Peroni_semantic_annotation_of_publication entities.pdf
[23] SKOS Simple Knowledge Organization System Reference. W3C Recommendation. 2009. August 18. http://www.w3.org/TR/skos-reference/

## Scientific Communications Based on Digital Libraries with Tools for Online Declaration of Semantic Relationships

Mikhail R. Kogalovsky, Sergey I. Parinov

A digital library having means for the online declaration of semantic linkage between its information objects, can serve also as a platform for new forms of scientific communications in community of its users. Such information systems created for scientific community differ from traditional ones. They provide to users an opportunity to express explicitly scientific relationship between results of researches, to navigate over information resources by structure of their semantic linkage, and also to make new kinds of scientometric researches based on accumulated linkage data including the information about citation motives, etc. In the paper we discuss an ability of such information systems to improve traditional scientific communications between authors and users of research outputs using a new mechanism for direct electronic notifications and followed scientific interactions. We provide also some details about implementation of this approach within a real digital library environment. Supporting the direct scientific communications between authors and users of research outputs on base of a digital library can essentially improve a cooperation and research activity coordination for the global scientific community. This research is funded by RHF, the project 14-07-12010-v.

# Методы автоматического построения формализованного представления содержания материалов электронных средств массовых коммуникаций для решения задачи мониторинга и оценки деятельности органов власти

© Ю.В. Никитин © Ал-др А. Хорошилов © Ал-ей А. Хорошилов
Институт проблем информатики Российской академии наук (ИПИ РАН),
Москва
yuri.v.nikitin@gmail.com khoroshilov@mail.ru alex_khoroshilov@mail.ru

**Аннотация**

В данной статье рассматриваются возможности создания формализованного представления информационных публикаций в сети Интернет для получения показателей количественной оценки деятельности органов власти по материалам таких публикаций. Также рассматриваются методы построения формализованного описания информационных сообщений и методы адаптации автоматизированных средств семантической обработки сообщений для получения наиболее адекватных результатов анализа в заданной предметной области.

## 1 Введение

В настоящее время приобретает все большую актуальность получение обратной связи от населения при оценке деятельности органов власти. Такая оценка востребована как самими органами власти для принятия оперативных решений, их вышестоящими и надзорными структурами, так и различными общественными и исследовательскими организациями по мониторингу общественного мнения.

Одним из возможных источников сбора информации для проведения подобных исследований является информационное пространство электронных средств массовых коммуникаций в сети Интернет, включающее такие источники информации, как электронные средства массовой информации (СМИ), публикации в блогах и на форумах, сообщения в социальных сетях, сервисы коротких сообщений (например, Twitter) и электронные ресурсы обратной связи с населением на государственных порталах по приему жалоб и обращений граждан.

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13-16 октября 2014 г.**

Все более широкое приобщение населения к электронным информационным ресурсам, в том числе рост популярности социальных сервисов среди молодежи и стимулирование государством использования гражданами электронных госуслуг, с одной стороны, предоставляет все больше возможностей для оперирования данными, получаемыми в электронном виде по сети Интернет, с другой стороны, все более возрастающий объем подобной информации с каждым годом в значительной мере усложняет задачу обработки этих сведений, затрудняя деятельность экспертов-аналитиков в области оценки общественного мнения о деятельности органов власти.

В этих условиях эксперты-аналитики все чаще ставят задачи перехода от качественных экспертных оценок по результатам изучения материалов информационных сообщений к автоматизированным количественным методам оценки общественного мнения о деятельности органов власти. Такие методы имеют два неоспоримых взаимосвязанных преимущества: во-первых, дают возможность получения более объективных статистических показателей, благодаря росту объема обрабатываемых данных, во-вторых, максимально освобождают результаты оценки от субъективного экспертного представления о действительной ситуации по тем же самым причинам и вследствие более глубокой автоматизации обработки этого объема информации, воспринять которую в ее исходном виде за короткий период времени ни один эксперт просто не в состоянии.

В то же время решение данной задачи сталкивается с рядом комплексных проблем. Прежде всего, необходимы современные и адекватные средства семантического анализа неструктурированной текстовой информации. Далее необходимо определить набор показателей для проведения количественного статистического анализа. В настоящее время имеется серьезная проблема с отсутствием определенного набора показателей для подобных оценок. Это связано с тем, что ранее у занимающихся этими вопросами исследователей не было большого опыта обработки

данных такого объема, а также не было четкого представления о возможностях формализации текстов, доступных методах преобразования неструктурированных данных в структурированные – т.е. не было понятия о доступном инструментарии.

В данной статье мы предлагаем возможный инструментарий для проведения таких оценок и предлагаем методы автоматической обработки неструктурированной текстовой информации, позволяющей разработать модель количественных показателей. Мы также рассматриваем условия, при которых приведенные методы будут адекватны поставленной задаче в рамках заданной предметной области, посредством настройки декларативных средств к заданной предметной области с учетом большого объема обрабатываемых данных.

Мы провели моделирование процесса автоматизированной обработки текстов с целью получения количественных показателей на примере информационных сообщений Интернет-СМИ и пользователей социальной сети «ВКонтакте» о деятельности органов власти Ханты-Мансийского автономного округа (ХМАО).

## 2 Программно-техническое обеспечение

### 2.1 Общие требования к процессу автоматизации

Для автоматизации решения задачи оценки деятельности органов власти по материалам Интернет-публикаций необходимо обеспечить программно-техническую реализацию следующих основных ее подзадач:

1. **Мониторинг** материалов Интернет-публикаций:

а) консолидация информационных потоков различных типов: ленты новостей на сайтах и порталах Интернет-СМИ, органов власти и комментарии пользователей к ним, публикации на форумах и в блогах, сообщения пользователей социальных сетей и сервисов коротких сообщений, электронные обращения граждан и другие доступные ресурсы;

б) оперативный мониторинг изменения информации на подключенных ресурсах (информационных источниках) и сбор (считывание и загрузка) содержимого текстовых материалов;

в) унификация форматов представления текстовых сообщений, извлечение возможных реквизитов публикаций, аналогичных библиографическому описанию (источник, рубрика, автор, наименование публикации, временной период и т.п.), в зависимости от типа источника.

2. **Лингвистическая обработка** неструктурированных текстов:

а) автоматическое создание формализованного представления смысловой структуры текста;

б) кластеризация сходных по смысловому содержанию текстов – группировка текстов публикаций по темам (информационным поводам);

в) выделение и классификация объектов, их признаков и отношений между ними и классификация текстов по типам отношений автора сообщения к основным объектам мониторинга;

г) автоматизированная настройка декларативных (словарных) средств лингвистического процессора на заданную предметную область.

3. **Обработка данных с применением технологий «Big Data»**:

а) обеспечение распределенной массово-параллельной лингвистической и статистической обработки загружаемых данных;

б) обеспечение масштабируемости на множество узлов обработки без деградации инфраструктуры обработки данных [2, 3].

Подзадача 1 (мониторинг материалов) является достаточно тривиальной задачей, имеющей множество программно-технических решений. Подзадача 3 (обработка данных с применением технологий «Big Data») является самостоятельным направлением исследований, и мы оставляем ее за рамками данной статьи, при этом учитывая требования по распределенности и масштабируемости к средствам лингвистической обработки текстов.

В данной статье мы рассматриваем подзадачу 2 (лингвистическая обработка неструктурированных текстов), методы ее обеспечения и требования к созданию декларативных (словарных) средств [1, 7].

### 2.2 Требования к лингвистическому программному обеспечению

Современные системы автоматизированной семантической обработки неструктурированной текстовой информации, разрабатываемые для решения задач данного типа, должны обеспечивать выполнение следующих процедур лингвистического анализа текстов [1–9]:

а) графематический анализ текста;

б) морфологический анализ слов;

в) семантико-синтаксический анализ текстов;

г) концептуальный анализ текстов;

д) дистрибутивно-статистический анализ текстов.

**Графематический анализ** предназначен для предварительного анализа текста по представляющей его последовательности символов [2, 3]. В результате этого анализа определяется язык текста, устанавливаются местоположения слов, предложений, абзацев, фамильно-именной группы, дат, адресов и т.п.

**Морфологический анализ** слов естественных языков предназначен для определения структуры слов и назначения им грамматических признаков, необходимых для выполнения последующих процедур автоматической обработки текстовой информации (например, синтаксического и концептуального анализа текстов) [8, 9].

**Семантико-синтаксический анализ** текстов проводится с целью формализованного представления их структуры – выделения в них смысловых единиц и установления связей между ними. При этом структура текстов может интерпретироваться по-разному и описываться на различных формализованных языках [7, 8].

**Концептуальный анализ** текстов предназначен для определения смысловой структуры текстов, выявления их понятийного (концептуального) состава текстов и установления связей между наименованиями понятий [7, 8].

**Дистрибутивно-статистический анализ** текстов естественных языков предназначен для установления статистических закономерностей совместной встречаемости наименований понятий [1, 9].

### 2.3 Требования к извлечению данных

Основной задачей при выполнении семантической обработки неструктурированной текстовой информации является представление смысловой структуры текста в формализованном виде [8, 9].

В классическом виде формализованное представление текстового содержания документа должно содержать:

а) библиографические реквизиты (например, информационный источник, рубрика, автор, наименование и дата публикации и т.п.);

б) аннотацию или реферат документа;

в) список ключевых выражений;

г) классификацию документа по смысловому содержанию – отнесение его к той или иной рубрике и кластеризация (группировка) текстов публикаций по темам (информационным поводам).

Рассмотрим более подробно каждый из реквизитов формализованного описания применительно к поставленной задаче.

**Библиографические реквизиты** выделяются на этапе мониторинга информационных публикаций. Структура реквизитов документа (статьи, сообщения, комментария) закладывается в шаблон загрузки и парсинга (разбора на структурные элементы) страницы с текстом публикации для каждого конкретного информационного источника. Соответственно, выделение данных реквизитов происходит на стадии унификации формата документа.

При этом на уровне данных реквизитов, определенных еще до проведения лингвистического анализа, проводится классификация документа, основанная на типе информационного источника. Такая классификация может проводиться по различным основаниям:

а) тип документа (например, статья в СМИ, официальное сообщение, публикация в блоге, комментарий в социальной сети, электронное обращение граждан и т.п.);

б) степень доверия (например, официальный источник, аккредитованное СМИ, «бульварная пресса», подписанное обращение гражданина, анонимное сообщение и т.п.);

в) вид сообщения (например, информационное сообщение, жалоба, комментарий к сообщению, мнение пользователей сети и т.п.);

г) отношение к власти (например, подведомственный источник, аффилированный источник, оппозиционный источник, независимая пресса и т.п.).

**Аннотация** к тексту документа в общем случае, может быть авторской и изначально сопровождать данный текст, а может отсутствовать в исходном тексте, тогда средствами лингвистического процессора создается автоматический реферат документа [9].

В отличие от автореферата (авторского реферата, отражающего авторское представление о важных, на его взгляд, тезизах документа), автоматический реферат выделяет значимое содержимое дистрибутивно-статистическим способом, основываясь на реальных смысловых акцентах в тексте и на значимости терминов (объектов) в заданной предметной области.

Так же в отличие от автореферата, который может представлять собой стандартный шаблон с аналитическими ответами на ключевые вопросы о содержании документов данного типа, т.е. фактически являющимся пересказом (например, автореферат диссертации), автоматический реферат содержит реальный, не измененный текст документа.

Зачастую документ, содержащийся в базе данных (БД), может интересовать пользователей в разрезе различных тематик. В таком случае необходимо подготовить автоматический контекстный реферат документа в разрезе рассматриваемой тематики или поискового запроса.

В отличие от аннотации такой реферат необходимо строить каждый раз заново с учетом потребности пользователя в той или иной информации, а также ограничений на объем реферата.

С помощью автоматического реферирования также можно выравнивать объемы сравниваемых по смыслу документов – для задач кластеризации документов с аналогичным содержанием.

**Ключевые выражения** применительно к материалам электронных публикаций в рамках нашей задачи не являются важными средством визуализации смыслового содержания текста, в отличие, например, от текстов научно-технических публикаций [4–8].

При этом список ключевых выражений (наиболее значимых для данного текста с учетом предметной области) и выделенных из текстов объектов (например, персоны, должности, должностные лица, организации, территории,

производственные объекты, географические объекты, бренды) играет ключевую роль в построении формализованного описания документа для его последующего семантического анализа [2, 3].

Как и в случае с аннотацией, в отличие от авторских ключевых выражений (указанных вручную и отражающих авторское представление о важных, на его взгляд, терминах документа), данные выражения выделяются семантико-синтаксическим и словарным методами, основываясь на реальном содержании текста и на значимости терминов в заданной предметной области (выявленной дистрибутивно-статистическим методом).

**Выделение объектов** из списка ключевых выражений, таких как должностные лица, организации, территории, производственные объекты, географические объекты и бренды, представляет отдельный интерес, т.к. данные выражения зачастую являются объектами мониторинга.

В отличие от обычных ключевых выражений, чаще предназначенных только для визуализации смыслового содержания, и чуть реже – фигурирующих в качестве тематических тегов документа, выделение объектов мониторинга из текстов позволяет определить основные опорные точки для формализованных показателей:

1) выделить из текста объекты и их предикаты;

2) дать классификацию объектов и отношений к объектам на основе классификации предикатов;

3) установить по тексту связи между объектами.

**Классификация объектов** проводится на основе базового классификатора лингвистического процессора и дополнительного специально созданного по корпусу текстов классификатора заданной предметной области.

Базовый классификатор содержит только основные классы (например, персона, должность, географический объект, дата-время, обычный концепт (термин) и т.п.)

Классификатор предметной области позволяет задать более конкретную классификацию применительно к заданной предметной области (например, орган федеральной, региональной, муниципальной власти, его подразделение или подведомственное предприятие и т.п.).

Классификация объектов позволяет отнести информационное сообщение к той или иной рубрике, а также применительно к нашей задаче определить к какой ветви структуры органов власти относится объект мониторинга для более нацеленного анализа высказываний, приведенных в публикациях.

**Классификация предикатов** позволяет установить отношения к основным объектам мониторинга (органам власти и государственным функциям) авторов суждений, приведенных в информационных сообщениях – позитивные, негативные или нейтральные.

**Кластеризация (группировка) документов** выполняется на последнем этапе лингвистической обработки документов.

## 3 Экспериментальные данные

### 3.1 Предметная область и исходные данные для испытаний

Для проверки изложенной гипотезы авторы провели моделирование процесса автоматизированной обработки текстов с целью получения количественных показателей на примере информационных сообщений Интернет-СМИ и пользователей социальной сети «Вконтакте» о деятельности органов власти Ханты-Мансийского автономного округа (ХМАО).

Сначала мы провели сбор и анализ региональных публикаций ХМАО с целью автоматизированного выделения наименований органов власти, объектов инфраструктуры и должностных лиц в объеме около 1000 статей по материалам Интернет-СМИ ХМАО, сайтов органов государственной власти ХМАО (www.admhmao.ru) и городского портала органов местного самоуправления Ханты-Мансийска (www.admhmansy.ru).

Далее мы собрали региональные пользовательские публикации социальной сети «ВКонтакте» в объеме около 650 Мб неформатированного текста для анализа лексики, определяющей тональности высказываний граждан.

Кроме того нами был проанализирован представительный набор коротких сообщений Twitter, который мы рассматривали только для анализа и сравнения лексического состава сообщений информационных источников различных типов, и не использовали в дальнейшем для проведения нашего эксперимента по извлечению данных.

По результатам первичного анализа текстов сделаны выводы о необходимости разделения настройки декларативных средств по трем отдельным корпусам текстов с учетом их особенностей:

1) официальные тексты, публикации СМИ, формальные заявления, обращения граждан, экспертные заключения – любые связные тексты со строгим языковым стилем;

2) тексты социальных сетей, блогосферы и форумов – характеризуются большим количеством орфографических ошибок, опечаток, неправильно употребляемых значений слов, сокращений, жаргонизмов и профессионализмов, неологизмов, молодежной неформальной, нецензурной лексики, несвязной структурой текстов и анафорическими связями с предыдущими комментариями и сообщениями;

3) короткие сообщения Twitter (твиты) – SMS-подобные текстовые сообщения строго ограниченного размера, характеризующиеся тезисным стилем изложения информации, большим количеством сокращений и наличием хэш-тегов вида #subject, при этом часть из этих тегов фигурирует в качестве дополнительных идентификаторов темы, расположенных в начале или в конце сообщения, а часть из низ являются непосредственными членами синтаксической структуры предложения (т.е. значимыми словами в связном тексте).

### 3.2 Использование лингвистического программного обеспечения

Для проведения экспериментов мы использовали разработанное авторами статьи лингвистическое программное обеспечение (ПО) МетаФраз [10].

Лингвистическое программное обеспечение МетаФраз R10 (Metafraz Lingware R10) разработано в виде единого интегрированного многофункционального программного комплекса (Системы), состоящего из нескольких программных продуктов, предназначенных для решения отдельных функциональных задач в области компьютерной лингвистики.

В состав ПО МетаФраз R10, используемого для проведения экспериментов, входят следующие компоненты:

1) Библиотека словарей МетаФраз (MF Dictionary Lib R10) – основной ресурс Системы, содержащий комплекс декларативных (словарных) средств для задач фразеологического машинного перевода и семантической (смысловой) обработки текстов, а также набор грамматических таблиц для базовых лингвистических процедур.

2) Ядро лингвистического процессора и системы перевода (Kernel) – основной модуль Системы, включающий набор лингвистических программных библиотек, обеспечивающих выполнение всех лингвистических процедур Системы.

3) Лингвистический комплекс МетаФраз (MF Lingware Complex R10) – программный продукт Системы, входящий в состав автоматизированных систем МетаФраз, поддерживающих функционал создания и верификации словарей МетаФраз, включает модули создания частотных словарей по корпусу текстов, конвертации текстовых словарей в формат словарей МетаФраз и модуль Системы перевода МетаФраз.

4) Система семантической обработки текстов МетаФраз (MF Text Analyst R10) – программный продукт Системы, входящий в состав автоматизированных систем МетаФраз, поддерживающих функционал семантической (смысловой) обработки неструктурированных текстов на естественном языке, извлечения сущностей и установления связей, рубрикацию и кластеризацию документов, морфологический и семантических поиск и подбор документов, их автоматическое реферирование и перевод.

5) Электронная библиотека документов МетаФраз (база данных) – ресурс, входящий в состав Системы семантической обработки текстов МетаФраз, предназначенный для загрузки и хранения в БД документов (текстовых файлов) и результатов их лингвистической обработки. Электронная библиотека документов реализована с использованием СУБД MS SQL Server.

ПО МетаФраз обладает всеми необходимыми программными процедурами лингвистической обработки неструктурированных текстов, необходимых для решения данной задачи, и позволяет адаптировать декларативные средства для настройки на заданную предметную область путем быстрого автоматизированного создания словарей по корпусу текстов.

### 3.3 Автоматизированная настройка декларативных (словарных) средств на заданную предметную область

Для решения задач автоматической обработки информации в заданной предметной области необходимо провести работу по составлению семантических декларативных средств, в которых представляется понятийный состав предметной области и фиксируются смысловые отношения между понятиями.

Общая технологическая схема составления концептуального словаря представляется в следующем виде.

Предварительно составленный корпус текстов подвергается обработке процедурой семантико-синтаксического и концептуального анализа текстов, в результате чего из текстов выделяются отдельные слова и словосочетания различной длины. После этого по массиву выделенных из текстов слов и словосочетаний составляется частотный словарь.

Полученный словарь обрабатывается процедурой орфографического и синтаксического контроля, в результате чего из этого словаря исключаются некорректные слова и словосочетания. Частотная часть словаря подвергается лингвистической обработке, в результате которой из словаря исключается малоинформативная и некорректная лексика.

Далее выполняется автоматическое приведение наименований понятий к их канонической форме и формируется частотный словарь наименований понятий. И, наконец, на завершающем этапе выполняется семантико-статистический анализ частотного словаря на основе статистических данных о количественном и качественном составе этого словаря. С этой целью автоматически формируется характеристическая таблица частотного словаря. Для этого частотный словарь предварительно упорядочиваются по убыванию частот встречаемости слов в текстах и для каждой

частоты вычисляются такие параметры как его кратность, накопленная частота, накопленная кратность и относительная накопленная частота. Эти параметры позволяют выявить частотный понятийный состав предметной области и соотнести его с параметром покрытием этой частотой текстов предметной области.

Автоматизация составления словарей позволяет в короткие сроки и с минимальными трудозатратами создать для заданной предметной области систему взаимосвязанных наименований понятий, основанную на корпусе реальных текстов в предметной области.

### 3.4 Создание классификаторов в предметной области

Для создания классификаторов в предметной области были реализованы следующие этапы обработки корпуса текстов собранных публикаций:

1) разработаны методы, алгоритмы и ПО для выявления предикативных (глагольных) словосочетаний, характеризующих направление деятельности органов государственной власти и оценки их качества по корпусу текстов публикаций СМИ;

2) разработаны методы, алгоритмы и ПО для выявления оценочных суждений о деятельности органов государственной власти и оценки их качества (по корпусу текстов сообщений социальных сетей);

3) проанализирован корпус текстов СМИ, относящихся к тематике деятельности органов власти ХМАО, автоматизированным способом выделено свыше 2000 объектов, связанных со структурой органов власти региона и их деятельностью (государственными функциями);

4) по корпусу текстов СМИ выявлено более 6000 предикативных (глагольных) словосочетаний, характеризующих направление деятельности органов государственной власти и оценки их качества;

5) определены типы выделенных объектов и проведена их классификация, определяющая их отношение к структуре органов власти:

– губернаторская структура (губернатор лично, пресс-служба, аппарат, аффилированные лица губернатора);

– исполнительная власть (правительство, министерства, госпредприятия);

– законодательная власть;

– муниципальная (городские и районные власти, коммунальные службы, муниципальные предприятия) –

и их структурным подразделениям и выполняемым государственным функциям;

6) определены типы сообщений СМИ и публикаций в социальных сетях с точки зрения их отношения к объектам мониторинга:

– информирующие сообщения (преимущественно СМИ) в отношении конкретных информационных поводов (фактов, событий, персон);

– сообщения, демонстрирующие осведомленность населения по конкретным информационным поводам;

– отношение населения (позитивное, негативное, нейтральное) к конкретным информационным поводам.

На основе результатов этой обработки созданы классификационные словари объектов и отношений в заданной предметной области.

### 3.5 Обработка текстов средствами МетаФраз

После настройки декларативных средств лингвистического программного обеспечения на заданную предметную область мы провели автоматическую обработку имеющегося массива тестовых данных.

Для каждого документа было автоматически сформировано формализованное представление документа, включающее:

1. Список ключевых выражений со следующим набором данных по каждому из них:

а) ключевое выражение;

б) нормализованное ключевое выражение (пословно в канонической форме);

в) хеш выражения;

г) вес выражения в тексте с учетом предметной области;

д) группа (группы) по классификатору;

е) адреса в исходном тексте (координаты всех вхождений в тексте).

2. Общий автоматический реферат (автоконспект) заданного объема по документу.

3. Список объектов со следующим набором данных по каждому из них:

а) наименование объекта;

б) нормализованное наименование объекта (пословно в канонической форме);

в) хеш выражения;

г) вес выражения в тексте с учетом предметной области;

д) класс объекта;

е) адреса в исходном тексте (координаты всех вхождений в тексте).

4. Список объектов с предикатами со следующим набором данных:

а) объект;

б) предикат;

в) класс предиката (для типизации отношений к объектам мониторинга).

5. Список установленных связей между объектами:

а) объект 1;

б) предикат-связь;

в) объект 2.

Также было проведено последовательное отождествление документов по степени смысловой близости для кластеризации документов – группировка текстов публикаций по темам (информационным поводам).

Для отождествления смысловой близости документов в нашей задаче применялись не полные тексты документов, а их автоконспекты.

Автоконспект документа, представляющий собой автоматический реферат по наиболее значимым для данного текста и всей предметной области ключевым выражениям, позволяет провести группировку документов по их основному (наиболее значимому) информационному поводу. этот процесс также облегчается фиксированным (изначально заданным в настройках лингвистического процессора) объемом автоконспекта.

### 3.6 Оценка применимости полученных показателей

Представленное формализованное описание документов и методы его построения, а также методы кластеризации документов по информационным поводам позволили получить следующие данные, пригодные для проведения количественного анализа по текстам публикаций:

1) все сообщения (публикации СМИ, сообщения пользователей «ВКонтакте») сгруппированы по информационным поводам – что дает количественно измеряемый параметр частоты встречаемости информационного повода и его доли значимости в общем объеме публикаций;

2) все сообщения классифицированы по нескольким основаниям:

а) тип информационного сообщения (публикации СМИ, высказывания пользователей соцсетей);

б) положение основного объекта мониторинга, определяющего информационный повод, в иерархической структуре органов власти и их подведомственных организаций;

в) оценка основного объекта мониторинга (позитивная, негативная, нейтральная) –

что дает возможность проводить многомерный анализ сообщений по интенсивности воздействия СМИ, откликам пользователей сети, оценкам деятельности власти в разрезе иерархии органов власти и функциональной принадлежности предприятий и государственных услуг и т.п.

Расширение возможностей классификации и группировки сообщений с привлечением экспертов позволит существенно увеличить количество предоставляемых показателей, позволяющих проводить количественный статистический анализ данных.

## 4 Заключение

В настоящей статье авторы рассмотрели возможности создания формализованного представления содержания информационных сообщений для получения показателей количественной оценки деятельности органов власти по материалам электронных средств массовых коммуникаций.

Авторы описали структуру и методы создания формализованного описания документа для этой задачи, а также методы создания декларативных средств для получения наиболее адекватных результатов анализа в заданной предметной области.

Проведенные эксперименты на реальных данных показали жизнеспособность данного подхода, на основе которого можно создавать действующие системы мониторинга и семантического анализа информационных сообщений.

В то же время необходимо более серьезно прорабатывать методы решения поставленных задач с привлечением экспертов-аналитиков, специализирующихся в данной предметной области.

Авторы данной статьи продолжают работы по этой проблеме.

## Литература


[1] Старовойтов А.В., Пошатаев О.Н., Прохоров С.Н., Хорошилов А.А. Методы автоматизированного составления и ведения словарей // Сб. Информатизация и связь / Центр информационных технологий и систем органов исполнительной власти. – 2013. – № 3. – С. 91–97.

[2] Богданов Ю.М., Пошатаев О.Н., Хорошилов А.А. Принципы создания высокопроизводительных систем обработки и анализа текстовой информации // Сб. Информатизация и связь / Центр информационных технологий и систем органов исполнительной власти. – 2013. – № 3. – С. 74–81.

[3] Пошатаев О.Н., Хорошилов А.А. Методы анализа текстов в технологиях «Big Data» // сб. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», XV Всероссийская научная конференция RCDL 2013, Ярославль, Россия, 14–17 октября. – С. 30–38.

[4] Белоногов Г.Г., Гиляревский Р.С., Селедков С.Н., Хорошилов А.А. О путях повышения качества поиска текстовой информации в системе Интернет // Научно-техническая информация. Сер. 2. Информационные процессы и системы / Всероссийский институт научной и технической информации РАН. – 2012. – № 8. – С. 15–22.

[5] Белоногов Г.Г., Гиляревский Р.С., Хорошилов А.А. Проблемы автоматической смысловой обработки текстовой информации // Научно-техническая информация. Сер. 2. Информационные процессы и системы / Всероссийский институт научной и технической информации РАН. – 2012. – № 11. – С. 24–28.

[6] Белоногов Г.Г., Гиляревский Р.С., Хорошилов А.А., Хорошилов-мл. А.А. Автоматическое распознавание смысловой близости документов // Научно-техническая информация. Сер. 2. Информационные процессы и системы / Всероссийский институт научной и технической информации РАН. – 2011. – № 7. – С. 15–22.

[7] Белоногов Г.Г., Гиляревский Р.С., Хорошилов Ал-др А., Хорошилов Ал-ей А. Развитие систем автоматической обработки текстовой информации // Нейрокомпьютеры: разработка, применение. – 2010. – № 8. – С. 4–13.

[8] Белоногов Г.Г., Хорошилов Ал-др А., Хорошилов Ал-ей А. Единицы языка и речи в системах автоматической обработки текстовой // Научно-техническая информация. Сер. 2. Информационные процессы и системы / Всероссийский институт научной и технической информации РАН. – 2005. – № 11. – С. 21–29.

[9] Белоногов Г.Г., Калинин Ю.П., Хорошилов А.А. Компьютерная лингвистика и перспективные информационные технологии. Теория и практика построения систем автоматической обработки текстовой информации. – Москва: Информационно-издательское агентство «Русский мир», 2004. – 247 с.

[10] Сайт МетаФраз. http://www.metafraz.ru

## Methods for Automatic Construction of a Formalized Representation of the Contents of Electronic Mass Communication Materials to Solve the Problem of Monitoring and Assessment of Authorities

Yury V. Nikitin, Alexander A. Khoroshilov, Alexei A. Khoroshilov

This paper addresses to the possibility of generating a formalized representation of the information publications in the Internet to derive a quantitative evaluation of the authorities based on the content of such publications. It also covers methods of constructing a representation of messages and methods of adaptation of automated semantic processing tools for obtaining the most appropriate analysis results in a given domain.

# Формализация фактоподобных высказываний в конкретно-исторических исследованиях

© Н.А.Маркова
Институт проблем информатики РАН,
Москва
MarkovaNatAlex@gmail.com

**Аннотация**

На основе анализа специфики конкретно-исторических исследований разработана модель представления фактоподобных высказываний, включающих не только точные утверждения, но и неполные сведения, результаты их аналитико-синтетической обработки, вопросы и гипотезы. Представление высказываний в виде метаданных является основой поддерживающей информационной технологии.

## 1 Введение

Массовые электронные публикации исторических источников и исследований открывают широчайшие возможности для работ по изучению конкретно-исторических вопросов. Для того чтобы ввести в научный оборот публикуемые материалы, требуется провести их фактографическое индексирование, оснастить метаданными, представляющими содержащие в них сведения – факты – в удобном для использования виде. Эта задача жизненно важна не только для рукописных и старопечатных документов, но и для поддающихся переводу в полнотекстовый вид нарративных источников. В той или иной степени ее решает каждый исследователь, изучая источник. Не дожидаясь пока библиографы и архивисты осуществят фактографическое индексирование источников, эту работу уже выполняют виртуальные сообщества исследователей, как профессиональных, так и любителей. Подавляющее большинство площадок для обмена фактографической информацией представляет собой бессистемный обмен текстовыми репликами на форумах.

Однако существуют примеры и хорошо продуманных информационных технологий в этой области. Фактографическое индексирование, выполняемое виртуальным сообществом в рамках крупнейшего международного проекта FamilySearch [5], насчитывает более миллиарда записей, в подготовке которых участвуют сотни тысяч волонтеров. В проекте фигурирует ограниченный круг хорошо структурированных источников (в основном, регистрационных), и выходом его служит ограниченная номенклатура фактов – основные даты биографий лиц и их родственные связи.

Чрезвычайно интересные результаты были получены в рамках проводимого в Петрозаводском университете комплекса работ по формализации информации, содержащейся в коллекциях текстов исторических документов, и построения информационной системы для упорядочивания и анализа накопленных знаний в рамках работы сетевого сообщества [2]. Модель предполагает глубокую и множественную разметку исходных документов. Её сфера применения в настоящее время ограничена сообществами исследователей рукописных средневековых текстов.

Ряд особенностей конкретно-исторических исследований не позволяет применить унифицированные готовые решения, опирающиеся на представлении о факте, как об утверждении, что, в частности, неявно предполагает семантический Web. Далеко не все факты, излагаемые в исторических источниках, в основанных на них исследованиях, а также в справочниках и энциклопедиях, соответствуют объективной истине. Документы нередко содержат предположения, гипотезы, частичное знание об интересующем предмете. При этом приближение к истине возможно за счет анализа противоречий, интеграции данных, извлекаемых из различных источников. Метаданные нередко используются для представления фактографической информации (например, в проекте dbPedia), однако в них не учитываются чрезвычайно важные для конкретно-исторических исследований особенности темпоральность и неточность.

Расширим понятие «факт», включив в него неточные и неполные сведения, результаты их аналитико-синтетической обработки, вопросы и гипотезы. Предложим общую форму для фиксации такого рода сведений в виде фактоподобных высказываний (ФПВ), представимых метаданными. Модель объединяет

данные фактографического индексирования исторических источников и их аналитико-синтетической обработки. Основные положения предлагаемого формализма будем выражать в терминах ER-модели, на концептуальном уровне совпадающих с категориями аппарата онтологий, повсеместно применяемых в настоящее время для формального представления фактических знаний.

Наши построения будут основаны на анализе специфики конкретно-исторических исследований. Их целью является создание основы для построения эффективной информационной технологии поддержки работы исследователей.

## 2 Специфика конкретно-исторических исследований

В рамках конкретно-исторических исследований изучаются определенные объекты, сведения о которых частично формализуемы. Предполагается, что в данной сфере имеется специальный (выбранный исследователем) понятийный аппарат, и предметом исследования является вполне определенный набор свойств объектов, часть из которых может быть определена математически множеством допустимых значений, а часть характеризуется нарративами или образами.

Мы рассматриваем исследования, опирающиеся на изучение документальных источников. Диапазон такого рода работ постоянно расширяется за счет того, что существенная часть документов получает электронные копии, к которым обеспечивается сетевой доступ.

Изучая источник, исследователь сохраняет метаданные: адресные ссылки, выдержки, выписки, а также, по возможности, некоторую формализованную в соответствии с задачами конкретного исследования форму извлеченного знания. В соответствии с классификацией, данной в работе [1], метаданные делятся на автономные и встроенные. В терминах традиционной бумажной технологии первые – соответствуют записям, сохраняемым в виде отдельных карточек или в рабочей тетради. Вторые – результат разметки документа-источника – очерчивания, закладок, заметок на полях, использования разноцветных маркеров или стикеров.

При переходе к современной информационной технологии эффективность работы исследователя будет тем выше, чем более систематизировано удастся представить эти метаданные.

Перечислим основные особенности изучаемых объектов, которые следует учитывать при создании информационной технологии, обслуживающей конкретно-исторические исследования.

- Имена объектов вариативны (объект может иметь несколько имен) и неоднозначны (различные объекты могут иметь совпадающие имена).
- Период существования объекта, а также периоды, в котором значение некоторого свойства его постоянно, представляют ограниченные временные интервалы.
- Номенклатура изучаемых свойств объектов специфична для определенного класса объектов и зависит от конкретного исследования. Причем наличие определенного свойства и его возможные значения, а также допустимые сочетания значений этого и других свойств объекта зависит от временного интервала даже в рамках одного исследования.
- При определении свойств объекта возможны искажения вследствие дефектов в содержании источников, в процессах их распознавания, интерпретации, интеграции.

С точки зрения процесса исследований к поддерживающей его информационной технологии целесообразно предъявить следующие требования.

- Необходимо фиксировать не только четко установленные факты, но и ФПВ, включающие предположения, неточные значения свойств, исследовательские вопросы.
- Каждое сформулированное высказывание должно быть соотнесено с источником или с цепочкой вывода, обобщающей другие ФПВ.
- Необходимо обеспечить информационную навигацию по связям между объектами, в том числе, для электронных документов – межтекстовые связи; многоаспектный поиск; возможности статистической обработки.
- Необходимо отслеживать процессы накопления данных, выявления дефектов, выдвижения/ опровержения гипотез по исследуемым источникам, времени, исполнителям, аргументации.

Концептуальной основой для создания информационной технологии, удовлетворяющей перечисленным требованиям, является предлагаемая формальная модель представления ФПВ.

## 3 Модель фактоподобных высказываний

Принципиальная проблема, которую необходимо решить при разработке модели ФПВ, состоит в выборе рационального уровня формализации. Малоэффективны как совсем неформальное текстовое представление (нарратив), так и попытка максимальной формализации. Сформулируем три положения, опора на которые позволит выбрать оптимальный уровень формализации.

1) Модель строится по ER-принципу, с определенными наборами объектов, атрибутов, отношений.

2) Для каждой сферы исследования выбираются свои наборы объектов и свойств, возможно, уточняемые для конкретного проекта.

3) Формализуются не все возможные свойства, а те, которые отражают поддающиеся типизации аспекты, возможные значения которых задаются диапазоном чисел, дат; словарным перечнем. Все, что не укладывается в эти рамки (а также малозначимые в рамках конкретного проекта сведения), представляется нарративным текстом.

Модель включает три группы элементов. Основные элементы модели – компоненты ER-модели изучаемого исторического процесса: объекты, атрибуты, отношения – фиксируются базовыми высказываниями. Высказывания-связки соотносят базовые высказывания с источниками и между собой. Наконец, информацию, включающую высказывания-ограничения, а также данные, относящиеся к процессу исследования, отнесем к служебным высказываниям. Рассмотрим эти группы высказываний подробнее.

### 3.1 Базовые высказывания

Множество объектов исследования ($O = \cup O_{class}$) включает объекты определенных классов. Для каждого класса объектов устанавливается набор свойств. Литеральные свойства – атрибуты – сопоставляют объекту некоторое значение из определенного множества (чисел, дат, номинальных шкал, текстов). Объектные свойства – отношения – сопоставляют объекту другой объект и литеральное значение, которое можно воспринимать, как метку на графе связей между объектами. Такая конструкция, вместо используемого в OWL строгого разделения на категории свойств, не предполагающего литеральных значений у объектных свойств, позволяет не вводить дополнительных объектов (*Отношение между Петровым и Гимназией*), а, оставаясь в рамках объектов исследования (*Петров*, *Гимназия*) специфицировать значение связи (*Должность = Инспектор*). Для конкретно-исторических задач такое представление существенно нагляднее.

Далеко не все высказывания, содержащиеся в историческом источнике, можно формализовать. Но даже для формализуемых высказываний, суждение о значении свойства объекта может быть сформулировано не только как равенство некой константе, но и как различные варианты неравенства, а также принадлежности (не принадлежности) некоторому набору констант. Оператор ФПВ, соотносящий значения свойства объекта с константой/списком констант (•), определим следующим образом:

$$\bullet \in \{=, \approx, \neq, <, >, \in, \notin\}$$

Наиболее важная особенность предлагаемой модели – включение в ФПВ временного интервала, в рамках которого оно предполагается справедливым. Такая конструкция предоставляет значительно более удобный базис для аналитико-синтетической работы в области конкретно-исторических исследований, чем фиксация отдельных событий. Действительно, подавляющее большинство событий, касающихся объекта, имеют свою пару – они фиксируют начало и конец периода, в котором некоторое свойство объекта имело некоторое значение. Даже для такого свойства, как титул, возможны события присвоение, лишение, восстановление, определяющие соответствующие временные интервалы. В любом случае, время ФПВ, касающегося некоторого свойства объекта, ограничено временем существования (жизни) объекта.

Будем определять периоды ($dt \in DT$), как

$$dt = (start, finish),$$

где start и finish – это либо даты (с некоторой степенью точности), либо оценки ограничений, налагаемые на эти даты. Подробно форма представления временных интервалов разной степени определенности в виде строки метаданных рассмотрена в [3].

Предложенный подход совсем не противоречит возможности в рамках конкретного исследования определить специальный класс объектов – события определенного рода (например, *Конференции*).

Рассмотрим основные виды базовых ФПВ и определим содержание метаданных, их фиксирующих.

1) Дефиниция – высказывание, определяющее существование объекта определенного класса в определенный период времени:

$$\forall t \in dt \ (o_d(t) \in O_{class}).$$

d = (nomen, class, dt) – метаданные, фиксирующие дефиницию.

Здесь nomen – имя объекта – неформальная и, возможно, неуникальная текстовая константа, служащая для удобства восприятия исследователем. Каждое ФПВ имеет свою уникальную идентификацию, которую для простоты описания мы опускаем. При адресации ФПВ будем использовать его обозначение (например, d).

2) Атрибут – высказывание, определяющее значение определенного литерального свойства объекта в определенный период времени:

$$\forall t \in dt \ (a_{aclass}(o_d, t) \bullet avalue).$$

a = (d, aclass, avalue, •, dt) – метаданные, фиксирующие атрибут.

3) Отношение – высказывание, определяющее связь объекта с другим объектом, а также

литеральное значение, сопоставляемое этой связи в определенный период времени:

$$\forall t \in dt \ (r_{rclass}(o_{dp}, o_{dq}, t) \bullet rvalue).$$

r = ($d_p$, $d_q$, rclass, rvalue, •, dt) – метаданные, фиксирующие отношение.

Во всех конкретно-исторических исследованиях рассматриваются классы *Лицо* и *Документальный* объект (*Д-объект)*. В большинстве случаев интерес представляют *Географические* и *Социальные* объекты. В специальных исследованиях классами изучаемых объектов являются *Архитектурные, Природные, Математические и пр.* объекты.

Наиболее общими для самых разных областей исследований являются свойства *Д-объектов*, под которыми мы понимаем не только документы, но и их совокупности, и их компоненты (от архивов, библиотек, интернет-порталов до абзацев текста). Атрибуты и связи документов хорошо специфицируют библиографические и археографические стандарты. В рамках современных стандартов IFLA (например, [5]) рассматриваются связи между документами, представляющие интерес для конкретно исторических исследований. К ним относятся: структурная (входит, следует за), деривативная (версии, переработки, переводы), дескриптивная (критика, комментарии, аннотации, рефераты) связи.

Атрибуты и связи *Лиц*, в основном, специфичны для сферы исследований. Универсальны атрибуты *пол* и связь с гео-объектами *местопребывание* (которое, например, в момент рождения – место рождения). Достаточно часто рассматриваются родственные связи, должностные отношения, отношения учитель-ученик. Связи *Лиц* и *Д-объектов* фиксируют сведения об авторстве, адресатах и упоминаниях.

### 3.2 Высказывания-связки

Утверждение о том, что некоторое ФПВ получено в результате интерпретации (⇒) определенного источника также является определенным высказыванием. Источник при этом адресуется дефиницией соответствующего *Д-объекта*. Сопоставляя ФПВ, исследователь конструирует новые выражения с помощью логических или темпоральных связок. Как интерпретация, так и логические операции над ФПВ не являются в полной мере формальными действиями. В рассуждениях исследователя есть доля интуиции. Однако степень уверенности в своих умозаключениях вполне оцениваема. Поэтому каждому ФПВ-связке будем сопоставлять оценку уверенности в фиксируемой им формулировке. Такую оценку рационально выражать в шкале нечеткой логики от 0 – FALSE до 1 – TRUE.

Пятно на рукописи, неразборчивый почерк, неизвестные сокращения – причины того, что исследователь неуверен в результатах интерпретации. Но и при уверенности в толковании источника, исследователь может быть не согласен со смыслом интерпретированного высказывания. В этом случае он должен зафиксировать противоречие между данными источника и более надежными сведениями, что послужит обоснованием для высказывания, фиксирующего ложность сведений источника. Пример цепочки такого рода размышлений, фиксируемых средствами ФПВ, будет приведен в следующем разделе.

Определим множество ФПВ (представленных метаданными) – F, как объединение вышеперечисленных видов ФПВ и специальных ФПВ-связок – L, определяемых рекурсивно.

$$F = D \cup A \cup R \cup L,$$

где $D = \{d\}$, $A=\{a\}$, $R=\{r\}$, $L = \{l\}$

$l = (fp, fq, \blacklozenge, estim)$

$fp \in F$, $fq \in F$,

$estim \in [0..1]$, (0 –TRUE, 1 – FALSE)

$\blacklozenge \in \{\Rightarrow\} \cup Logical \cup Temporal$

Logical = {AND, OR, XOR, …}

Temporal = {BEFORE, AFTER, SAMETIME, INTERSECT}

⇒ – интерпретация.

### 3.3 Служебные высказывания

Каждой сфере исследования соответствует свой набор классов объектов, их свойств, зависимостей между значениями свойств. Часть из этих ограничений легко формализуема. Например, спецификация перечней классов объектов и классов свойств, в зависимости от классов объектов; списки возможных значений свойств. Несколько сложнее, но все же возможно формализовать ограничения на возможные сочетания значений свойств, а также на временные характеристики. Примерами такого рода ограничений являются накладываемые биологическими законами разности в возрасте родителей и детей, или формулируемые конкретным социальным устройством регламент продвижения по службе.

Важнейшим служебным высказыванием является перечень классов (например, class ∈ {*Лицо*, *Д-объект*, *Гео-объект*, *Соц-объект*}). Ограничения на атрибуты формулируются указанием области определения и области значений (domain и range). Например, domain(*Пол*)= *Лицо*; range(*Пол*)={м, ж, ?}. Для отношений область определения задается парой, например, domain(*Родство*)= (*Лицо*, *Лицо*).

Для фиксации ограничений на возраст детей может потребоваться формализация высказывания

$\forall$ $d_0$, d (($d_0$, d, *Родство*, *Родитель*) $\rightarrow$
(d.start – $d_0$.start > 10) AND
(d.start – $d_0$.start < 90))

Должны ли фиксироваться подобные ограничения в виде метаданных, интерпретируемых некоторым унифицированным инструментом, или они представляют специализированные процедуры контроля (своего рода сложных алгоритмических высказываний) – зависит от конкретных обстоятельств. Во многих случаях контроль ограничения вообще может быть выполнен только вручную. В целом, полезно хотя бы в неформальном, текстовом виде фиксировать ограничения, как своего рода памятку для исследователя (нарративное высказывание).

В соответствии с выдвинутыми требованиями необходимо отслеживать процесс накопления данных и их аналитико-синтетической обработки. Для этого целесообразно применить типовой прием, используемый, в частности, в wiki-технологии. Каждая запись ФПВ сопровождается временной меткой и указанием автора. Вместо изменения записи производится формирование ее новой версии.

## 4 Пример рассуждений, фиксируемых ФПВ

Рассмотрим пример интерпретации источника, выявления противоречия, формулировки новых ФПВ. Источник – книга, посвященная 100-летию Первой московской гимназии [6].

$d_0$ = («Столетие 1-й гимназии», *Д-объект*, 1903)

$d_1$ = («1-я гимназия», *Соц-объект*, 1804-1904..)

В источнике содержатся, в частности, списки выпускников по годам выпуска, а также списки печатных работ, авторами которых являются выпускники гимназии. Два однофамильца – Алексей М. и Александр М. окончили гимназию соответственно в 1896 и в 1888 годах:

$d_2$ = («Алексей М.», *Лицо*, 1874..1878-1903..)

$r_1$ = ($d_2$, $d_1$, *Ученик*, ..-1896)

$d_3$ = («Александр М.», *Лицо*, 1866..1870-1903..)

$r_2$ = ($d_3$, $d_1$, *Ученик*, ..-1888)

Оценка времени жизни дана, исходя из ограничения на возраст учеников.

В источнике допущена ошибка. В комментарии, относящимся к Александру М., сказано «Известный этнограф, исследователь былин сев. края». При этом работ у Александра М. не отмечено, а вот у Алексея М. отмечено несколько работ, посвященных северным былинам.

$a_1$ = ($d_3$, «исслед. былин сев. края», 1888..-)[1]

$l_1$ = ($d_{0\ c.295}$, $a_1$, $\Rightarrow$, 1)

$a_2$ = ($d_2$, «автор Беломорские былины», 1901)

$l_2$ = ($d_{0\ c.295}$, $a_1$, $\Rightarrow$, 1)

Итак, ФПВ $a_1$ и $a_2$ противоречивы:

$l_3$ = ($a_1$, $a_2$, *AND*, 0.01)

Формальную возможность того, что и Александр М. был «известным этнографом», но не публиковал своих исследований, мы оценили в 1 процент.

$l_4$ = ($l_3$, $a_2$, $\Rightarrow$, 0.01)

$l_5$ = ($l_4$, $r_1$, *AND*, 0.99)

Теперь мы можем сформулировать новое высказывание, корректирующее ошибочное $a_1$:

$a_3$ = ($d_2$, «исслед. былин сев. края», 1896..-)

$l_6$ = ($l_4$, $a_3$, $\Rightarrow$, 0.99)

Строго говоря, приведенная цепочка рассуждений, равно как и операция интерпретации источника не являются формальными. Однако возможность формализованной фиксации результатов мыслительных операций существенно дисциплинирует исследователя, а также позволяет осуществиться научной коммуникации, что служит залогом взаимного контроля и способствует повторному использованию данных исследования.

## 5 Заключение

В рамках данной работы модель ФПВ представлена концептуально. При ее использовании в конкретной информационной технологии она должна быть выражена в терминах соответствующего аппарата, в качестве которого могут выступать как современные языки онтологий, так и инструменты баз данных.

Опора на языки онтологий позволит организовать обмен информацией с другими информационными системами. В частности, это позволит импортировать конечные (или хотя бы стабилизированные) данные исследования в качестве фактографического индекса в библиографическую/археографическую информационную систему.

Технология баз данных обеспечит эффективность накопления и аналитико-синтетической обработки ФПВ. Однако наилучшего результата, как показала практика разработки и эксплуатации инструментального комплекса Фактограф [4], можно добиться, сочетая автономные метаданные, хранимые в базе данных, и встроенные, размечающие документ-источник. При этом предполагается, что исследователь имеет свою копию источника, которую он может «чиркать» разметкой. Взаимные связи между ФПВ, хранимыми в базе данных, и фрагментами текста источника достигаются средствами гиперссылок. В документах-источниках границы фрагментов,

[1] Для краткости мы опускаем • = «=», класс атрибута – *Упоминание*, а также определения Д-объектов – страниц, входящих в книгу-источник.

связанных с высказываниями, хранимыми в базе, определяются либо явно (для xml и html форматов), либо закладками, применимыми не только в офисных документах, но и в документах форматов pdf и djvu. В свою очередь, гиперссылками на форму, представляющую конкретный объект в базе данных, целесообразно оснастить текст источников в точках его упоминания. В случае не редактируемых документов (pdf и djvu) такую ссылку можно поместить в комментарий.

Вычленяя ФПВ из источника, мы обеспечиваем удобство его контроля, анализа, интеграции, но в то же время, теряем контекст, который может быть чрезвычайно полезен для создания целостной картины. С другой стороны, возможность получения оперативной справки по ходу чтения источника, касающейся его текущего участка, способствует пониманию текста. Сравнение обладающего внутренним единством линейного текста со структурной картиной связанных объектов, в нем упоминаемым, дает возможность как уточнить идентификацию объектов, сформулировать новые ФПВ, так и глубже понять подтекст, неподдающийся формализации.

Повторное обращение к источнику (адресное, и поэтому эффективное), равно как и повторное использование выявленных сведений, чрезвычайно полезно уже индивидуальному исследователю. Тем важнее эти возможности для организации информационного обмена в сообществах, изучающих историю. Предложенный в работе метод формализации данных может служить основой для создания информационной технологии, существенно повышающих эффективность работы коллектива исследователей.

## Литература


[1] Когаловский М.Р. Метаданные в компьютерных системах // Программирование, МАИК/Наука «Интерпериодика». 2013. Т. 39, № 4. С. 28–46.

[2] Кравцов А.В. Информационные модели и технологии в организации работы научного сообщества по публикации и анализу коллекций исторических документов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XI Всероссийской научной конференции RCDL'2009. Петрозаводск: КарНЦ РАН, 2009. С. 210–218.

[3] Маркова Н.А. Логика биографических фактов //Информатика и ее применения, 2012. Т. 6, вып. 2. С. 49–58.

[4] Маркова Н.А. Программа Средства интеграции, хранения и анализа биографических данных (Фактограф). Свидетельство о государственной регистрации программы для ЭВМ № 2013617234 от 06.08.2013.

[5] Руководство пользователя по программе FamilySearch Indexing. © 2009, 2014 by Intellectual Reserve, Inc. URL: http://broadcast.lds.org/elearning/FHD/Local_Support/FamilySearchIndexing/RU/fsi_user_guide.pdf

[6] Столетие Московской 1-й гимназии. 1804-1904 гг. / сост. И. Гобза. – М.: Синод. тип., 1903. URL: http://dlib.rsl.ru/viewer/01003711731#?page=1

[7] Функциональные требования к библиографическим записям / Рос. библ. ассоц., РГБ. – М.: Пашков дом, 2008.


## Formalization of the Fact-like Propositions in Specific Historical Studies


Natalia A. Markova



The paper proposes a model of metadata representation of the fact-like propositions that specify not only true statements, but suggestions, hypothesis, incomplete information, the results of analytic/synthetic processing. Requirements to provide efficiency of the specific historical studies are under consideration. The metadata are considered as the base of supporting IT.

# Introduction into Analysis of Methods and Tools for Hypothesis-Driven Scientific Experiment Support

© Kalinichenko L.A. © Kovalev D.Y. © Kovaleva D.A. © Malkov O.Y.
Institute of Informatics Problems RAS Institute of Astronomy RAS
Moscow
leonidk@synth.ipi.ac.ru dm.kovalev@gmail.com dana@inasan.ru malkov@inasan.ru

## Abstract

Data intensive sciences (DIS) are being developed in frame of the new paradigm of scientific study known as the Fourth paradigm, emphasizing an increasing role of observational, experimental and computer simulated data practically in all fields of scientific study. The principal goal of data intensive research (DIR) is an extraction (inference) of knowledge from data.

The intention of this work is to make an overview of the existing approaches, methods and infrastructures of the data analysis in DIR accentuating the role of hypotheses in such research process and efficient support of hypothesis formation, evaluation and selection in course of the natural phenomena modeling and experiments carrying out. An introduction into various concepts, methods and tools intended for effective organization of hypothesis driven experiments in DIR is presented in the paper.

## 1 Hypotheses, theories, models and laws in data intensive science

Data intensive science (DIS) is being developed in accordance with the 4th Paradigm [29] of scientific study (following three previous historical paradigms of the science development (empirical science, theoretical science, computational science)) emphasizing that science as a whole is becoming increasingly dependent on data as the core source for discovery. Emerging of the 4th Paradigm is motivated by the huge amounts of data coming from scientific instruments, sensors, simulations, as well as from people accumulating data in Web or social nets. The basic objective of DIS is to infer knowledge from the integrated data organized in networked infrastructures (such as warehouses, grids, clouds). At the same time, "Big Data" movement has emerged as a recognition of the increased significance of massive data in various domains. Open access to large volumes of data therefore becomes a key prerequisite for discoveries in the 21st century. Data Intensive Research (DIR) denotes a crosscut of DIS/IT areas aimed at the creation of effective data analysis technologies for DIS and other data intensive domains.

Science endeavors to give a meaningful description of the world of natural phenomena using what are known as laws, hypotheses and theories. Hypotheses, theories and laws in their essence have the same fundamental character (Fig. 1) [48].

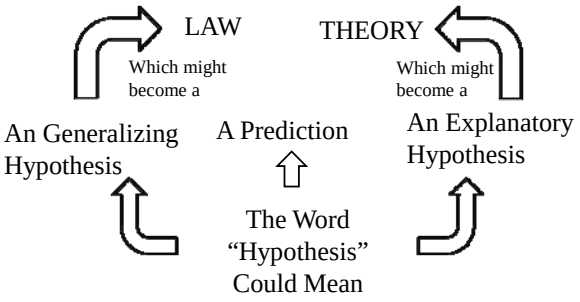



Fig. 1. Multiple incarnations of hypotheses

A *scientific hypothesis* is a proposed explanation of a phenomenon which still has to be rigorously tested. In contrast, *a scientific theory* has undergone extensive testing and is generally accepted to be the accurate explanation behind an observation. A *scientific law* is a proposition, which points out any such orderliness or regularity in nature, t*he prevalence of an invariable association between a particular set of conditions and particular phenomena.* In the exact sciences laws can often be expressed in the form of mathematical relationships. Hypotheses explain laws, and well-tested, corroborated hypotheses become theories (Fig. 1). At the same time the laws do not cease to be laws, just because they did not appear first as hypotheses and pass through the stage of theories.

Though theories and laws are different kinds of knowledge, actually they represent different forms of the same knowledge construct. Laws are generalizations, principles or patterns in nature, and theories are the explanations of those generalizations. However, classification expressed at the Fig. 1 is subjective. [40] provides examples showing that the differences between laws, hypotheses and theories consist only in that they stand at different levels in their claim for acceptance depending on how much empirical evidence is amassed. Therefore there is no essential difference between constructs used for expressing

**Proceedings of the 16th All-Russian Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" — RCDL-2014, Dubna, Russia, October 13–16, 2014.**

hypotheses, theories and laws. Important role of hypotheses in scientific research can scarcely be overestimated. In the edition of M. Poincaré's book [52] it is stressed that *without hypotheses there is no science*. Thus it is not surprising that so much attention in the scientific research and the respective publications is devoted to the methods for hypothesis manipulation in experimenting and modeling of various phenomena applying the means of informatics. The idea that the new approaches are needed that can address both data driven and *hypothesis driven sciences* runs all through this paper. Such symbiosis alongside with the hypothesis-driven tradition of science (‘‘first hypothesize-then-experiment’’) might cause wide application of another one that is typified by ‘‘first experiment-then-hypothesize’’ mode of research. Often the “first experiment” ordering in DIS is motivated by the necessity of analysis of the existing massive data to generate a hypothesis.

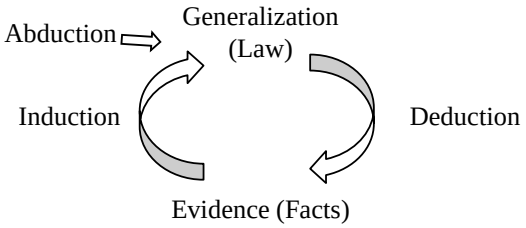



Fig. 2 Enhanced knowledge production diagram

In the course of our study paying attention to the issue of inductive and deductive reasoning in hypothesis driven sciences will be emphasized. On Fig. 2 such ways of knowledge production are shown [48]. “Generalization” here means any subset of hypotheses, theories and laws and “Evidence” is any subset of all facts accumulated in a specific DIS.

All researchers collect and interpret empirical evidence through the process called *induction*. This is a technique by which individual pieces of evidence are collected and examined until a law is discovered or a theory is invented. Frances Bacon first formalized induction [4]. The method of (naïve) induction (Fig. 2) he suggested is in part the principal way by which humans traditionally have produced generalizations that permit predictions. The problem with induction is that it is both impossible to collect all observations pertaining to a given situation in all time – past, present and future.

The formulation of a new law begins through induction as facts are heaped upon other relevant facts. Deduction is useful in checking the validity of a law. The Fig. 2 shows that a valid law would permit the accurate prediction of facts not yet known. Also an *abduction* [49] is the process of validating a given hypothesis through reasoning by successive approximation. Under this principle, an explanation is valid if it is the best possible explanation of a set of known data. Abductive validation is common practice in hypothesis formation in science. Hypothesis related logic reasoning issues are considered in more details in section 3.

In [52] the useful hypotheses of science are considered to be of two kinds:

1. The hypotheses which are valuable *precisely* because they are either verifiable or else refutable through a definite appeal to the tests furnished by experience;

2. The hypotheses which, despite the fact that experience suggests them, are valuable *despite*, or even *because*, of the fact that experience can *neither* confirm nor refute them.

Aspects of science which are determined by the use of the hypotheses of the second kind are considered in the M. Poincaré's book [52] as “constituting an essential human way of viewing nature, an interpretation rather than a portrayal or a prediction of the objective facts of nature, an adjustment of our conceptions of things to the internal needs of our intelligence”. According to M. Poincaré's discussion, the central problem of the logic of science becomes the problem of the relation between the two fundamentally distinct kinds of hypotheses, i.e., between those which cannot be verified or refuted through experience, and those which can be empirically tested.

The analysis in this paper will be focused mostly on the modeling of hypotheses of the first kind, leaving issues of analysis the relations between such two kinds of hypotheses to further study.

The rest of the paper is organized as follows. Section 2 discusses the basic concepts defining the role of hypotheses in the formation of scientific knowledge and the respective organization of the scientific experiments. Approaches for hypothesis formulation, logical reasoning, hypothesis modeling and testing are briefly introduced in Section 3. In Section 4 a general overview of the basic facilities provided by informatics for the hypothesis driven experimentation scenarios, including conceptual modeling, simulations, statistics and machine learning methods is given. Into Section 5 several examples of organization of hypothesis driven scientific experiments are included. Conclusion summarizes the discussion.

## 2 Role of hypotheses in scientific experiments: basic principles

Normally, scientific hypotheses have the form of a mathematical model. Sometimes one can also formulate them as existential statements, stating that some particular instance of the phenomenon under examination has some characteristic and causal explanations, which have the general form of universal statements, stating that every instance of the phenomenon has a particular characteristic (e.g., *for all x, if x is a swan, then x is white*). Scientific hypothesis considered as a declarative statement identifies the predicted relationship (associative or causal) between two or more variables (independent and dependent). In causal relationship a change caused by the independent variable is predicted in the dependent variable. Variables are more commonly related in non-causal (associative) way [25].

In experimental studies the researcher manipulates the independent variable. The dependent variable is often referred to as consequence or the presumed effect that varies with a change of the independent variable. The dependent variable is not manipulated. It is observed and assumed to vary with changes in the independent variable. Predictions are made from the independent variable to the dependent variable. It is the dependent variable that the researcher is interested in understanding, explaining or predicting [25].

In case when a possible correlation or similar relation between variables is investigated (such as, e.g., whether a proposed medication is effective in treating a disease, that is, at least to some extent and for some patients), a few cases in which the tested remedy shows no effect do not falsify the hypothesis. Instead, statistical tests are used to determine how likely it is that the overall effect would be observed if no real relation as hypothesized exists. If that likelihood is sufficiently small, the existence of a relation may be assumed. In statistical hypothesis testing two hypotheses are compared, which are called the *null hypothesis* and the *alternative hypothesis*. The null hypothesis states that there is no relationship between the phenomena (variables) whose relation is under investigation, or at least not of the form given by the alternative hypothesis. The alternative hypothesis, as the name suggests, is the alternative to the null hypothesis: it states that there *is* some kind of relation.

Alternative hypotheses are generally used more often than null hypotheses because they are more desirable to state the researcher's expectations. But in any study that involves statistical analysis the underlying null hypothesis is usually assumed [25]. It is important, that the conclusion "do not reject the null hypothesis" does not necessarily mean that the null hypothesis is true. It suggests that there is not sufficient evidence against the null hypothesis in favor of the alternative hypothesis. Rejecting the null hypothesis suggests that the alternative hypothesis may be true.

Any useful hypothesis will enable *predictions by reasoning* (including *deductive reasoning*). It might predict the outcome of an experiment in a laboratory setting or the observation of a phenomenon in nature. The prediction may also invoke statistics assuming that a hypothesis must be *falsifiable* [53], and that one cannot regard a proposition or theory as scientific if it does not admit the possibility of being shown false. The way to demarcate between hypotheses is to call *scientific* those for which we can specify (beforehand) one or more potential falsifiers as the respective experiments. Falsification was supposed to proceed deductively instead of inductively.

Other philosophers of science have rejected the criterion of falsifiability or supplemented it with other criteria, such as verifiability (only statements about the world that are empirically confirmable or logically necessary are cognitively meaningful). They claim that science proceeds by "induction"— that is, by finding confirming instances of a conjecture. Popper treated confirmation as never certain [53]. However, a falsification can be sudden and definitive. Einstein said: "No amount of experimentation can ever prove me right; a single experiment can prove me wrong". To scientists and philosophers outside the Popperian belief [53], science operates mainly by induction (confirmation), and also and less often by disconfirmation (falsification). Its language is almost always one of induction. For this survey both philosophical treatment of hypotheses are acceptable. Sometimes such way of reasoning is called the *hypothetico-deductive method***.** According to it, scientific inquiry proceeds by formulating a hypothesis in a form that could conceivably be falsified by a test on observable data. A test that could and does run contrary to predictions of the hypothesis is taken as a falsification of the hypothesis. A test that could but does not run contrary to the hypothesis corroborates the theory.

A scientific method involves experiment, to test the ability of some hypothesis to adequately answer the question under investigation. A prediction enabled by hypothesis suggests a test (observation or experiment) for the hypothesis thus becoming testable. If a hypothesis does not generate any observational tests, there is nothing that a scientist can do with it.

For example, not testable hypothesis: "Our universe is surrounded by another, larger universe, with which we can have absolutely no contact"; not verifiable (though testable) hypothesis: "There are other inhabited planets in the universe"; scientific hypothesis (both testable and verifiable): "Any two objects dropped from the same height above the surface of the earth will hit the ground at the same time, as long as air resistance is not a factor" (http://www.batesville.k12.in.us/physics/phynet/aboutscience/hypotheses.html).

A *problem* (*research question*) should be formulated as an issue of what relation exists between two or more variables. The problem statement should be such as to imply possibilities of empirical testing otherwise this will not be a scientific problem. Problems and hypotheses being generalized relational statements enable to deduce specific empirical manifestations implied by the problem and hypotheses. In this process hypotheses can be deduced from theory and from other hypotheses. A problem cannot be scientifically solved unless it is reduced to hypothesis form, because a problem is not directly testable [37].

Most formal hypotheses connect concepts by specifying the expected relationships between *propositions*. When a set of hypotheses are grouped together they become a type of *conceptual framework*. When a conceptual framework is complex and incorporates causality or explanation it is generally referred to as a *theory* [28]. In general, hypotheses have to reflect the multivariate complexity of the reality. A scientific theory summarizes a hypothesis or a group of hypotheses that have been supported with repeated testing. A theory is valid as long as there is no evidence to dispute it. *Scientific paradigm* explains the working set of theories under which science operates.

Elements of hypothesis-driven research and their relationships are shown on Fig. 3 [23, 57]. The hypothesis triangle relations, *explains, formulates, represents* are functional in the scientist's final decision in adopting a particular model *m1* to formulate a hypothesis *h1*, which is meant to explain phenomenon *p1*.

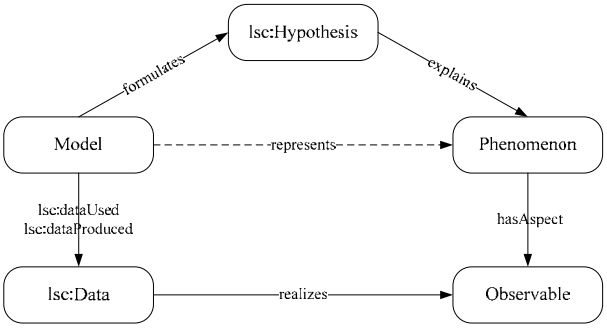



Fig. 3. Elements of hypothesis-driven research

In [23] the lattice structure for hypothesis interconnection is proposed as shown on Fig. 4. A hypothesis lattice is formed by considering a set of hypotheses equipped with *wasDerivedFrom* as a strict order < (from the bottom to the top). Hypotheses directly derived from exactly one hypothesis are *atomic*, while those directly derived from at least two hypotheses are *complex*.

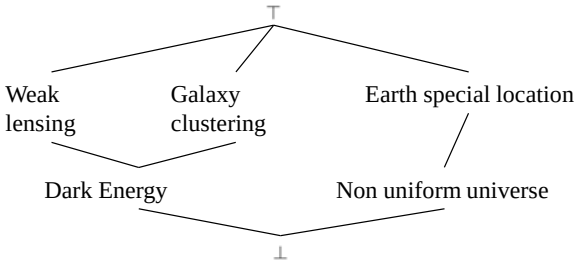



Fig. 4. A lattice theoretic representation for hypothesis relationship

The hypothesis lattice is unfolded into model and phenomena isomorphic lattices according to the hypothesis triangle (Fig. 3) [23]. The lattices are isomorphic if one takes subsets of *M* (Model), *H* (Hypotheses) and *P* (Phenomenon) such that *formulates, explains and represents* are both one-to-one and onto mappings (i.e., bijections), seen as structure-preserving mappings (morphisms). Example of the isomorphic lattice is shown on the Fig. 5 [23]. This particular lattice corresponds to the case in Computational Hemodynamics considered in [23]. Here model *m1* formulates hypothesis *h1*, which explains phenomenon *p1*. Similarly, *m2* formulates *h2*, which explains *p2*, and so on. Properties of the hypothesis lattices and operations over them are considered in [24].

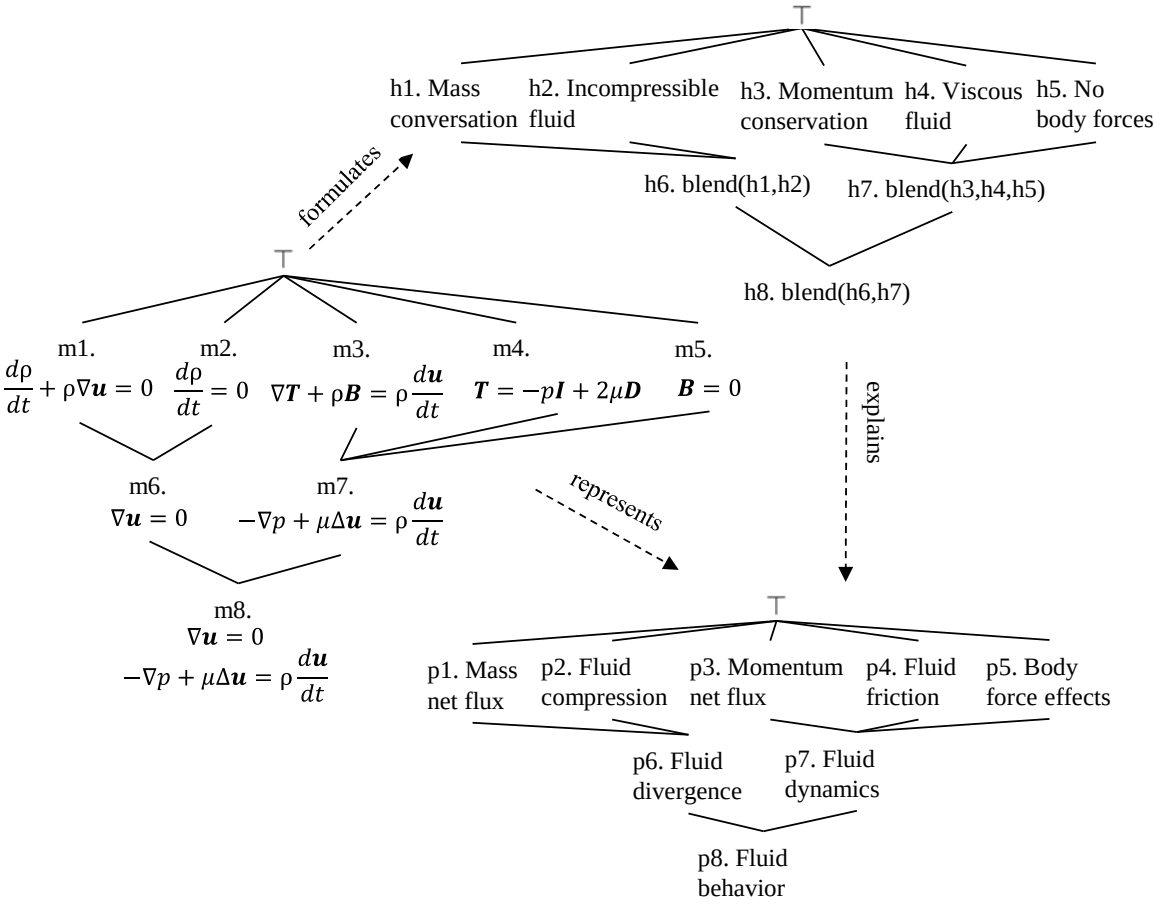



Fig. 5. Hypothesis lattice unfolded into model and phenomenon isomorphic lattice

*Models* are one of the principal instruments of modern science. Models can perform two fundamentally different representational functions: a model can be a representation of a selected part of the world, or a model can represent a theory in the sense that it interprets the laws and hypotheses of that theory.

Here we consider scientific models to be representations in both senses at the same time. One of the most perplexing questions in connection with models is how they relate to theories. In this respect models can be considered as a complement to theories, as preliminary theories, can be used as substitutions of theories when the latter are too complicated to handle. Learning about the model is done through experiments, thought experiments and simulation. Given a set of parameters, a model can generate expectations about how the system will behave in a particular situation. A model and the hypotheses it is based upon are supported when the model generates expectations that match the behavior of its real-world counterpart.

A *law* generalizes a body of observations. Generally, a law represents a group of related undisputable hypotheses using a handful of fundamental concepts and equations to define the rules governing a set of phenomena. A law does not attempt to explain why something happens – it simply states that it does.

Facilities for support of the hypothesis-driven experimentation will be discussed in the remaining sections.

## 3 Hypothesis manipulation in scientific experiments

### 3.1 Hypothesis generation

Researchers that support rationality of scientific discovery presented several methods for hypothesis generation, including discovery as abduction, induction, anomaly detection, heuristics programming and use of analogies [73].

*Discovery as abduction* characterizes reasoning processes that take place before a new hypothesis is justified. The abductive model of reasoning that leads to plausible hypotheses formulation is conceptualized as an inference beginning with data. According to [50] an abduction happens as follows: 1) Some phenomena *p1, p2, p3,* … are encountered for which there is no or little explanation; 2) However, *p1, p2, p3,* … would not be surprising if a hypothesis *H* were added. They would certainly follow from something like *H* and would be explained by it; 3) Therefore there is good reason for elaborating an hypothesis *H* – for proposing it as a possible hypothesis from which the assumption *p1, p2, p3,* … might follow. The abductive model of reasoning is primarily a process of explaining anomalies or surprising phenomena [63]. The scientists' reasoning proceeds abductively from an anomaly to an explanatory hypothesis in light of which the phenomena would no longer be surprising. There can be several different hypotheses that can serve as the explanations for phenomena, so additionally some criteria for choosing among different hypotheses are required.

One way to implement abductive model of reasoning is the abductive logic programming [36]. Hypothesis generation in abduction logical framework is organized as follows. During the experiment, some new observations are encountered. Let *B* represents the background knowledge; *O* is the set of facts that represents observations. Both *B* and *O* are logic programs (set of rules in some rule language). In addition, *Γ* stands for a set of literals representing the set of abducibles, which are candidate assumptions to be added to *B* for explaining *O*. Given *B*, *O* and *Γ*, the hypothesis-generation problem is to find a set *H* of literals (called a hypothesis) such that: 1) *B* and *H* entail *O*, 2) *B* and *H* is consistent, and 3) *H* is some subset of *Γ*. If all conditions are met then *H* is an explanation of *O* (with respect to *B* and *Γ*). Examples of abductive logic programming systems include ACLP [35], A-system [71], ABDUAL [2] and ProLogICA [59]. Abductive logic programming can also be implemented by means of Answer Set Programming systems, e.g. by the DLV system [14].

The example abductive logic program in ProLogICA describes a simple model of the lactose metabolism of the bacterium E.Coli [59]. The background knowledge *B* describes that E. coli can feed on the sugar lactose if it makes two enzymes permease and galactosidase. Like all enzymes (E), these are made if they are coded by a gene (G) that is expressed. These enzymes are coded by two genes (lac(y) and lac(z)) in cluster of genes (lac(X)) called an operon that is expressed when the amounts (amt) of glucose are low and lactose are high or when they are both at medium level. The abducibles, *Γ*, declare all ground instances of the predicates "amount" as assumable. This reflects the fact that in the model it is not known what are the amounts at any time of the various substances. This is incomplete information that we want to find out in each problem case that we are examining. The integrity constraints state that the amount of a substance (S) can only take one value.

```
##  Background Knowledge (B)
feed(lactose):-
make(permease),make(galactosidase).
make(Enzyme):- code(Gene,Enzyme),express(Gene).
express(lac(X)):-
amount(glucose,low),amount(lactose,hi).
express(lac(X)):-
amount(glucose,medium),amount(lactose,medium).
code(lac(y),permease).
code(lac(z),galactosidase).
temperature(low):-amount(glucose,low).
false :- amount(S,V1), amount(S,V2), V1 != V2.

##  Abducibles (Γ)
abducible_predicate(amount).

## Observation (O)
feed(lactose).

This goal generates two possible hypotheses:
{amount(lactose,hi), amount(glucose,low)}
{amount(lactose,medium),amount(glucose,medium)}
```

Just a couple of another examples of real rule-based systems, where abductive logic programming is used.

Robot Scientist (see 4.4) abductively hypothesizes new facts about the yeast functional biology by inferring what is missing from a model [38]. In [68], both abduction and induction are used to formulate hypotheses about inhibition in metabolic pathways. Augmenting background knowledge is done with abduction, after that induction is used for learning general rules. In [33] authors use SOLAR reasoning system to abductively generate hypotheses about the inhibitory effects of toxins on the rat metabolisms.

The process of discovery is deeply connected also with the search of *anomalies*. There are a lot of methods and algorithms to discover anomalies. Anomaly detection is an important research problem in data mining that aims to find objects that are considerably dissimilar, exceptional and inconsistent with respect to the majority data in an input database [6].

*Analogies* play several roles in science. Not only do they contribute to discovery but they also play a role in the development and evaluation of scientific theories (new hypotheses) by analogical reasoning.

### 3.2 Hypothesis evaluation

Being testable and falsifiable, a scientific hypothesis provides a solid basis to its further modeling and testing. There are several ways to do it, including the use of statistics, machine learning and logic reasoning techniques.

#### 3.2.1 Statistical testing of hypotheses

The classical (*frequentist*) and *Bayesian* statistic approaches are applicable for hypothesis testing and selection. Brief summary of the basic differences between these approaches are as follows [34].

Classical (frequentist) statistics is based on the following beliefs:

− Probabilities refer to *relative frequencies of events*. They are objective properties of the real world;

− Parameters of hypotheses (models) are *fixed, unknown constants*. Because they are not fluctuating, probability statements about parameters are meaningless;

− Statistical procedures should have well-defined long-run frequency properties.

In contrast, Bayesian approach takes the following assumptions:

− Probability describes the degree of subjective belief, not the limiting frequency. Probability statements can be made about things other than data, including hypotheses (models) themselves as well as their parameters;

− Inferences about a parameter are made by producing its probability distribution — this distribution quantifies the uncertainty of our knowledge about that parameter. Various point estimates, such as expectation value, may then be readily extracted from this distribution.

The Bayesian interpretation of probability can be seen as an extension of propositional logic that enables reasoning with hypotheses, i.e., the propositions whose truth or falsity is uncertain.

Bayesian probability belongs to the category of evidential probabilities; to evaluate the probability of a hypothesis, the Bayesian probabilist specifies some prior probability, which is then updated in the light of new, relevant data (evidence) [64]. The Bayesian interpretation provides a standard set of procedures and formulae to perform this calculation.

**Hypothesis testing in classical statistic style**. After null and alternative hypotheses are stated, some statistical assumptions about data samples should be done, e.g. assumptions about statistical independence or distributions of observations. Failing to provide correct assumptions leads to the invalid test results.

A common problem in classical statistics is to ask whether a given sample is consistent with some hypothesis. For example, we might be interested in whether a measured value $x_i$, or the whole set $\{x_i\}$, is consistent with being drawn from a Gaussian distribution $N(\mu,\sigma)$. Here $N(\mu,\sigma)$ is our *null hypothesis.*

It is always assumed that we know how to compute the probability of a given outcome from the null hypothesis: for example, given the cumulative distribution function, $0 \leq H_0(x) \leq 1$, the probability that we would get a value at least as large as $x_i$ is $p(x > x_i) = 1 - H_0(x_i)$, and is called the *p-value*. Typically, a threshold $p$ value is adopted, called *the significance level* $\alpha$, and the null hypothesis is rejected when $p \leq \alpha$ (e.g., if $\alpha = 0.05$ and $p < 0.05$, the null hypothesis is rejected at a 0.05 significance level). If we fail to reject a hypothesis, it does not mean that we proved its correctness because it may be that our sample is simply not large enough to detect an effect.

When performing these tests, we can meet with two types of errors, which statisticians call *Type I and Type II errors*. Type I errors are cases when the null hypothesis is true but incorrectly rejected. In the context of source detection, these errors represent spurious sources, or more generally, false positives (with respect to the alternative hypothesis). The false-positive probability when testing a single datum is limited by the adopted significance level $\alpha$. Cases when the null hypothesis is false, but it is not rejected are called Type II errors (missed sources, or false negatives (again, with respect to the alternative hypothesis)). The false-negative probability when testing a single datum is usually called $\beta$, and is related to *the power of* $\alpha$ *test* as $(1 - \beta)$. Hypothesis testing is intimately related to comparisons of distributions.

As the significance level $\alpha$ is decreased (the criterion for rejecting the null hypothesis becomes more conservative), the number of false positives decreases and the number of false negatives increases. Therefore, there is a trade-off to be made to find an optimal value of $\alpha$, which depends on the relative importance of false negatives and positives in a particular problem. Both the acceptance of false hypotheses and the rejection of true ones are errors that scientists should try to avoid. There is discussion as to what states of affairs is *less* desirable;

many people think that the acceptance of a false hypothesis is always worse than failure to accept a true one and that science should in the first place try to avoid the former kind of error.

When many instances of hypothesis testing are performed, a process called *multiple hypothesis testing*, the fraction of false positives can significantly exceed the value of $\alpha$. The fraction of false positives depends not only on $\alpha$ and the number of data points, but also on the number of true positives (the latter is proportional to the number of instances when an alternative hypothesis is true).

Depending on data type (discrete vs. continuous random variables) and what we can assume (or not) about the underlying distributions, and the specific question we ask, we can use different statistical tests. The underlying idea of statistical tests is to use data to compute an appropriate statistic, and then compare the resulting data-based value to its expected distribution. The expected distribution is evaluated by *assuming that the null hypothesis is true*. When this expected distribution implies that the data-based value is unlikely to have arisen from it by chance (i.e., the corresponding $p$ value is small), the null hypothesis is rejected with some threshold probability $\alpha$, typically 0.05 or 0.01 ($p < \alpha$). Note again that $p > \alpha$ does *not* mean that the hypothesis is *proven* to be correct.

The number of various statistical tests in the literature is overwhelming and their applicability is often hard to decide (see [19, 31] for variety of statistical methods in SPSS). When the distributions are not known, tests are called nonparametric, or distribution-free tests. The most popular nonparametric test is the Kolmogorov–Smirnov (K-S) test, which compares the cumulative distribution function, $F(x)$, for two samples, $\{x_{1i}\}$, $i = 1, ..., N_1$ and $\{x_{2i}\}$, $i = 1, ..., N_2$. The K-S test is not the only option for nonparametric comparison of distributions. The Cramér-von Mises criterion, the Watson test, and the Anderson–Darling test are similar in spirit to the K-S test, but consider somewhat different statistics. The Mann–Whitney–Wilcoxon test (or the Wilcoxon rank-sum test) is a nonparametric test for testing whether two data sets are drawn from distributions with different location parameters (if these distributions are known to be Gaussian, the standard classical test is called the $t$ test). A few standard statistical tests can be used when we know, or can assume, that both $h(x)$ and $f(x)$ are Gaussian distributions (e.g., the Anderson–Darling test, the Shapiro–Wilk test) [34]. More on statistical tests can be found in [19, 31, 32, 34].

**Hypothesis (model) selection and testing in Bayesian style.** The Bayesian approach can be thought of as formalizing the process of continually refining our state of knowledge about the world, beginning with no data (as encoded by the *prior*), then updating that by multiplying in the likelihood once the data are observed to obtain the *posterior*. When more data are taken, then the posterior based on the first data set can be used as the prior for the second analysis. Indeed, the data sets can be different.

The question often arises as to which is the 'best' model (hypothesis) to use; '*model selection*' is a technique that can be used when we wish to discriminate between competing models (hypotheses) and identify the best model (hypothesis) in a set, $\{M_1, ..., M_n\}$, given the data.

We need to remind the basic notation. The Bayes theorem can be applied to calculate the posterior probability $p(M_j|d)$ for each model (or hypothesis) $M_j$ representing our state of knowledge about the truth of the model (hypothesis) in the light of the data $d$ as follows:

$$p(M_j|d) = p(d|M_j)\, p(M_j) / p(d),$$

where $p(M_j)$ is the prior belief in the model (hypothesis) that represents our state of knowledge (or ignorance) about the truth of the model (hypothesis) before we have analyzed the current data, $p(d|M_j)$ is the model (hypothesis) *likelihood* (represents the probability that some data are produced under the assumption of this model) and $p(d)$ is a normalization constant given by:

$$p(d) = \sum_i p(d|M_i)\, p(M_i)\,.$$

The relative 'goodness' of models is given by a comparison of their posterior probabilities, so to compare two models $M_a$ and $M_b$, we look at the ratio of the model posterior probabilities:

$$p(M_a|d) / p(M_b|d) = p(d|M_a)\, p(M_a) / p(d|M_b)\, p(M_b).$$

The Bayes factor, $B_{ab}$ can be computed as the ratio of the model likelihoods:

$$B_{ab} = p(d|M_a) / p(d|M_b).$$

Empirical scale for evaluating the strength of evidence from the Bayes factor $B_{ij}$ between two models is shown in Tabl. 1 [45].

Tabl. 1. Strength of evidence for Bayes factor $B_{ij}$ for two models

| $\lvert\ln B_{ij}\rvert$ | Odds | Strength of evidence |
|---|---|---|
| < 1.0 | < 3 : 1 | Inconclusive |
| 1.0 | ~ 3 : 1 | Weak evidence |
| 2.5 | ~ 12 : 1 | Moderate evidence |
| 5.0 | ~ 150 : 1 | Strong evidence |

The Bayes factor gives a measure of the 'goodness' of a model, regardless of the prior belief about the model; the higher the Bayes factor, the better the model is. In many cases, the prior belief in each model in the set of proposed models will be equal, so the Bayes factor will be equivalent to the ratio of the posterior probabilities of the models. The 'best' model in the Bayesian sense is the one which gives the best fit to the data with the smallest parameter space.

A special case of model (hypothesis) selection is *Bayesian hypothesis testing* [34, 62]. Taking $M1$ to be the "null" hypothesis, we can ask whether the data supports the alternative hypothesis $M2$, i.e., whether we can reject the null hypothesis. Taking equal priors $p(M_1) = p(M_2)$, the odds ratio is

$$B_{21} = p(d|M_1) / p(d|M_2).$$

The inability to reject $M_1$ in the absence of an alternative hypothesis is very different from the hypothesis testing procedure in classical statistics. The latter procedure rejects the null hypothesis if it does not provide a good description of the data, that is, when it is very unlikely that the given data could have been generated as prescribed by the null hypothesis. In contrast, the Bayesian approach is based on the posterior rather than on the data likelihood, and cannot reject a hypothesis if there are no alternative explanations for observed data [34].

Comparing classical and Bayesian approaches [34], it is rare for a mission-critical analysis be done in the "fully Bayesian" manner, i.e., without the use of the frequentist tools at the various stages. Philosophy and beauty aside, the reliability and efficiency of the underlying computations required by the Bayesian framework are the main practical issues. A central technical issue at the heart of this is that it is much easier to do optimization (reliably and efficiently) in high dimensions than it is to do integration in high dimensions. Thus the usable machine learning methods, while there are ongoing efforts to adapt them to Bayesian framework, are almost all rooted in frequentist methods.

Most users of Bayesian estimation methods, in practice, are likely to use a mix of Bayesian and frequentist tools. The reverse is also true—frequentist data analysts, even if they stay formally within the frequentist framework, are often influenced by "Bayesian thinking," referring to "priors" and "posteriors." The most advisable position is probably to know both paradigms well, in order to make informed judgments about which tools to apply in which situations [34]. More details on Bayesian style of hypothesis testing can be found in [34, 62, 64].

### 3.2.2 Logic-based hypothesis testing

According to the hypothetico-deductive approach the hypotheses are tested by deducing predictions or other empirical consequences from general theories. If these predictions are verified by experiments, this supports the hypothesis. It should be noted that not anything that is logically entailed by a hypothesis can be confirmed by a proper test for it. The relation between hypothesis and evidence is often *empirica l* rather than logical. A clean deduction of empirical consequences from a hypothesis, as it may sometimes exist in physics, is practically inapplicable in biology. Thus, entailment of the evidence by hypotheses under test is neither sufficient nor necessary for a good test. *Inference to the best explanation* is usually construed as a form of inductive inference (see abduction in 3.1) where a hypothesis' explanatory credentials are taken to indicate its truth [72].

An inductive logic is a system of evidential support that extends deductive logic to less-than-certain inferences. For valid deductive arguments the premises *logically entail* the conclusion, where the entailment means that the truth of the premises provides a *guarantee* of the truth of the conclusion. Similarly, in a good inductive argument the premises should provide some *degree of support* for the conclusion, where such support means that the truth of the premises indicates with some *degree of strength* that the conclusion is true. If the logic of good inductive arguments is to be of any real value, the measure of support it articulates should meet the *Criterion of Adequacy (CoA)*: as evidence accumulates, the *degree* to which the collection of true evidence statements comes to *support* a hypothesis, as measured by the logic, should tend to indicate that the hypotheses are probably false or probably true. In [27] the extent to which a kind of logic based on the Bayes theorem can estimate how the *implications of hypotheses about evidence claims* influences the degree to which hypotheses are supported is discussed in detail. In particular, it is shown how such a logic may be applied to satisfy the CoA: as evidence accumulates, false hypotheses will very probably come to have evidential support values (as measured by their *posterior probabilities*) that approach 0; and as this happens, a true hypothesis will very probably acquire evidential support values (measured by their *posterior probabilities*) that approach 1.

### 3.2.3 Parameter estimation

Models (hypotheses) are typically described by parameters $\boldsymbol{\theta}$ whose values are to be estimated from data. We describe this process according to [34]. For a particular model $M$ and prior information $I$ we get:

$$p(M, \boldsymbol{\theta}|d, I) = p(d|M, \boldsymbol{\theta}, I)\, p(M, \boldsymbol{\theta}|I) / p(d|I)$$

The result $p(M, \boldsymbol{\theta}|d, I)$ is called the *posterior* probability density function (pdf) for model $M$ and parameters $\boldsymbol{\theta}$, given data $d$ and other prior information $I$. This term is a $(k + 1)$-dimensional pdf in the space spanned by $k$ model parameters and the model $M$. The term $p(d|M, \boldsymbol{\theta}, I)$ is the *likelihood* of data *given* some model $M$ and some fixed values of parameters $\boldsymbol{\theta}$ describing it, and all other prior information $I$. The term $p(M, \boldsymbol{\theta}|I)$ is the a priori joint probability for model $M$ and its parameters $\boldsymbol{\theta}$ in the absence of any of the data used to compute likelihood, and is often simply called the *prior*.

In the Bayesian formalism, $p(M, \boldsymbol{\theta}|d, I)$ corresponds to the state of our *knowledge* (i.e., belief) about a model and its parameters, given data $d$. To simplify the notation, $M(\boldsymbol{\theta})$ will be substituted by $M$ whenever the absence of explicit dependence on $\boldsymbol{\theta}$ is not confusing. A completely Bayesian data analysis has the following conceptual steps:

1. Formulation of the data likelihood $p(d|M, I)$.
2. Choice of the prior $p(\boldsymbol{\theta}|M,I)$, which incorporates all other knowledge that might exist, but is *not* used when computing the likelihood (e.g., prior measurements of the same type, different measurements, or simply an uninformative prior). Several methods for constructing "objective" priors have been proposed. One of them is the *principle of maximum entropy* for assigning uninformative priors by maximizing the entropy over a suitable set of pdfs, finding the distribution that is least informative (given

the constraints). Entropy maximization with no testable information takes place under a single constraint: the sum of the probabilities must be one. Under this constraint, the maximum entropy for a discrete probability distribution is given by the uniform distribution.

3. Determination of the posterior $p(M|d, I)$, using Bayes theorem above. In practice, this step can be computationally intensive for complex multidimensional problems.

4. The search for the best model $M$ parameters, which maximizes $p(M|d, I)$, yielding the *maximum a posteriori* (MAP) estimate. This *point estimate* is the natural analog to the *maximum likelihood estimate* (MLE) from classical statistics.

5. Quantification of uncertainty in parameter estimates, via *credible regions*. As in MLE, such an estimate can be obtained analytically by doing mathematical derivations specific to the chosen model. Also as in MLE, various numerical techniques can be used to simulate samples from the posterior. This can be viewed as an analogy to the frequentist approach, which can simulate draws of samples from the true underlying distribution of the data. In both cases, various descriptive statistics can then be computed on such samples to examine the uncertainties surrounding the data and estimators of model parameters based on that data.

6. *Hypothesis testing* as needed to make other conclusions about the model (hypothesis) or parameter estimates.

### 3.3 Algorithmic generation and evaluation of hypotheses

Two cultures of data analysis (*formulaic modeling*[1] and *algorithmic modeling*) distinguished here in accordance with [10] can be applied to the hypothesis extraction and generation based on data.

*Formulaic modeling* is a process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the formulae $\boldsymbol{y} = f(\boldsymbol{x})$ that give a relation specifying a vector of dependent variables $\boldsymbol{y}$ in terms of a vector of independent variables $\boldsymbol{x}$. In a statistics experiment (based on various regression techniques) the dependent variable defines the event studied and is expected to change whenever the independent variable (*predictor* variables, extraneous variables) is altered. Such methods as linear regression, logistic regression, multiple regression are well-known examples of the representatives of this modeling approach.

In the *algorithmic modeling* culture the approach is to find an algorithm that operates on $\boldsymbol{x}$ to predict the responses $\boldsymbol{y}$. What is observed is a set of $\boldsymbol{x}$'s that go in and a subsequent set of $\boldsymbol{y}$'s that come out. Predictive accuracy and properties of the algorithms (such as, e.g., their convergence if they are iterative) are the issues to be investigated. *Machine learning algorithms* focus on prediction, based on known properties learned from the training data. Such machine learning algorithms as decision tree, association rule, neural networks, support vector machines as well as other techniques of learning in Bayesian and probabilistic models [5, 26] are examples of the methods that belong to this second culture.

The models that best emulate the nature in terms of predictive accuracy are also the most complex and inscrutable. Nature forms the outputs $\boldsymbol{y}$ from the inputs $\boldsymbol{x}$ by means of a black box with complex and unknown interior. Current accurate prediction methods are also *complex black boxes* (such as neural nets, forests, support vectors). So we are facing two black boxes, where ours seems only slightly less inscrutable than nature's [10]. In a choice between *accuracy* and *interpretability*, in applications people sometimes prefer interpretability.

However, the goal of a model is not interpretability (a way of getting information), but getting useful, accurate information about the relation between the response and predictor variables. It is stated in [10] that algorithmic models can give better predictive accuracy than formulaic models, providing also better information about the underlying mechanism. And actually this is what the goal of statistical analysis is. The researchers should be focused on solving the problems instead of asking what regression model they can create.

An objection to this idea (expressed by Cox) is that prediction without some understanding of underlying process and linking with other sources of information becomes more and more tentative. Due to that it is suggested to construct the stochastic calculation models that summarize the understanding of the phenomena under study. One of the objectives of such approach might be an understanding and test of hypotheses about underlying process. Given the relatively small sample size following such direction could be productive. But data characteristics are rapidly changing. In many of the most interesting current problems, the idea of starting with a formal model is not tenable. The methods used in statistics for small sample sizes and a small number of variables are not applicable. Data analytics need to be more pragmatic. Given a statistical problem, find a good solution, whether it is a formulaic model, an algorithmic model or a Bayesian model or a completely different approach.

In the context of the hypothesis driven analysis we should pay attention to the question how far can we go applying the algorithmic modeling for hypothesis generation and testing. Various approaches to machine learning use related to hypothesis formation and selection can be found in [5, 10, 34].

Besides machine learning, an interesting example of algorithmic generation of hypotheses can be found in the IBM Watson project [18] where the symbiosis of the general-purpose reusable natural language processing

[1] In [10] instead of "formulaic modeling" the term "data modeling" is used that looks misleading in the computer science context.

(NLP) and knowledge representation and reasoning (KRR) technologies (under the name DeepQA) is exploited for answering arbitrary questions over the existing natural language documents as well as structured data resources. Hypothesis generation takes the results of question analysis and produces candidate answers by searching the available data sources and extracting answer-sized snippets from the search results. Each candidate answer plugged back into the question is considered a hypothesis, which the system has to prove correct with some degree of confidence. After merging, the system must rank the hypotheses and estimate confidence based on their merged scores. A machine-learning approach adopted is based on running the system over a set of training questions with known answers and training a model based on the scores. An important consideration in dealing with NLP-based scorers is that the features they produce may be quite sparse, and so accurate confidence estimation requires the application of confidence-weighted learning techniques [18] – a new class of online learning methods that maintain a probabilistic measure of confidence in each parameter. It is important to note that instead of statistics based hypothesis testing, contextual evaluation of a wide range of loosely coupled probabilistic question and semantic based content analytics is applied for scoring different questions (hypotheses) and content interpretations. Training different models on different portions of the data in parallel and combining the learned classifiers into a single classifier allows to make the process applicable to the large collections of data. More details on that can be found in [17, 18] as well as in other Watson project related publications.

### 3.4 Bayesian motivation for discovery

One way for discriminating between competing models of some phenomenon is to use Bayesian model selection approach (3.2.1), the Bayesian evidences for each of the proposed models (hypotheses) can be computed and the models can then be ranked by their Bayesian evidence. This is a good method for identifying which is the best model in a given set of models, but it gives no indication of the *absolute goodness* of the model. Bayesian model selection says nothing about the *overall quality* of the set of models (hypotheses) as a whole —the best model in the set may merely be the best of in a set of poor models. Knowing that the best model in the current set of models is not particularly good model would provide *motivation to search for a better model*, and hence may lead to model discovery.

One way of assigning some measure of the absolute goodness of a model is to use the concept of Bayesian doubt, first introduced by [67]. Bayesian doubt works by comparing all the known models in a set with an idealized model, which acts as a benchmark model.

An application of the Bayesian doubt method for the cosmological model building is given in [44, 45]. One of the most important questions in cosmology is to identify the fundamental model underpinning the vast amount of observations nowadays available. The so-called 'cosmological concordance model' is based on the cosmological principle (i.e. the Universe is isotropic and homogeneous, at least on large enough scales) and on the hot big bang scenario, complemented by an inflationary epoch. This remarkably simple model is able to explain with only half a dozen free parameter observations spanning a huge range of time and length-scales. Since both a cold dark matter (CDM) and a cosmological constant ($\Lambda$) component are required to fit the data, the concordance model is often referred to as 'the $\Lambda$CDM model'.

Several different types of explanation are possible for the apparent late time acceleration of the Universe, including different classes of dark energy model such as $\Lambda$CDM, *w*CDM; theories of modified gravity; void models or the back reaction [45]. The methodology of Bayesian doubt which gives an absolute measure of the degree of goodness of a model has been applied to the issue of whether the $\Lambda$CDM model should be doubted.

The methodology of Bayesian doubt dictates that an unknown idealized model $X$ should be introduced against which the other models may be compared. Following [67], 'doubt' may be defined as the posterior probability of the unknown model:

$$D \equiv p(X|d) = p(d|X)\, p(X) / p(d).$$

Here $p(X)$ is the prior doubt, i.e. the prior on the unknown model, which represents the degree of belief that the list of known models does not contain the true model. The sum of all the model priors must be unity.

The methodology of Bayesian doubt requires a baseline model (the best model in the set of known models), for which in this application the $\Lambda$CDM has been chosen. The average Bayes factor between $\Lambda$CDM and each of the known models is given by:

$$<B_{i\Lambda}> \equiv 1/\mathrm{N} \sum_{i=1}^{N} B_{i\Lambda} \ .$$

The ratio $R$ between the posterior doubt and prior doubt, which is called the relative change in doubt, is:

$$R \equiv D/p(X).$$

For doubt to grow, i.e. the posterior doubt to be greater than the prior doubt ($R << 1$), the Bayes factor between the unknown model $X$ and the baseline model must be much greater than the average Bayes factor:

$$<B_{i\Lambda}> / B_{X\Lambda} << 1.$$

To genuinely doubt the baseline model, $\Lambda$CDM, it is not sufficient that $R > 1$, but additionally, the probability of $\Lambda$CDM must also decrease such that its posterior probability is greater than its prior probability, i.e. $p(\Lambda|d) < p(\Lambda)$. We can define:

$$R_\Lambda \equiv p(\Lambda|d) / p(\Lambda).$$

For $\Lambda$CDM to be doubted, the following two conditions must be fulfilled:

$$R > 1, \ R_\Lambda < 1.$$

If these two conditions are fulfilled, then it suggests that the set of known models is incomplete, and gives motivation to search for a better model not yet included, which may lead to model discovery.

In [67] a way of computing an absolute upper bound for $p(d|X)$ achievable among the class of known models has been proposed. Finally it was found that current cosmic microwave background (CMB), matter power spectrum (mpk) and Type Ia supernovae (SNIa) observations.do not require the introduction of an alternative model to the baseline ΛCDM model. The upper bound of the Bayesian evidence for a presently unknown dark energy model against ΛCDM gives only weak evidence in favor of the unknown model. Since this is an absolute upper bound, it was concluded that ΛCDM remains a sufficient phenomenological description of currently available observations.

# 4 Facilities for the scientific hypothesis-driven experiment support

## 4.1 Conceptualization of scientific experiments

DIS increasingly becomes dependent on computational resources to aid complex researches. It becomes paramount to offer scientists mechanisms to manage the variety of knowledge produced during such investigations. Specific conceptual modeling facilities [54] are investigated to allow scientists to represent scientific hypotheses, models and associated computational or simulation interpretations which can be compared against phenomena observations (Fig. 3). The model allows scientists to record the existing knowledge about an observable investigated phenomenon, including a formal mathematical interpretation of it, if any. Model evolution and model sharing need also to be supported taking either a mathematical or computational view (e.g., expressed by scientific workflows). Declarative representation of scientific model allows scientists to concentrate on the scientific issues to be investigated. Hypotheses can be used also to bridge the gap between an ontological description of studied phenomena and the simulations. Conceptual views on scientific domain entities allow for searching for definitions supporting scientific models sharing among different scientific groups.

In [23] the engineering of hypothesis as linked data is addressed. A semantic view on scientific hypotheses shows their existence apart from a particular statement formulation in some mathematical framework. The mathematical equation is considered as not enough to identify the hypothesis, first because it must be physically interpreted, second because there can be many ways to formulate the same hypothesis. The link to a mathematical expression, however, brings to the hypothesis concept higher semantic precision. Another link, in addition, to an explicit description of the explained phenomenon (emphasizing its “physical interpretation") can bring forth the intended meaning. By dealing with that hypothesis as a conceptual entity, the scientists make it possible to change its statement formulation or even to assert a semantic mapping to another incarnation of the hypothesis in case someone else reformulates it.

In [54] the following elements related to hypothesis driven science are conceptualized: a phenomenon observed, a model interpreting this phenomenon, the metadata defining the related computation together with the simulation definition (for simulation a declarative logic-based language is proposed). Specific attention in this work is devoted to hypothesis definition. The explanation a scientific hypothesis conveys is a relationship between the causal phenomena and the simulated one, namely, that the simulated phenomenon is caused by or produced under the conditions set by the causal phenomena. By running the simulations defined by the antecedents in the causal relationship, the scientist aims at providing hypothetical analysis of the studied phenomenon.

Thus, the scientific hypothesis becomes an element of the scientific model that may replace a phenomenon. When computing a simulation based on a scientific hypothesis, i.e. according to the causal relationship it establishes, the output results may be compared against phenomenon observations to assess the quality of the hypothesis. Such interpretation provides for bridging the gap between qualitative description of the phenomenon domain (scientific hypotheses may be used in qualitative (i.e., ontological) assertions) and the corresponding quantitative valuation obtained through simulations. According to the approach [54], complex scientific models can be expressed as the composition of computation models similarly to database views.

## 4.2 Hypothesis space browsers

In the HyBrow (Hypothesis Space Browser) project [58] the hypotheses for the biology domain are represented as a set of first-order predicate calculus sentences. In conjunction with an axiom set specified as rules that model known biological facts over the same universe, and experimental data, the knowledge base may contradict or validate some of the sentences in hypotheses, leaving the remaining ones as candidates for new discovery. As more experimental data is obtained and rules identified, discoveries become positive facts or are contradicted. In the case of contradictions, the rules that caused the problems must be identified and eliminated from the theory formed by the hypotheses. In such model-theoretical approach, the validation of hypotheses considers the satisfiability of the logical implications defined in the model with respect to an interpretation. This might be applicable also for simulation-based research, in which validation is decided based on the quantitative analysis between the simulation results and the observations [54]. HyBrow is based on an OWL ontology and application-level rules to contradict or validate hypothetical statements. HyBrow provides for designing hypotheses, and evaluating them for consistency with existing knowledge, uses an ontology of hypotheses to represent hypotheses in machine understandable form as relations between objects (agents) and processes [65].

As an upgrade of HyBrow, the HyQue [12] framework adopts linked data technologies and employs Bio2RDF linked data to add to HyBrow semantic

interoperability capabilities. HyBrow/HyQue's hypotheses are domain-specific statements that correlate biological processes (seen as events) in the First-Order Logic (FOL). Hypotheses are formulated as instances of the HyQue Hypothesis Ontology and are evaluated through a set of SPARQL queries against biologically-typed OWL and HyBrow data. The query results are scored in terms of how the set of events correspond to background expectations. A score indicates the level of support the data lends the hypothesis. Each event is evaluated independently in order to quantify the degree of support it provides for the hypothesis posed. Hypothesis scores are linked as properties to the respective hypothesis.

OBI (the Ontology for Biomedical Investigations) project (http://obi-ontology.org) aims to model the design of an investigation: the protocols, the instrumentation, and materials used in experiments and the data generated [20]. Ontologies such as EXPO and OBI enable the recording of the whole structure of scientific investigations: how and why an investigation was executed, what conclusions were made, the basis for these conclusions, etc. As a result of these generic ontology development efforts, the Minimum Information about a Genotyping Experiment (MIGen) recommends the use of terms defined in OBI. The use of a generic or a compliant ontology to supply terms will stimulate cross-disciplinary data-sharing and reuse. As much detail about an investigation as possible in order to make the investigation more reproducible and reusable can be collected [39].

Hypothesis modeling is embedded into the knowledge infrastructures being developed in various branches of science. One example of such infrastructure is considered under the name SWAN – a SemanticWeb Application in Neuromedicine [20]. SWAN is a project for developing an integrated knowledge infrastructure for the Alzheimer disease (AD) research community. SWAN incorporates the full biomedical research knowledge lifecycle in its ontological model, including support for personal data organization, hypothesis generation, experimentation, laboratory data organization, and digital pre-publication collaboration. The common ontology is specified in an RDF Schema. SWAN's content is intended to cover all stages of the "truth discovery" process in biomedical research, from formulation of questions and hypotheses, to capture of experimental data, sharing data with colleagues, and ultimately the full discovery and publication process.

Several information categories created and managed in SWAN are defined as subclasses of Assertion. They include Publication, Hypothesis, Claim, Concept, Manuscript, DataSet, and Annotation. An Assertion may be made upon any other Assertion, or upon any object specifiable by URL. For example, a scientist can make a Comment upon, or classify, the Hypothesis of another scientist. Linking to objects "outside" SWAN by URL allows one to use SWAN as metadata to organize – for example – all one's PDFs of publications, or the Excel files in which one's laboratory data is stored, or all the websites of tools relevant to Neuroscience. Annotation may be structured or unstructured. Structured annotation means attaching a Concept (tag or term) to an Assertion. Unstructured annotation means attaching free text. Concepts are nodes in controlled vocabularies, which may also be hierarchical (taxonomies).

### 4.3 Scientific hypothesis formalization

An example showing on Fig. 6 the diversity of the components of a scientific hypothesis model has been borrowed from the applications in Neuroscience [54, 55] and in a human cardiovascular system in Computational Hemodynamics [23, 56]. The formalization of a scientific hypothesis was provided by a mathematical model, by a set of differential equations for continuous processes, quantifying the variations of physical quantities in continuous space-time and by the mathematical solver (HEMOLAB) for discrete processes. The mathematical equations were represented in MathML, enabling models interchange and reuse.

In [3] the formalism of quantitative process models is presented that provides for encoding of scientific models formally as a set of equations and informally in terms of processes expressing those equations. The model revision works as follows. For input it is required an initial model; a set of constraints representing acceptable changes to the initial model in terms of processes; a set of generic processes that may be added to the initial model; observations to which the revised model should fit. These data provide the approach with a heuristic that guides search toward parts of the model space that are consistent with the observations. The algorithm generates a set of revised models that are sorted by their distance from the initial model and presented with their mean squared error on the training data. The distance between a revised model and the initial model is defined as the number of processes that are present in one but not in the other. The abilities of the approach have been successfully checked in several environmental domains.

Formalisms for hypothesis formation are mostly monotonic and are considered to be not quite suitable for knowledge representation, especially in dealing with incomplete knowledge, which is often the case with respect to biochemical networks. In [69] knowledge based framework for the general problem of hypothesis formation is presented. The framework has been implemented by extending BioSigNet-RR – a knowledge based system that supports elaboration tolerant representation and non-monotonic reasoning. The main features of the extended system provide: (1) seamless integration of hypothesis formation with knowledge representation and reasoning; (2) use of various resources of biological data as well as human expertise to intelligently generate hypotheses; (3) support for ranking hypotheses and for designing experiments to verify hypotheses. The extended system is positioned as a prototype of an intelligent research assistant of molecular biologists.

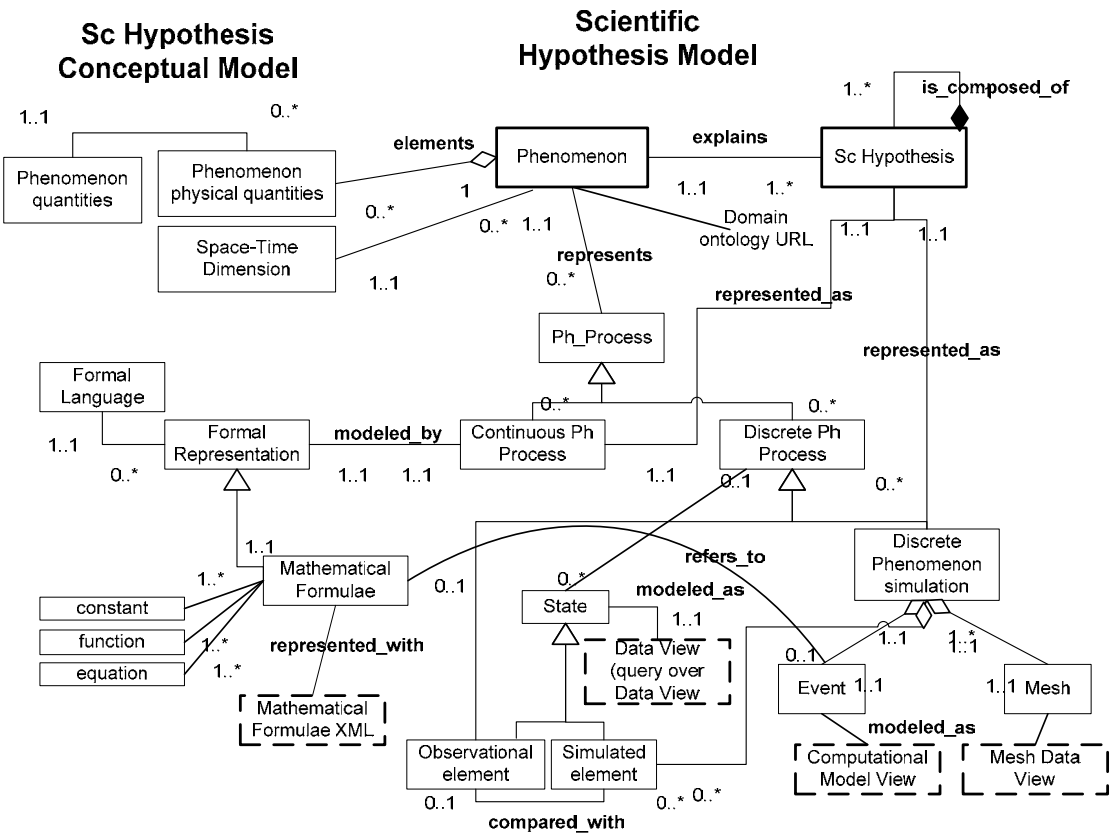



Fig. 6. Elements of the scientific hypothesis model

### 4.4 Hypothesis-driven robots

The Robot Scientist [66] oriented on genomic applications is a physically implemented system which is capable of running cycles of scientific experimentation and discovery in a fully automatic manner: hypothesis formation, experiment selection to test these hypotheses, experiment execution using robotic system, results analysis and interpretation, repeating the cycle (closed-loop in which the results obtained are used for learning from them and feeding the resulting knowledge back into the experimental models). Deduction, induction and abduction are types of logical reasoning used in scientific discovery (section 3). The full automation of science requires 'closed-loop learning', where the computer not only analyses the results, but learns from them and feeds the resulting knowledge back into the next cycle of the process (Fig. 6).

In the Robot Scientist the automated formation of hypotheses is based on the following key components:

1. Machine–computable representation of the domain knowledge.

2. Abductive or inductive inference of novel hypotheses.

3. An algorithm for the selection of hypotheses.

4. Deduction of the experimental consequences of hypotheses.

Adam, the first Robot Scientist prototype, was designed to carry out microbial growth experiments to study functional genomics in the yeast *Saccharomyces cerevisiae*, specifically to identify the genes encoding 'locally orphan enzymes'. Adam uses a comprehensive logical model of yeast metabolism, coupled with a bioinformatic database (Kyoto Encyclopaedia of Genes and Genomes – KEGG) and standard bioinformatics homology search techniques (PSI-BLAST and FASTA) to hypothesize likely candidate genes that may encode the locally orphan enzymes. This hypothesis generation process is abductive.

To formalize Adam's functional genomics experiments, the LABORS ontology (LABoratory Ontology for Robot Scientists) has been developed. LABORS is a version of the ontology EXPO (as an upper layer ontology) customized for Robot scientists to describe biological knowledge. LABORS is expressed in OWL-DL. LABORS defines various structural research units, e.g. trial, study, cycle of study and replicate as well as design strategy, plate layout, expected actual results. The respective concepts and relations in the functional genomics data and metadata are also defined. Both LABORS and the corresponding database (used for storing the instances of the classes) are translated into Datalog in order to use the SWI-Prolog reasoner for required applications [39].

There were two types of hypotheses generated. The first level links an orphan enzyme, represented by its enzyme class (E.C.) number, to a gene (ORF) that potentially encodes it. This relation is expressed as a two place predicate where the first argument is the ORF and the

second the E.C. number. An example of hypothesis at this level is: *encodesORFtoEC('YBR166C', '1.1.1.25')*.

The second level of hypothesis involves the association between a specific strain, referenced via the name of its missing ORF, and a chemical compound which should affect the growth of the strain, if added as a nutrient to its environment. This level of hypothesis is derived from the first by logical inference using a specific model of yeast metabolism. An example of such a hypothesis is: *affects growth('C00108','YBR166C')*, where the first argument is the compound (names according to KEGG) and the second argument is the strain considered.

Adam then designs the experimental assays required to test these hypotheses for execution on the laboratory robotic system. These experiments are based on a two-factor design that compares multiple replicates of the strains with and without metabolites compared against wild type strain controls with and without metabolites.

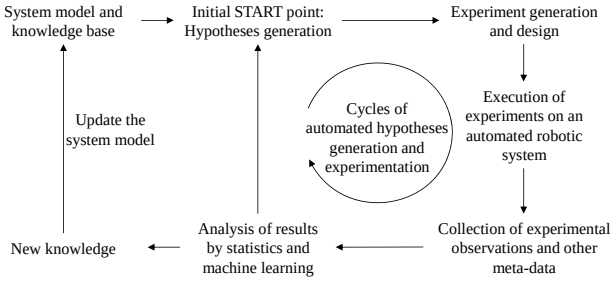



Fig. 6. Hypothesis driven closed-loop learning

Adam follows a hypothetico-deductive methodology (section 2). Adam abductively hypothesizes new facts about yeast functional biology, then it deduces the experimental consequences of these facts using its model of metabolism, which it then experimentally tests. To select experiments Adam takes into account the variable cost of experiments, and the different probabilities of hypotheses. Adam chooses its experiments to minimize the expected cost of eliminating all but one hypothesis. This is in general a NP complete problem and Adam uses heuristics to find a solution [65].

It is now likely that the majority of hypotheses in biology are computer generated. Computers are increasingly automating the process of hypothesis formation, for example: machine learning programs (based on induction) are used in chemistry to help design drugs; and in biology, genome annotation is essentially a vast process of (abductive) hypothesis formation. Such computer-generated hypotheses have been necessarily expressed in a computationally amenable way, but it is still not common practice to deposit them into a public database and make them available for processing by other applications [65].

The details describing the software and informatics decisions in the Robot Scientist project can be found in [65, 66] and online at the website http://www.aber.ac.uk/compsci/Research/bio/robotsci/data/informatics/). The details for developing the formalization used for Adam's functional genomics investigations can be found in [13, 39]. An ontology-based formalization based on graph theory and logical modeling makes it possible to keep an accurate track of all the result units used for different goals, while preserving the semantics of all the experimental entities involved in all the investigations. It is shown how experimentation and machine learning are used to identify additional knowledge to improve the metabolic model [13].

### 4.5 Hypotheses as data in probabilistic databases

Another view of hypotheses encoding and management is presented in [51]. Authors use probabilistic database techniques for hypotheses systematic construction and management. MayBMS [30], a probabilistic database management system, is used as a core for hypothesis management. This methodology (called $\gamma$-DB) enables researchers to maintain several hypotheses explaining some phenomena and provides evaluation mechanism based on Bayesian approach to rank them.

The construction of $\gamma$-DB database comprises several steps. In the first step, phenomenon and hypothesis entities are provided as input to the system. Hypothesis is a set of mathematical equations expressed as functions in W3C MathML-based format and is associated with one or more simulation trial dataset, consisting of tuples with input variables of equation and its corresponding output as functionally dependent (FD) variables (the predictions). Phenomenon is represented by at least one empirical dataset similar to simulation trials. In the next step, the system deals with hypotheses and phenomena in the following way: 1) researcher has to provide some meta data about hypotheses and phenomena; e.g., hypotheses need to be associated with the respective phenomena and assigned a prior confidence distribution (uniform by default according to the principle of maximum entropy (3.2.3)); 2) functional dependencies (FD) are extracted from equations in order to obtain database schema to store simulations and experimental data; it should be mentioned that to precisely identify hypothesis formulation the special attributes for phenomena and hypothesis references are introduced into FD; 3) tuples are synthesized from simulation trials and observational data by uncertain pseudo-transitive closure and reasoning; 4) finally, the probabilistic γ-DB database is formed. Once phenomenon and hypothesis (with empirical datasets and simulation trials) are produced it becomes possible to manipulate them with database tools.

MayBMS provides tools to evaluate competing hypotheses for the explanation of a single phenomenon. With prior probabilities already provided the system allows to make one or more (if new observational data appears) Bayesian inference steps. In each step the prior probability is updated to posterior according to Bayes' theorem. As a result, hypotheses which better explain phenomenon get higher probabilities enabling researchers to make more confident decisions (see also 3.2.1). The $\gamma$-DB approach provides a promising way to analyse hypotheses in large scale DIR as uncertain predictive database in face of empirical data.

## 5 Examples of hypothesis-driven scientific research

### 5.1 Besançon Galaxy model

Various models in astronomy heavily rely on hypotheses. One of the most impressive is the Besançon galaxy model (BGM) [16, 60, 61] evolving for many years and representing the population and structure synthesis model for the Milky Way. It allows astronomers to test hypotheses on the star formation history, star evolution, and chemical and dynamical evolution of the Galaxy. From the beginning, the aim of the BGM was not only to be able to simulate reasonable star counts but further to test scenarios of Galactic evolution from assumptions on the rate of star formation (SFR), initial mass function (IMF), and stellar evolution.

We will further focus on the renewed BGM [16], in which authors draw their attention to the Galaxy thin disk treatment and use of Tycho-2 as a testing dataset. The parameters of BGM (such as IMF, SFR and evolutionary track sets) explicitly and model ingredients implicitly can be treated as hypotheses. Model ingredients include the treatment of binarity, the local stellar mass densities of thin disk, extinction model, age-metallicity and age-velocity relations, radial scale length, the age of the Galaxy thin disc, different sets of the star atmosphere models, etc.

Tycho-2 dataset and $\chi^2$-type statistics test is used to test various versions of these hypotheses in order to choose the most appropriate ones and update model to better fit the provided data. The tests were made by comparing star counts and $(B-V)_T$ colour distributions between data and simulations. Two different tests were used to evaluate the adequacy of the stellar densities globally and to test the shape of the colour distribution.

Due to the fact, that some ingredients of the model are highly correlated (such as the IMF, SFR and the local mass density) the authors defined default models as a combination of a new set of ingredients that significantly improve the fit to Tycho data. So, 11 IMF functions, 2 SFR functions, 2 evolutionary track sets, 3 sets of atmosphere models, 3 values for the age of the formation of the thin disk, 3 sets of values of the thin disk local stellar volume mass density were tested. As a result of testing, the two most appropriate IMS and SFR hypotheses were chosen. Based on this experience, an investigation of the thick disc is underway using SDSS and 2MASS surveys.

### 5.2 Connectome analysis based on network data

In the neuroscience community the development of common paradigms for interrogating the myriad functional systems in the brain remains to be the core challenge. Building on the term "*connectome*," coined to describe the comprehensive map of neural connections in the human brain, the "functional connectome" denotes the collective set of functional connections in the human brain (its "wiring diagram") [7]. More broadly, a connectome would include the mapping of all neural connections within an organism's nervous system. The production and study of connectomes, known as *connectomics*, may range in scale from a detailed map of the full set of neurons and synapses within part or all of the nervous system of an organism to a macro scale description [15] of the functional and structural connectivity between all cortical areas and subcortical structures. The ultimate goal of connectomics is to map the human brain. In functional magnetic resonance imaging (fMRI), associations are thought to represent functional connectivity, in the sense that the two regions of the brain participate together in the achievement of some higher-order function, often in the context of performing some task. fMRI has emerged as a powerful tool used to interrogate a multitude of functional circuits simultaneously. This has elicited the interest of statisticians working in that area. At the level of basic measurements, neuroimaging data can be considered to consist typically of a set of signals (usually time series) at each of a collection of pixels (in two dimensions) or voxels (in three dimensions). Building from such data, various forms of higher-level data representations are employed in neuroimaging. In recent years a substantial interest in network-based representations has emerged in neuroimaging to use *networks* to summarize relational information in a set of measurements, typically assumed to be reflective of either functional or structural relationships between regions of interest in the brain. With neuroimaging now a standard tool in clinical neuroscience, quickly moving towards a time in which we will have available databases composed of large collections of secondary data in the form of *network-based data objects* is predictable.

One of the most basic tasks of interest in the analysis of such data is the testing of hypotheses, in answer to questions such as "Is there a difference between the networks of these two groups of subjects?" Networks are not Euclidean objects, and hence classical methods of statistics do not directly apply. Network-based analogues of classical tools for statistical estimation and hypothesis testing are investigated [21, 22]. Such research is motivated by the 1000 Functional Connectomes Project (FCP) launched in 2010 [7].The 1000 FCP [74] constitutes the largest data set of its kind similarly to large data sets in genetics. Other projects (such as the Human Connectome Project (HCP)) are aimed to build a network map of the human brain in healthy, living adults. The total volume of data produced by the HCP will likely be multiple petabytes [46]. HCP informatics platform includes data management system ConnectomeDB that is based on the XNAT imaging informatics platform [47], a widely used open source system for managing and sharing imaging and related data.

Visualization, processing and analysis of high-dimensional data such as images often requires some kind of preprocessing to reduce the dimensionality of the data and find a mapping from the original representation to a low-dimensional vector space. The assumption is that the original data resides in a low-

dimensional subspace or manifold [11], embedded in the original space. This topic of research is called dimensionality reduction, non-linear dimensionality reduction, including methods for parameterization of data using low-dimensional manifolds as models. Within the neural information processing community this has become known as manifold learning. Methods for manifold learning are able to find non-linear manifold parameterizations of datapoints residing in high-dimensional spaces, very much like Principal Component Analysis (PCA) is able to learn or identify the most important linear subspace of a set of data points (projecting data on a n-dimensional linear subspace which maximizes the variance of the data in the new space).

In [21] necessary mathematical properties associated with a certain notion of a 'space' of networks used to interpret functional neuroimaging connectome-oriented data are established. Extension of the classical statistics tools to network-based datasets, however, appeared to be highly non-trivial. The main challenge in such an extension is due to the fact that networks are not Euclidean objects (for which classical methods were developed) – rather, they are combinatorial objects, defined through their sets of vertices and edges. In [21] it was shown that networks can be associated with certain natural subsets of Euclidean space, and demonstrated that through a combination of tools from geometry, probability on manifolds, and high-dimensional statistical analysis it is possible to develop a principled and practical framework in analogy to classical tools. In particular, an asymptotic framework for one- and two-sample hypothesis testing has been developed. Key to this approach is the correspondence between an undirected graph and its Laplacian, where the latter is defined as a matrix (associating with a network). Graph Laplacian appeared to be particularly appropriate to be used for such matrices. The space of graph Laplacians is used working in certain subsets of Euclidian space which are some submanifolds of the standard Euclidian space.

The 1000 FCP describes functional neuroimaging data from 1093 subjects, located in 24 community-based centers. The mean age of the participants is 29 years, and all subjects were 18 years-old or older. It is of interest to compare the subject-specific networks of males and females in the 1000 FCP data set. In [21] for the 1000 FCP database comparing of networks with respect to the sex of the subjects, over different age group, and over various collection sites is considered. It is shown that it is necessary to compute the means in each subgroup of networks. This was done by constructing the Euclidean mean of the Laplacians for each group of subjects in different age groups. Such group-specific mean Laplacians can then be interpreted as the mean functional connectivity in each group. Such approach provides for building the hypothesis tests about the average of networks or groups of networks to investigate the effect of sex differences on entire networks.

For the 1000 FCP data set it was tested using the two-sample test for Laplacians whether sex differences were significant to influence patterns of brain connectivity. The null hypothesis of no group differences was rejected with high probability. Similarly for the three different age cohorts the null hypothesis of no cohort differences also was rejected with high probability.

On such examples it was shown [21] that the proposed global test has sufficient power to reject the null hypothesis in cases when mass-univariate approach (considered to be the gold standard in fMRI research [43]) fails to detect the differences at the local level. According to the mass-univariate approach statistical analysis is performed iteratively on all voxels to identify brain regions whose fMRI detected responses display significant statistical effects. Thus it was shown that a framework for network-based statistical testing is more statistically powerful, than a mass-univariate approach.

It is expected that in the near future there will be a plethora of databases of network-based objects in neuroscience motivating the development and extension of various tools from classical statistics to global network data.

In the [70] paper discussion the relationship between neuroimaging and Big Data areas it is analyzed how modern neuroimaging research represents a multifactorial and broad ranging data challenge, involving the growing size of the data being acquired; sociological and logistical sharing issues; infrastructural challenges for multi-site, multi-datatype archiving; and the means by which to explore and mine these data. As neuroimaging advances further, e.g. aging, genetics, and age-related disease, new vision is needed to manage and process this information while marshalling of these resources into novel results. It is predicted that on this way "big data" can become "big" brain science.

### 5.3 Climate in Australia

Another view on hypothesis representation and evaluation is presented in [41]. Authors argue, that as long as in DIS data relevant to some hypotheses gets continuously aggregated as time passes, hypotheses should be represented as programs that are executed repeatedly, as new relevant amounts of data gets aggregated. Their method and techniques are illustrated by examining hypotheses about temperature trends in Australia during the 20th century. The hypothesis being tested comes from [42], stated that the temperature series is not stationary and is integrated of order 1 (I(1)). Non-stationarity means that the level of the time series is not stable in time and can show increasing and decreasing trends. I(1) means that by differentiating the stochastic process a stationary process (main statistical properties of the series remain unchanged) is obtained. Phillips-Perron test and the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test are used and both of them are executed in R. Several data sources are crawled: 1) The National Oceanographic and Atmospheric Administration marine and weather information, 2) Australian Bureau of Meteorology dataset. The framework consists of R interpreter and R *SPARQL*, *tseries* packages. Authors also used agINFRA for

computation and rich semantics to support traditional scientific workflows for natural sciences. Authors received further evidence on different independent dataset that time series is integrated of order 1.

### 5.4 Financial market

Efficient-market hypothesis (EMH) is one of the most prominent in finance and *"asserts that financial markets are "informationally efficient""*. In [8] authors test the weak form of EMH, stating that prices on traded assets (e.g., stocks, bonds, or property) already reflect all past publicly available information. The null hypothesis states that successive prices changes are independent (random walk). The alternative hypothesis states that they are dependent. To check if the successive closing prices are dependent of each other the following statistical tests were used: a serial correlation test, a runs test, an augmented Dickey-Fuller test and the multiple variance ratio test. Tests were performed on daily closing prices from the six European stock markets (France, Germany and UK, Greece, Portugal and Spain) during the period between 1993 and 2007. The result of each test states whether successive closing prices are dependent of each other.

Test provides evidence that for monthly prices and returns the null hypothesis should not be rejected for all six markets. If daily prices are concerned the null hypothesis is not rejected for France, Germany, UK and Spain, but this hypothesis is rejected for Greece and Portugal. However, on the 2003-2007 dataset the null hypothesis for these two countries is not rejected as well.

In [8] Bollen et al. use different approach to test EMH. Authors investigate whether public sentiment, as expressed in large-scale collections of daily Twitter posts, can be used to predict the stock market. They build public mood time series by sentiment analysis of tweets from February 28, 2008 to December 19, 2008 and try to show that it can predict Dow Jones Index corresponding values. The null hypothesis states that the mood time series do not predict DJIA values. Granger causality analysis in which Dow Jones values and mood time series are correlated is used to test the null hypothesis. Granger causality analysis is used to determine if one time series can predict another time-series. Its results reject the null hypothesis and claim that public opinion is predictive of changes in DJIA closing values.

## 6 Conclusion

The objective of this study is to analyze, collect and systematize information on the role of hypotheses in the data intensive research process as well as on support of hypothesis formation, evaluation, selection and refinement in course of the natural phenomena modeling and scientific experiments. The discussion is started with the basic concepts defining the role of hypotheses in the formation of scientific knowledge and organization of the scientific experiments. Based on such concepts, the basic approaches for hypothesis formulation applying logical reasoning, various methods for hypothesis modeling and testing (including classical statistics, Bayesian hypothesis and parameter estimation methods, hypothetico-deductive approaches) are briefly introduced. Special attention is given to discussion of the data mining and machine learning methods role in process of generation, selection and evaluation of hypotheses as well as the methods for motivation of new hypothesis formulation. Facilties of informatics for support of hypothesis-driven experiments, considered in the paper, are aimed at the conceptualization of scientific experiments, hypothesis formulation and browsing in various domains (including biology, biomedical investigations, neuromedicine, astronomy), automatic organization of hypothesis-driven experiments. Examples of scientific researches applying hypotheses considered in the paper include modeling of population and structure synthesis of the Galaxy, connectome-related hypothesis testing, studying of temperature trends in Australia, analysis of stock markets applying the EMN (Efficient market hypothesis), as well as algorithmic generation of hypotheses in the IBM Watson project applying the NLP and knowledge representation and reasoning technologies. An introduction into the state of the art of the hypothesis-driven research presented in the paper opens a way for investigation of the generalized approaches for efficient organization of hypothesis-driven experiments applicable for various branches of DIS.

## References


[1] Agresti, A., Finlay, B. Statistical Methods for the Social Sciences (4th Edition), 2008. – P. 624.

[2] Alferes, J. J., Pereira, L. M., Swift, T. Abduction in well-founded semantics and generalized stable models via tabled dual programs. In: TPLP, 2004. – Vol. 4, No. 4. – P. 383–428.

[3] Asgharbeygi, N., Langley, P., Bay, S., Arrigo, K. Inductive revision of quantitative process models. In: Ecological modelling – Vol. 194, No. 1. – P. 70–79.

[4] Bacon, F. The new organon. In: R. M. Hutchins, (ed.), Great books of the western world. The works of Francis Bacon. Chicago, Encyclopedia Britannica, Inc., 1952 – Vol. 30. – P. 107–195.

[5] Barber, D. Bayesian Reasoning and Machine Learning. Cambridge University Press, 2010. – P. 720.

[6] Bartha, P. Analogy and Analogical Reasoning. In: The Stanford Encyclopedia of Philosophy, 2013. – http://plato.stanford.edu/archives/ fall2013/entries/reasoning-analogy/

[7] Biswal, B.B., Mennes, M., Zuo, X.N., Gohel, S., Kelly, C., Smith, S.M., Windischberger, C. Toward discovery science of human brain function. In: Proceedings of the

National Academy of Sciences, 2010. – V. 107, No. 10. – P. 4734–4739.

[8] Bollen, J., Mao, H., Zeng, X. Twitter mood predicts the stock market. In: Journal of Computational Science, 2011. – V. 2, No. 1. – P. 1–8.

[9] Borges, M. R. Efficient market hypothesis in European stock markets. In: The European Journal of Finance, 2010. – V. 16, No. 7. – P. 711–726.

[10] Breiman, L. Statistical Modeling: The Two Cultures. In: Statistical Science, 2001. – V. 16, No. 3. – P. 199–231.

[11] Brun, A. Manifold learning and representations for image analysis and visualization. Department of Biomedical Engineering, Linköpings universitet, 2006.

[12] Callahan, A., Duumontier, M., Shah, N. HyQue: Evaluating hypotheses using Semantic Web technologies. In: J. Biomedical Semantics, 2011. – V. 2, No. S-2. – P. S3.

[13] Castrillo, J.I. , S.G. Oliver (eds.). Yeast Systems Biology: Methods and Protocols. In: Methods in Molecular Biology, Springer, 2011. – V. 759. – P. 535.

[14] Citrigno, S., Eiter, T., Faber, W., Gottlob, G., Koch, C., Leone, N., Scarcello, F. The dlv system: Model generator and application frontends. In: Proceedings of the 12th Workshop on Logic Programming, 1997. – P. 128–137.

[15] Craddock, R.C., Jbabdi, S., Yan, C.G., Vogelstein, J.T., Castellanos, F.X., Di Martino, A., Milham, M.P. Imaging human connectomes at the macroscale. In: Nature methods, 2013. – V. 10, No. 6. – P. 524–539.

[16] Czekaj, M.A., Robin, A.C., Figueras, F., Luri, X., Haywood, M. The Besançon Galaxy model renewed I. Constraints on the local star formation history from Tycho data. In: arXiv preprint arXiv:1402.3257, 2014.

[17] Dredze, M, Crammer, K., Pereira, F. Confidence-Weighted Linear Classification. In: Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 2008. – P. 264-271.

[18] Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Welty, C. Building Watson: An overview of the DeepQA project. In: AI magazine, 2010. –V. 31, No. 3. – P. 59–79.

[19] Field, A. Discovering statistics using IBM SPSS statistics. In: Sage, 2013. – P. 915.

[20] Gao, Y., Kinoshita, J., Wu, E., Miller, E., Lee, R., Seaborne, A., Clark, T. SWAN: A distributed knowledge infrastructure for Alzheimer disease research. In: Web Semantics: Science, Services and Agents on the World Wide Web, 2006. - V. 4, No. 3. – P. 222–228.

[21] Ginestet, C.E., Balanchandran, P., Rosenberg, S., Kolaczyk, E.D. Hypothesis Testing For Network Data in Functional Neuroimaging. In: arXiv preprint arXiv:1407.5525, 2014.

[22] Ginestet, C. E., Fournel, A. P., Simmons, A. Statistical network analysis for functional MRI: summary networks and group comparisons. In: Frontiers in computational neuroscience, 2014. – Vol. 8.

[23] Gonçalves, B., Porto, F. A Lattice-Theoretic Approach for Representing and Managing Hypothesis-driven Research. In: AMW, 2013.

[24] Gonçalves, B., Porto, F., Moura, A. M. C. On the semantic engineering of scientific hypotheses as linked data. In: Proceedings of the 2nd International Workshop on Linked Science, 2012.

[25] Haber, J. Research Questions, Hypotheses, and Clinical Questions. In: Evolve Resources for Nursing Research, 2010. – P. 27–55.

[26] Hastie, T., Tibshirani, R., Friedman, J., Franklin, J. The elements of statistical learning: data mining, inference and prediction. In: The Mathematical Intelligencer, 2005. – Vol. 27, No. 2. – P. 83–85.

[27] Hawthorne, J. Inductive Logic. In: The Stanford Encyclopedia of Philosophy, 2014 – http://plato.stanford.edu/archives/sum2014/entries/logic-inductive/

[28] Hempel, C. G. Fundamentals of concept formation in empirical science. In: Int. Encyclopedia Unified Science, 1952. – V. 2, No. 7.

[29] Hey, T., Tansley, S., Tolle, K. (eds.). The fourth paradigm: Data-intensive scientific discovery. Redmond, Microsoft Research, 2009. – P. 252.

[30] Huang, J., Antova, L., Koch, C., Olteanu, D. MayBMS: a probabilistic database management system. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of data, 2009. – P. 1071–1074.

[31] IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp. IBM SPSS Statistics base. IBM Corp., 2013.

[32] Ihaka, R., Gentleman, R. R: a language for data analysis and graphics. In: Journal of computational and graphical statistics, 1996. – Vol. 5, No. 3. – P. 299–314.

[33] Inoue K., Sato T., Ishihata M., Kameya Y., Nabeshima H. Evaluating abductive hypotheses using and EM algorithm on BDDs. In: Proceedings of IJCAI-09, 2009. – P. 810–815.

[34] Ivezić, Ž., Connolly, A. J., VanderPlas, J. T., Gray, A. Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data. Princeton University Press, 2014. – P. 552.

[35] Kakas, A.C., Michael, A., Mourlas, C. ACLP: Abductive constraint logic programming. In: The Journal of Logic Programming, 2000. – Vol. 44, No. 1. – P. 129–177.

[36] Kakas, A.C., Kowalski, R.A., Toni, F. Abductive Logic Programming. In: Journal of Logic and Computation, 1993. – Vol. 2, No. 6. – P. 719–770.

[37] Kerlinger, F.N., Lee, H.B. Foundations of behavioral research: Educational and psychological inquiry. New York: Holt, Rinehart and Winston, 1964. – P. 739.

[38] King, R.D., Liakata, M., Lu, C., Oliver, S.G., Soldatova, L.N. On the formalization and reuse of scientific research. In: Journal of The Royal Society Interface, 2011. – Vol. 8, No. 63. – P. 1440–1448.

[39] King, R.D., Whelan, K.E., Jones, F.M., Reiser, P.G., Bryant, C.H., Muggleton, S.H., Oliver, S.G. Functional genomic hypothesis generation and experimentation by a robot scientist. Nature, 2004. – Vol. 427, No. 6971. – P. 247–252.

[40] Lakshmana Rao, J R. Scientific 'Laws', 'Hypotheses' and 'Theories'. In: Meanings and Distinctions. Resonance, 1998. – Vol. 3. – P. 69–74.

[41] Lappalainen, J., Sicilia, M.Á., Hernández, B. Automatic Hypothesis Checking Using eScience Research Infrastructures, Ontologies, and Linked Data: A Case Study in Climate Change Research. In: Procedia Computer Science, 2013. – Vol. 18. – P. 1172–1178.

[42] Lenten, L.J., Moosa, I.A. An empirical investigation into long-term climate change in Australia. In: Environmental Modelling & Software, 2003. – Vol. 18, No. 1. – P. 59–70.

[43] Mahmoudi, A., Takerkart, S., Regragui, F., Boussaoud, D., Brovelli, A. Multivoxel Pattern Analysis for fMRI Data: A Review. In: Computational and mathematical methods in medicine, 2012.

[44] March, M.C. Advanced Statistical Methods for Astrophysical Probes of Cosmology. In: Springer Theses, 2013. – Vol. 20. – P. 177.

[45] March, M.C., Starkman, G.D., Trotta, R., Vaudrevange, P. M. Should we doubt the cosmological constant?. In: Monthly Notices of the Royal Astronomical Society, 2011. – Vol. 410, No. 4. – P. 2488–2496.

[46] Marcus, D.S., Harwell, J., Olsen, T., Hodge, M., Glasser, M.F., Prior, F., Van Essen, D.C. Informatics and data mining tools and strategies for the human connectome project. In: Frontiers in neuroinformatics, 2011. – Vol. 5.

[47] Marcus, D.S., Olsen, T.R., Ramaratnam, M., Buckner, R.L. The extensible neuroimaging archive toolkit. In: Neuroinformatics, 2007. – Vol. 5, No. 1. – P. 11–33.

[48] McComas, W.F. The principal elements of the nature of science: dispelling the myths. In: The Nature of Science in Science Education, 1998. – P. 53–70.

[49] Menzies, T. Applications of Abduction: Knowledge-Level Modeling, In: International Journal of Human-Computer Studies, 1996. – V. 45, No. 3. – P. 305–335.

[50] Nickles, T. (ed.). Scientific discovery: Case studies. Taylor & Francis, 1980. – Vol. 2. – P. 501.

[51] Plotkin, G.D. A note on inductive generalization. In: Machine Intelligence. Edinburgh University Press, 1970. – Vol. 5. – P. 153–163.

[52] Poincaré, Henri. The Foundations of Science: Science and Hypothesis, The Value of Science, Science and Method. The Project Gutenberg EBook, 2012. – Vol. 39713. – P 554.

[53] Popper, K.. The Logic of Scientific Discovery (Taylor & Francis e-Library ed.). London and New York: Routledge / Taylor & Francis e-Library, 2005.

[54] Porto, F. Big Data in Astronomy. The LIneA-DEXL case. Presentation at the EMC Summer School on BIG DATA – NCE/UFRJ, 2013.

[55] Porto, F. Moura, A. M. C., Gonçalves, B., Costa, R., Spaccapietra, S. A Scientific Hypothesis Conceptual Model. In: ER Workshops, 2012. – Vol. 7518. – P. 101–110.

[56] Porto, F., Moura, A. M. C. Scientific Hypothesis Database. Report, 2011.

[57] Porto, F., Spaccapietra, S. Data model for scientific models and hypotheses. In: The evolution of conceptual modeling, 2011. – Vol. 6520. – P. 285–305.

[58] Racunas, S.A., Shah, N.H., Albert, I., Fedoroff, N.V. Hybrow: a prototype system for computer-aided hypothesis evaluation. In: Bioinformatics, 2004. – Vol. 20, No. 1. – P. 257–264.

[59] Ray, O., Kakas, A. ProLogICA: a practical system for Abductive Logic Programming. In: Proceedings of the 11th International Workshop on Non-monotonic Reasoning, 2006. – P. 304–312.

[60] Robin, A.C., Reylé, C., Derrière, S., Picaud, S. A synthetic view on structure and evolution of the Milky Way. arXiv preprint astro-ph/0401052, 2004.

[61] Robin, A., Crézé, M. Stellar populations in the Milky Way-A synthetic model. In: Astronomy and Astrophysics, 1986. – Vol. 157. – P. 71–90.

[62] Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D., Iverson, G. Bayesian t tests for accepting and rejecting the null hypothesis. In: Psychonomic bulletin & review, 2009. – Vol. 16, No. 2. – P. 225–237.

[63] Schickore, J. Scientific Discovery. The Stanford Encyclopedia of Philosophy, 2014 – http://plato.stanford.edu/archives/spr2014/entries/scientific-discovery/

[64] Sivia, D.S., Skilling, J. Data Analysis. A Bayesian Tutorial. Oxford University Press Inc., New York, 2006. – P. 264.

[65] Soldatova, L.N., Rzhetsky, A., King, R. D. Representation of research hypotheses. In: J. Biomedical Semantics, 2011. – Vol. 2, No. S-2. – P. S9.

[66] Sparkes, A., Aubrey, W., Byrne, E., Clare, A., Khan, M. N., Liakata, M., King, R. D. Towards Robot Scientists for autonomous scientific discovery. In: Autom Exp, 2010. – Vol. 2, No 1.

[67] Starkman, G.D., Trotta, R., Vaudrevange, P.M. Introducing doubt in Bayesian model comparison. arXiv preprint arXiv:0811.2415, 2008.

[68] Tamaddoni-Nezhad, A., Chaleil, R., Kakas, A., Muggleton, S.H. Application of abductive ILP to learning metabolic network inhibition from temporal data. In: Machine Learning, 2006. – Vol. 64. – P. 209–230.

[69] Tran, N., Baral, C., Nagaraj, V.J., Joshi, L. Knowledge-based integrative framework for hypothesis formation in biochemical networks. In: Data Integration in the Life Sciences, 2005. – P. 121–136.

[70] Van Horn, J.D., Toga, A.W. Human neuroimaging as a “Big Data” science. In: Brain imaging and behavior, 2014. – Vol. 8, No. 2. – P. 323–331.

[71] Van Nuffelen, B., Kakas, A. A-system: Declarative programming with abduction. In: Logic Programming and Nonmotonic Reasoning, 2001. – P. 393–397.

[72] Weber, M. Experiment in Biology. The Stanford Encyclopedia of Philosophy, 2014. – http://plato.stanford.edu/archives/fall2014/entries/biology-experiment/

[73] Woodward, J. Scientific Explanation. The Stanford Encyclopedia of Philosophy, 2011. – http://plato.stanford.edu/archives/win2011/entries/scientific-explanation/

[74] Yan, C.G., Craddock, R.C., Zuo, X.N., Zang, Y.F., Milham, M.P. Standardizing the intrinsic brain: towards robust measurement of inter-individual variation in 1000 functional connectomes. In: Neuroimage, 2013. – Vol. 80. – P. 246–262.

# The Besancon Galaxy Model, a Population Synthesis Tool for Galactic Structure and Evolution Studies

© Annie C. Robin
Institut UTINA

© Céline Reylé
OSU THETA
Besançon, France

© Bernard Debray
Université de Franche-Comté

annie.robin@obs-besancon.fr

## Abstract

Understanding the Milky Way structure and evolution is a major objective in astrophysics. Our Galaxy is an object of highest interest for learning about galaxy formation and evolution in general. Many data of different kinds (photometry, astrometry, spectroscopy) are available for millions of stars in the Milky Way, which interpretation in terms of evolution is not easy. Stars in the Milky Way have been recognized to be part of several populations with typical characteristics and spatial distributions, such as the disc, the halo and the bulge, which shape the overall Galaxy and classify it as a barred spiral galaxy.

We present here an approach for understanding the Galaxy by simulating stars in these various populations. The population synthesis scheme is used to simulate a scenario of formation of the Galaxy. It allows to confront such scenario with real data. It is used to prepare new observations (define the best protocole of observations to answer a given question) and to interpret observations in terms of Galaxy structure and evolution. The main hypotheses to build the model are based on up-to-date knowledge on stellar evolution models, grids of atmosphere models, and galactic dynamics. Many parameters are needed, but the availability of many sources of data allow to constraint step by step theses parameters. The model produces simulations at a variety of wavelengths from X rays to mid infrared. It has been confronted to large scale surveys, ground based (Sloan Digital Sky Survey [1], 2MASS near-infrared sky survey [2], among others), and space based surveys (GALEX UV telescope, Hubble Space Telescope, etc.). It has been successfully used to constrain galactic structure parameters [3, 4], such as the disc scale length, scale heights, bar structure, halo shape, and evolution parameters such as the star formation history [5] in the solar neighborhood. It is also a useful tool to prepare future surveys, in particular the Gaia mission launched by the European Space Agency in December 2013, or further projects like PLATO and EUCLID.

The model simulator is available through a dedicated web interface, which allows the users to run the model and prepare their own simulations for direct comparisons with real data. The simulations are computed on the cluster of the Institut UTINAM and deposited on the ftp server. VOTable and ascii format are available for these simulations. Further developments are envisaged for the web service, such as a data base of simulations, and a service for bayesian stellar classification. We shall present various applications of the simulations in the different astrophysical domains.

## References

[1] Aihara, H., Allende Prieto, C., An, D., et al. 2011. The Eighth Data Release of the Sloan Digital Sky Survey: First Data from SDSS-III. Astrophys. J. Suppl, 193, 29.

[2] Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006. The Two Micron All Sky Survey (2MASS). Astronomical J., 131, 1163.

[3] Robin, A. C., Reyle, C., Derriere, S., & Picaud, S. 2003. A synthetic view on structure and evolution of the Milky Way. Astron. Astrophys., 409, 523.

[4] Robin, A. C., Marshall, D. J., Schultheis, M., & Reyle, C. 2012. Stellar populations in the Milky Way bulge region: towards solving the Galactic bulge and bar shapes using 2MASS data. Astron. Astrophys., 538, A106.

[5] Czekaj, M. A., Robin, A. C., Figueras, F., Luri, X., & Haywood, M. 2014. The Besancon Galaxy model renewed. I. Constraints on the local star formation history from Tycho data. Astron. Astrophys., 564, A102.

**Proceedings of the 16th All-Russian Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" — RCDL-2014, Dubna, Russia, October 13–16, 2014.**

# Проблемы обозначения и кросс-идентификации кратных объектов в астрономии

© Д.А. Ковалева © П.В. Кайгородов © О.Ю. Малков
Институт астрономии РАН, Москва
dana@inasan.ru pasha@inasan.ru malkov@inasan.ru
© Л.А.Калиниченко © Н.А.Скворцов
Институт проблем информатики РАН, Москва
leonidandk@gmail.com nskv@mail.ru

**Аннотация**

Работа продолжает и развивает направление исследований, ориентированных на решение задач всестороннего анализа массивов разнородных данных имеющимся арсеналом научных методов и инструментов для выявления полезной информации и получения новых знаний. Здесь рассмотрены проблемы обозначения и кросс-идентификации кратных объектов в астрономии, с акцентом на упорядочение обозначений и интеграцию многомерных наблюдательных данных различных типов. В работе кратко обозреваются существующие методики обозначения одиночных и кратных астрономических объектов (с акцентом на применяемую в Базе данных двойных звезд BDB схему обозначений BSDB), а также принципы кросс-идентификации.

## 1 Идентификация небесных объектов

Астрономические объекты стали каталогизироваться еще во II в. н. э. Астрономические каталоги, созданные до начала XVII в., насчитывали до полутора тысяч объектов, а затем, с изобретением телескопа, число объектов стало стремительно расти. Каталоги 70-х годов прошлого века, когда стали образовываться центры астрономических данных и создаваться астрономические базы данных, насчитывали до 2 млн объектов, а современные каталоги включают млрд объектов, и в ближайшем будущем это число увеличится на 1-2 порядка.

Необходимо заметить, что постоянно растет не только число объектов, подлежащих каталогизации, но и количество каталогизируемых параметров. Если первые каталоги включали 3-4 параметра на объект (две координаты, визуальная оценка блеска и грубый классификатор уровня «звезда–туманность»), то современные каталоги характеризуют объект десятками и сотнями параметров, не считая различных функций (распределение энергии в спектре, функция изменения блеска со временем и пр.).

При наименовании небесных объектов (необходимом для каталогизации) астрономы не следовали какому-то одному правилу, точнее, применяли различные подходы. Хронологически первыми нужно считать методы, опирающиеся на систему созвездий – поименованных участков небесной сферы, содержащих группы звезд. Были введены в обращение, в частности, система Байера, именующая звезды в созвездии греческими буквами в порядке убывания яркости; система Флемстида, просто нумерующая звезды в созвездии с увеличением их прямого восхождения (т.е., для северного полушария Земли, слева направо); система Аргеландера-Хартвига (для переменных звезд, в порядке их открытия), использующая сложную комбинацию двухбуквенных обозначений и, в случае их нехватки, порядковых номеров. Нехватка букв в греческом алфавите (особенно для протяженных созвездий) стала очевидной довольно быстро, а система Флемстида теряла свою стройность, если в созвездии обнаруживали новую, более слабую звезду, расположенную на небесной сфере среди уже перенумерованных.

С введением систем небесных координат исследователи получили возможность создавать имена (идентификаторы), базирующиеся на значениях координат. Идентификаторы компилировались из округленных (или, иногда, сокращенных) значений координат, при этом для объектов Галактики применялась экваториальная, а для внегалактических объектов – галактическая система координат. Такие системы уже не были связаны с границами созвездий, но тоже оказались не совсем устойчивыми: из-за прецессии и нутации земной оси, а также из-за собственных движений

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2013, Дубна, Россия, 13–16 октября 2014 г.**

координаты небесных тел изменяются (у некоторых – весьма значительно). Таким образом, координатно-ориентированные имена звезд оказывались, вообще говоря, разными в разных системах, если введение этих систем в строй отстояло друг от друга на более чем 10 лет (это число зависит, конечно, от позиционной точности каталога).

Кроме того, каталоги, базирующиеся на таких системах и традиционно отсортированные по значению прямого восхождения (аналог земной долготы), оказались весьма неудобны для планирования наблюдений, так как в них могли соседствовать объекты с очень разными значениями склонений (аналог земной широты). С увеличением числа каталогизируемых небесных объектов этот недостаток становился все более досадным, и авторы каталогов, содержащих более 100 тыс. объектов в конце концов перешли на зонную систему идентификации. Небесная сфера «нарезалась» на зоны (пояса) по склонению (а позже, для каталогов с числом объектов, превышающим 10 млн, и эти пояса пришлось «порезать» на более мелкие зоны), и объекты в каталоге сортировались по зонам, а внутри зоны – по прямому восхождению.

Наконец, в «тематических» каталогах, содержащих объекты (или параметры объектов) определенного типа, принято просто нумеровать их в порядке открытия. Помимо упомянутых выше переменных звезд можно привести в пример каталог спектроскопических двойных систем и (с некоторыми оговорками) ставший в последнее время популярным, в связи с открытием экзопланет, каталог близких звезд Глизе (V/70A, указана нумерация каталога в Базе данных каталогов VizieR [1]).

Таким образом, небесные объекты оказались поименованы и включены в различные каталоги в соответствии с их:

- положением в созвездии,
- координатами,
- координатами в зонах небесной сферы,
- блеском (разным в различных фотометрических системах),
- очередностью открытия
- или вообще без всякой системы, как, например, объекты в широко известном каталоге туманностей Мессье [2].

Нужно отметить, что практически все эти (даже самые архаичные) системы наименований сохранились, успешно применяются по сей день и продолжают создавать проблемы для успешной кросс-идентификации небесных объектов.

Так, самые яркие (и следовательно, наиболее хорошо изученные и включенные во многие каталоги) звезды имеют 4-5 десятков общеупотребительных наименований.

Данная работа посвящена идентификации звезд, однако практически все вышесказанное справедливо для туманностей и скоплений нашей Галактики и для более удаленных объектов (галактик и квазаров). В статье рассматриваются, с одной стороны, схемы идентификации одиночных и кратных звёзд как системы их обозначения в каталогах, а с другой стороны, – методы кросс-идентификации звёзд, использующие как их идентификаторы, так и наблюдательные и астрофизические параметры для отождествления звёзд в разных каталогах.

## 2 Кросс-идентификация небесных объектов

Разнообразие методов создания астрономических каталогов поставило астрономов перед задачей выработки схем кросс-идентификации содержащихся в них объектов. Действительно, точная позиционная информация, содержащаяся в каталоге А, будучи проанализированной совместно с высокоточной фотометрией из каталога Б и данными радионаблюдений из каталога В, позволяет получить более полную картину образования, строения и эволюции Галактики, чем данные из каждого из этих каталогов по отдельности. При этом три упомянутых каталога создавались разными коллективами по разным методикам, вообще говоря, в разное время и использовали различные системы идентификации. Базы астрономических данных, объединяющие неоднородную информацию, требуют в первую очередь решения проблемы кросс-идентификации объектов.

Проблема кросс-идентификации (КИ) астрономических объектов состоит в отождествлении одних и тех же объектов среди неоднородных данных из разных каталогов. Она является частным случаем рассматриваемого в информатике направления разрешения сущностей (entity resolution). Комплексный подход к разрешению сущностей обычно представляется в виде определённой последовательности действий, включающей [3, 4]:

- связывание элементов схем данных разных источников, соответствующих по смыслу, предварительную очистку и приведение к единообразному представлению неоднородных данных из связанных атрибутов;
- индексирование данных с целью уменьшения попарного перебора сравниваемых кортежей из разных источников;
- применение методов сравнения данных различных типов, включая строки, числовые типы, даты, пространственные координаты, множества, записи в целом;
- выделение набора данных, однозначно определяющих уникальные объекты, а при невозможности его выделения оценка близости

данных из разных источников по определённым критериям и принятие решения об отождествлении описанных ими объектов.

Естественной основой разрешения объектов применительно к решению задачи кросс-идентификации небесных объектов являются их пространственные координаты. Данные атрибутов, содержащие координаты объектов, приводятся к единообразному представлению с учётом используемых форматов координат и разности эпох наблюдения. Близость координат с определённым допущением решает проблему уменьшения перебора попарного сравнения кортежей в каталогах. Координатное совмещение данных различных каталогов позволяет решать до 80% проблем, связанных с КИ.

Однако чисто координатного подхода оказывается недостаточно, если речь идет о плотных звездных полях (скоплениях или кратных системах, см. ниже), о быстро движущихся и/или переменных объектах, о данных с различающимся угловым разрешением, а соответственно, различной точностью координат и т.п. В таких случаях почти всегда для КИ удается использовать атрибуты каталогов, содержащие фотометрическую информацию. При приведении фотометрических данных к единообразному представлению приходится принимать во внимание тот факт, что блески (или цвета, т.е., разницы блесков) объектов в различных фотометрических системах, вообще говоря, различны и, хотя и подчиняются неким корреляционным соотношениям, соотношения эти опять-таки различны для объектов различной природы (которая, как правило, при КИ остается неизвестной). Более детальное описание общих принципов кросс-идентификации объектов в астрономии можно найти в [5].

Полезным дополнительным параметром является также классификатор объекта, однако он присутствует в каталогах далеко не всегда и представляет, как правило, достаточно грубую оценку природы объекта (точечный/протяженный). Естественно, используется и вся другая информация (например, спектральный тип объектов), если она присутствует в обоих кросс-идентифицируемых каталогах.

Таким образом, критерии отождествления и оценки близости объектов при кросс-идентификации разрабатываются с учётом позиционных и фотометрических параметров, рассчитанных астрофизических величин, параметров фотометрических систем и углового разрешения оборудования, используемого для обзоров.

Одной их особенностей данных астрономических каталогов является то, что некоторые из них уже включают данные об именах описываемых объектов в соответствии с определёнными системами идентификации, принятыми в других каталогах. Приведённые в каталогах имена идентифицируют соответствующие объекты в других каталогах, и таким образом, являются результатом уже решённой при их составлении задачи отождествления между конкретными парами каталогов.

Однако кросс-идентификация для разных пар каталогов может производиться различными методами, различные виды идентификаторов имеют разный смысл. Использование разных критериев отождествления объектов, различных систем идентификации и способов связывания имён в каталогах может само по себе рождать конфликты идентификации. В том числе, точность используемых методов отождествления невозможно выяснить из самих идентификаторов. Поэтому для проверки корректности существующих связей идентификаторов объектов и для разрешения конфликтов между идентификаторами при отождествлении объектов более, чем в двух каталогах, приходится заново прибегать к решению задачи КИ на основе наблюдательных и астрофизических параметров.

С учетом изложенных выше соображений астрономам удалось решить (и удается решать, с появлением новых каталогов и обзоров) большинство проблем КИ, что находит свой результат в публикуемых таблицах КИ и создаваемых базах астрономических данных различных типов. Приведённые выше критерии, однако, не во всех случаях достаточны для решения задачи кросс-идентификации. Нередко возникает необходимость в более специфических методах. Для некоторых типов объектов задача кросс-идентификации далека от окончательного разрешения, и, в первую очередь, это относится к двойным и кратным звездам.

## 3 Особенности идентификации и кросс-идентификации кратных звезд

Двойные и кратные звезды весьма многочисленны и не исключено, что их доля среди звезд Галактики (если включать в их число и планетные системы) весьма близка к 100 процентам. Столь высокая кратность объясняется особенностями звездообразования, в частности, необходимостью для вращающегося и сжимающегося протозвездного газо-пылевого облака избавиться от осевого момента инерции, что проще всего осуществить за счет фрагментации на компоненты и/или образования планетной системы.

Далеко не все двойные звезды наблюдаются именно как двойные. Для этого паре нужно либо находиться достаточно близко к наблюдателю и быть достаточно широкой (тогда компоненты будут наблюдаться по отдельности), либо демонстрировать доплеровское смещение линий в спектре и/или переменность блеска из-за орбитального движения компонентов, либо проявлять себя как источник рентгеновского излучения (возникающего из-за аккреции вещества на один из компонентов) и т.п. Эти и другие типы

двойных регистрируются с помощью различных методик различными коллективами и им, естественно, присваиваются обозначения в рамках различных схем идентификации. Таким образом, обозначение одиночного объекта в некотором каталоге должно быть приписано двойной системе в другом каталоге, когда эти каталоги имеют разное пространственное разрешение (т.е., речь идет о кросс-идентификации объектов различных категорий). Задача еще сильнее усложняется для объектов большей кратности. В результате к «традиционным» проблемам схем идентификации прибавляется несколько новых, характерных именно для двойных (кратных) систем.

Прежде всего, требуется разработать методику обозначений для идентификации компонентов кратной системы. Для двойных звезд эту проблему решали традиционно, добавляя к идентификатору системы в качестве суффиксов буквы *A* и *B*. Но уже с тройными системами поступали по-разному. В тесных системах, когда оказывалось, что компонент *A* представляет собой на самом деле двойную звезду, новые два компонента получали обозначения *Aa* и *Ab*. Этот принцип, помимо прочего, отражал и тот факт, что кратные системы (кроме самых широких) должны быть иерархическими, иначе они будут динамически нестабильными и просуществуют недолго. Исследователи же широких систем, где компоненты, как правило, наблюдаются по отдельности, а уровни иерархии не очевидны, обозначали вновь открытый компонент буквой *C*. Аналогично эти принципы распространялись на системы более высокой кратности.

Эти схемы, естественно, не идеальны. Помимо того, что появляются трудно форматируемые обозначения типа *Aa*1, а в неиерархических системах особенно высокой кратности (которые некоторые исследователи, впрочем, предпочитают называть скорее скоплениями) не хватает букв латинского алфавита. Открытие компонента на промежуточном иерархическом уровне является более редким событием, но также нарушает описанные выше принципы наименования объектов.

Еще одной, дополнительной трудностью, присущей даже двойным системам является порядок присвоения букв *A* (главному) и *B* (вторичному) компонентам, точнее, неоднозначность ответа на вопрос, какой компонент в паре является главным. Для исследователей визуально-двойных звезд это – более яркий компонент (оставим в стороне вопрос о порядке присвоения букв в парах с компонентами одинаковой яркости, а также то обстоятельство, что в разных фильтрах относительная яркость компонентов может быть разной, а в некоторых случаях самым ярким агентом в системе является даже не звезда, а аккреционный диск вокруг одной из звезд), для исследователей переменных звезд – более горячий. При моделировании тесных двойных систем принято считать главным компонентом более массивную на сегодняшний день звезду, а с точки зрения звездной эволюции главный компонент – изначально более массивная звезда (из-за переноса массы в системе в процессе эволюции это могут быть разные компоненты). С точки зрения кинематики двойной системы главный компонент – меньший по массе. И существуют, наконец, задачи, для которых удобно считать главным компонентом больший по размерам. Присвоение букв (*B, C,* …) компонентам в системах большей кратности также может осуществляться по-разному: в порядке уменьшения блеска, в порядке удаления от главного компонента *A* и т.д.

Все эти обстоятельства приводят к тому, что компоненты (и сами системы) двойных и кратных звезд получают в различных каталогах весьма различные обозначения (присваиваемые в соответствии с различными схемами идентификации), и задача КИ, более-менее решенная для одиночных объектов, становится гораздо более сложной для двойных и кратных систем.

## 4 Поиск информации в Базе данных двойных звезд

В данном разделе описаны принципы поиска и отображения информации в Базе данных двойных звезд (Binary star DataBase, BDB, [6]), которая разрабатывается в настоящий момент в Институте Астрономии РАН и содержит данные о порядка 110 000 звездных систем с кратностью от 2 и выше. Данные, содержащиеся в BDB, получаются путем объединения множества каталогов (с разными принципами организации и разными системами идентификации), содержащих информацию о двойных и кратных звездах разных наблюдательных типов: визуальных двойных, спектральных, рентгеновских, астрометрических, интерферометрических, спектроскопических, фотометрических и т.д. Из этих каталогов извлекаются как наблюдательные данные о координатах, собственных движениях, периодах, переменности, звездных величинах, так и астрофизические параметры – эволюционный статус, массы и т.п.

Запрос данных в BDB возможен либо с использованием идентификатора, либо при помощи поиска по параметрам. При поиске по идентификатору пользователь может выбрать систему идентификации из нескольких, включенных в BDB: ADS, Bayer, CCDM, DM, Flamsteed, GCVS, HD, HIP, IDS, IGR, Name, SBC9 или WDS; либо ввести идентификатор в свободной форме. В последнем случае BDB прежде всего попытается найти введенный идентификатор среди имеющихся в базе. Если найти идентификатор не удалось, будет сделан запрос (с использованием протокола SOAP) к системе, связывающей обозначения (name resolver) Sesame [7], и среди выданных ею результатов будет выбран идентификатор, имеющийся в BDB.

Здесь необходимо отметить, что иногда, когда таких идентификаторов несколько, и не все они принадлежат одному и тому же объекту BDB, мы получаем указание либо на наличие кратной системы, состоящей из двух других систем, либо на ошибки, содержащиеся в Sesame (исключая, конечно, тривиальный случай, когда эти идентификаторы принадлежат двум разным компонентам одной системы).

Результаты запросов к Sesame кэшируются. После получения идентификатора:

1. BDB находит все записи (системы, пары или компоненты), ссылающиеся на данный идентификатор. Найденные записи включаются в список (здесь и далее в список включаются только ранее отсутствовавшие в нем элементы).

2. Для каждой найденной записи находятся (и также включаются в список) записи с тем же внутренним идентификатором, что и у нее (каждому объекту в BDB может соответствовать несколько записей, поступивших из разных каталогов, но имеющих один внутренний идентификатор). В качестве внутреннего идентификатора используется имя BSDB (см. ниже).

3. Если найденная запись относится к системам, то в список включаются все пары, ссылающиеся на нее. Для пары включаются в список оба ее компонента, а также система, на которую она ссылается. Для компонента в список включается его родительская пара.

4. Составляется (пополняется) список внешних идентификаторов, на которые ссылаются записи, найденные на предыдущих этапах.

Перечисленные этапы повторяются циклически для всех найденных идентификаторов до тех пор, пока формируемый список не перестанет расти.

## 5 Система идентификации BSDB

При создании Базы данных двойных звезд BDB авторами разработана схема обозначений BSDB, призванная разрешить существующие проблемы идентификации и кросс-идентификации двойных систем. BSDB должна была удовлетворять (и удовлетворяет) следующим критериям:

- ни один объект не должен носить более одного идентификатора;
- ни один идентификатор не должен быть присвоен более, чем одному объекту;
- открытие новых компонентов в системе не должно нарушать принципы присвоения идентификаторов;
- система должна быть несложной, близкой к традиционным и интуитивно понятной исследователям двойных.

При присвоении идентификатора по системе BSDB мы выделяем три категории объектов: система, пара, компонент. Это подход следует считать пионерским, и диктуется он тем обстоятельством, что каждая из трех категорий характеризуется своим набором наблюдательных данных. Компонент характеризуется массой, радиусом, температурой, светимостью, и т.п. (то есть, тем набором астрофизических параметров, которым характеризуется, например, одиночная звезда). Пара – это два объекта (каждый из которых, кстати, тоже может оказаться парой), связанных гравитационно. Эта категория характеризуется такими параметрами как относительное положение членов пары на небесной сфере (для визуальных двойных), орбитальными параметрами (период обращения, эксцентриситет орбиты и пр. – для орбитальных и части спектроскопических двойных), интегральным блеском (для фотометрически неразрешенных) и спектром (для спектроскопически неразрешенных). Здесь следует заметить, что наблюдатели имеют дело преимущественно именно с парами, и именно информация о парах, как правило, и включается в каталоги. Наконец, такая категория как система характеризуется общими параметрами: возраст, металличность, расстояние, кинематика в Галактике и пр. Некоторые параметры могут приписываться различным категориям: так координаты характеризуют каждый из компонентов в случае разрешенной двойной и пару – в случае неразрешенной.

Идентификатор BSDB состоит из цифровой части, компилируемой из значений небесных координат и предваряемой символом ‘*J*’ (означающим, что координаты относятся к эпохе 2000.0 года); индикатора «система–пара–компонент» (“*s*”, “*p*”, “*c*”), отделяемого двоеточием; и буквенным обозначением, в общих чертах напоминающим знакомые исследователям двойных звезд схемы обозначений. Так, обозначения объектов некой тройной системы будут выглядеть следующим образом:

*J*000144.48+590527.1:*s*

*J*000144.48+590527.1:*pAa-Ab*

*J*000144.48+590527.1:*cAa*

*J*000144.48+590527.1:*cAb*

*J*000144.48+590527.1:*pA-B*

*J*000144.48+590527.1:*cB*

Отметим, что в списке отсутствует компонент *A*, поскольку в данной кратной системе объект *A* является не звездой, а парой звезд (*Aa-Ab*) и, соответственно, описывается параметрами, характерными для пары (например, период обращения), а не для компонента (например, масса).

Координатная часть обозначения BSDB внутри системы не меняется, несмотря на то, что координаты компонентов, вообще говоря, могут различаться.

Принципы создания идентификатора BSDB удовлетворяют правилам, утвержденным Международным астрономическим союзом.

## 6 Решение проблем кросс-идентификации двойных и кратных систем

Схема обозначений BSDB должна быть всеобъемлющей, поэтому необходимо позаботиться о том, чтобы кратные системы всех наблюдательных типов могли получить в ней соответствующие обозначения. Для этого коллективом ведется работа по созданию общего каталога идентификаций двойных звезд (предварительное название – Identification List of Binaries, ILB), который должен включать обозначения BSDB для всех каталогизированных в настоящее время двойных систем, а также предоставить такую возможность и для будущих списков/каталогов/обзоров двойных.

К каталогу ILB постепенно подключаются каталоги двойных систем, начиная с самых широких и, одновременно, самых представительных. Каждому объекту, встречающемуся впервые, приписывается уникальное обозначение BSDB. Объекты, уже имеющиеся в ранее исследованных каталогах, дописываются в соответствующие строки ILB. Новые объекты, входящие в уже существующие в ILB звездные системы, приводят к корректировкам соответствующих разделов каталога. При этом приходится решать проблемы КИ, которые возникают даже в том случае, когда кратная система принадлежит только одному наблюдательному типу и, следовательно, ее составляющие поименованы хоть, возможно, и по-разному, но, по крайней мере, в соответствии с одной и той же схемой идентификации. Задача усложняется, когда в системе присутствуют объекты, проявляющие свою двойственность по-разному (т.е., принадлежащие различным наблюдательным типам, изучаемые различными группами исследователей и, в результате, имеющие весьма разные обозначения). Более того, объект, представляющийся одиночным с точки зрения одной методики наблюдений, может оказаться двойным или кратным с точки зрения другой; это является следствием разницы в позиционной и фотометрической точности используемых каталогов (методик наблюдения), сказывается на приписываемых идентификаторах и усложняет проблему КИ.

Для решения проблем КИ привлекается вся имеющаяся в каталогах информация, в первую очередь – позиционная и фотометрическая, а также уже содержащаяся в некоторых каталогах кросс-идентификация. При этом попутно решается (сама по себе весьма актуальная [8], в частности, в астрономии [9]) проблема достоверности (согласованности, непротиворечивости) каталогизируемых многомерных числовых измерительных данных, поступающих из разных источников. В частности, в процессе КИ нами обнаруживаются ошибки как в оригинальных каталогах, так и в базах данных общего назначения, о чем мы сообщаем их авторам.

Около 90 процентов всех проблем КИ удается разрешить автоматически, и это делает проблему создания унифицированных методов КИ в принципе решаемой. Для оставшихся 10 процентов все же требуется ручной подход – опять-таки, в сотрудничестве с авторами оригинальных каталогов.

Каталог ILB будет постоянно пополняться, станет основой для базы данных BDB, а также может служить для других приложений. Методика КИ компонентов, пар и систем двойных и кратных звезд также должна считаться оригинальной (как и система обозначений BSDB); она постоянно модифицируется и станет полезной для будущих астрономических обзоров.

## Заключение

Проблема обозначения небесных объектов появилась в астрономии давно и окончательного решения не нашла до сих пор. Параллельное существование и интенсивное использование десятков систем обозначений приводит, кроме всего прочего, к необходимости постоянно решать проблемы кросс-идентификации. Особенно остро эти вопросы стоят для двойных и кратных систем – объектов, выглядящих, обозначаемых и каталогизируемых различными группами исследователей по-разному.

В работе дан обзор систем и стандартов идентификации астрономических объектов, описаны сложности и особенности для кратных объектов. Обсуждаются существующие методы и средства кросс-идентификации объектов. Описаны методика обозначений BSDB и общий каталог идентификаций двойных звезд ILB, применяемые в Базе данных двойных звезд BDB, а также иллюстрируются примеры разрешения конфликтов идентификации.

## Благодарности

Мы благодарны анонимным рецензентам за ценные замечания, которые позволили улучшить текст статьи. Работа выполнена при поддержке грантов РФФИ 12-02-31904, 12-07-00528, и Программы Президиума РАН Поддержка ведущих научных школ (грант НШ-3620.2014.2)

## Литература

[1] VizieR database: http://vizier.u-strasbg.fr/viz-bin/VizieR

[2] O'Meara S. J. The Messier objects. – Cambridge University Press, 1998. – P. 3. – 304 p.

[3] Christen P. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. – Springer, 2012.

[4] Christen P. A survey of indexing techniques for scalable record linkage and deduplication // IEEE Transactions on Knowledge and Data Engineering. – 2011.

[5] Karpov S.V., Malkov O.Yu., Mironov A.V. 2012, Astrophysical Bulletin, 67, 82.

[6] Malkov O.Yu., Kaygorodov P.V., Kovaleva D.A., Oblak E., Debray E. 2014, Astronomical and Astrophysical Transactions, 28, 235.

[7] Sesame: http://cds.u-strasbg.fr/cgi-bin/Sesame

[8] Ежела В.В., *данный сборник*

[9] Авраменко А.Е. Концептуальные модели данных в отождествлении физических свойств пульсаров на вековом масштабе // Труды RCDL 2012. – CEUR, 2012. – Т. 934. – С. 245–251.

## Problems of Designation and Cross-Identification of Multiple Objects in Astronomy

Dana A. Kovaleva, Pavel V. Kaygorodov, Leonid A. Kalinichenko, Oleg Yu. Malkov, Nikolay A. Skvortsov

In this work we continue and develop research focused on solving of problems arising in the comprehensive analysis of heterogeneous data sets by available arsenal of scientific methods and tools to identify useful information and to gain new knowledge. Here the problems of designation and cross-identification of multiple objects in astronomy are discussed, with a focus on streamlining of designation schemes and integrating of multidimensional observational data of different types. We shortly review existing methods of designation of single and multiple astronomical objects, describe BSDB schemes, implemented in Binary star database (BDB) and discuss problems and solutions of cross-identification.

# Опыт идентификации персон для CRIS-систем

© А.А. Князева
Институт вычислительных технологий СО РАН, Томск
aknjazeva@ict.nsc.ru

© И.Ю. Турчановский
tur@hcei.tsc.ru

© О.С. Колобов
Институт сильноточной электроники
СО РАН, Томск
okolobov@hcei.tsc.ru

О.Л. Жижимов
Институт вычислительных технологий
СО РАН, Новосибирск
zhizhim@mail.ru

**Аннотация**

В данной работе приводится описание системы идентификации персон, которая создавалась в процессе разработки Единого репозитория результатов научно-технической деятельности (РНТД) в ИВТ СО РАН. Кратко описываются принципы и методы, используемые при создании системы, а также ее структура. Описан алгоритм создания авторитетной базы данных с описаниями персон в автоматическом режиме, без участия пользователя. Для выявления нечетких дубликатов в упоминаниях персон использовались индексирование по биграммам и расстояние редактирования.

## 1 Введение

Разработка информационных систем (ИС), предназначенных для сбора и хранения информации о результатах научной деятельности, в настоящее время крайне актуальна [1]. К таким системам относятся научные сети (например, Scopus [2], Web of Science [3], ResearchGate [4], SciVerse [5], Cross-Ref [6]), и целый класс информационных систем CRIS (Current Research Information Systems) [7]. Предметом рассмотрения в данной работе будут CRIS-системы.

С 2000 года существует организация *EuroCRIS*, объединяющая разработчиков и исследователей ИС текущих исследований в странах Европейского Союза. EuroCRIS занимается созданием и поддержкой стандартов и методологий создания CRIS-систем.

В настоящее время распространение CRIS-систем не ограничивается географическими рамками. В EuroCRIS более 300 делегатов из 43 стран. Членами данной организации являются пять российских организаций: Центральный экономико-математический институт РАН, Институт вычислительных технологий СО РАН, Институт нефтегазовой геологии и геофизики им. А.А. Трофимука СО РАН, Уральский федеральный университет и Научная библиотека «КиберЛенинка». Разработкой CRIS-систем занимаются также Астраханский государственный университет [8], Институт математики и механики им. Н.Н. Красовского УрО РАН, Институт вычислительных технологий СО РАН [9] и другие. Существуют также системы, взаимодействующие с локальными CRIS-системами и расширяющие их функциональность. В качестве примера можно привести систему, разрабатываемую в среде крупного отечественного онлайнового научно-образовательного пространства, поддерживаемого системой Соционет [10, 11].

Исходя из широкого диапазона пользователей возникает необходимость учета самой разнообразной научно-исследовательской информации, а также и большой набор требований к CRIS-системам.

В данной статье описывается система идентификации персон, которая разрабатывалась в рамках создания системы агрегирования данных по научным проектам в Институте вычислительных технологий СО РАН[1]. Задача идентификации персон в данных CRIS-систем, объединенных в единый репозиторий, близка к задачам идентификации сущностей (entity identification) [12], установления связей (record linkage) [13], выявления дубликатов (duplicate detection) [14–16].

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

[1] Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации (Государственный контракт № 14.521.11.0004 от 14.08.2013 «Разработка системы агрегирования данных по научным проектам из различных источников для обеспечения мониторинга реализации мероприятий и программ», шифр 2013-2.1-14-521-0017-002).

Перечисленные задачи актуальны для широкого диапазона ресурсов, распределенных и локальных. В самых разнообразных данных часто встречаются упоминания одних и тех же объектов реального мира, которые необходимо связывать между собой для обеспечения более качественной работы с информацией. В частности, в нашей работе используются методы нечеткого сопоставления строк и общие принципы связывания документов.

## 2 Обзор работ в области идентификации объектов

Для идентификации объектов в CRIS-системах в настоящее время используется подход, основанный на принципах LOD[2]. Он позволяет использовать собственные идентификаторы, которые при этом становятся внешними. Созданные идентификаторы могут однозначно разрешаться и сторонними системами благодаря использованию механизма URI. Такой подход разрабатывается в рамках проекта CERIF-Linked-Data[3]. В дальнейшем связать данные CRIS-систем, преобразованные в соответствии с принципами LOD, можно связать с другими источниками LOD (например, библиографическими) с помощью инструмента автоматического установления RDF-ссылок Silk[4].

Использование семантических связей при создании и отображении CERIF-документов рассматривается также в работах С.И. Паринова [17].

Вопросы интеграции информационных ресурсов, формирования наборов метаданных и онтологий для научных информационных ресурсов рассматриваются в работах А.Н. Бездушного, М.В. Кулагина, В.А. Серебрякова и др. [18, 19].

Задача создания собственной CRIS-системы, в которой производится идентификация персон, рассматривается также в работах А.С. Умарова и др. [1].

Различные системы учета публикаций, например, Scopus, Web of Science, SCIENCE INDEX (на базе РИНЦ) используют различные идентификационные коды авторов[5]. Обзор различных систем идентификаторов и их сравнительный анализ приводится в работах [20, 21].

## 3 Постановка задачи

### 3.1 Источники данных

В процессе работы использовались данные Единого репозитория результатов научно-технической деятельности (РНТД), разрабатываемого в ИВТ СО РАН. РНТД объединяет данные по научным проектам нескольких организаций: ФГБНУ «Научно-исследовательский институт – Республиканский исследовательский научно-консультационный центр экспертизы», Российский фонд фундаментальных исследований (РФФИ), Российский гуманитарный научный фонд (РГНФ), ФГБНУ «Дирекция научно-технических программ», Национальный фонд подготовки кадров, ООО «Инконсалт», ФГАНУ «Центр информационных технологий и систем органов исполнительной власти».

### 3.2 Описание задачи идентификации

Имеется несколько независимых источников, содержащие данные о научно-исследовательской деятельности. Данные из источников агрегируются в выделенную базу данных (репозиторий). Агрегированные данные могут содержать дублетное описание базовых сущностей, т.е. описывать один объект реального мира в различных вариантах. Это отражается на качестве поиска, так как результаты поиска документов, относящихся к отдельной сущности, не будут достаточно полными. Необходимо решить задачу идентификации объектов реального мира в данных.

### 3.3 Варианты идентификации объектов

Идентификацию объектов реального мира можно организовать различными способами, в зависимости от наличия авторитетных данных:

1. Существует авторитетная база данных (своя или сторонняя), в ней описываются объекты, которые необходимо идентифицировать. Задача сводится к установлению связи с авторитетным документом путем указания его идентификатора.

2. Нет такой базы, необходимо создавать ее в процессе идентификации.

Первый вариант идентификации актуален для организаций, в которых существуют развитые авторитетные базы данных (библиотек, архивов) и не всегда подходит для научно-исследовательских институтов, в которых такие базы, зачастую, не сформированы на этапе разработки CRIS-систем.

Использование сторонних баз данных для идентификации сущностей может быть особенно полезным в тех случаях, когда в самих документах приводится мало информации об объекте. Тогда можно идентифицировать объект (например, персону) не столько по его описанию, сколько по его связям с другими объектами (персонами, организациями и т.п.). С этой точки зрения могут быть полезны социальные сети, которые активно развиваются в настоящее время, например, *LinkedIn*, *Facebook*, *ВКонтакте* и др. Поскольку основное внимание в них уделяется именно связям между объектами. Можно использовать профили пользователей для их идентификации при условии, что существует API. Этот подход представляется перспективным и будет развиваться в нашей дальнейшей работе.

[2] Linked Open Data.

[3] http://code.google.com/p/cerif-linked-data/

[4] Silk Link Discovery Framework – http://www4.wiwiss.fu-berlin.de/bizer/silk/

[5] ORCID, ResearcherID, SPIN-код соответственно.

В данной статье рассматривается второй вариант идентификации, при котором создаются авторитетные базы данных. Он может быть полезен в том случае, если нет готовых авторитетных баз данных и при этом нет уверенности, что пользователи уже зарегистрированы в социальных сетях или в системах учета публикаций. Для реализации данного подхода необходимо решать следующие задачи:

1. Формальный контроль входных данных.

2. Автоматическое формирование авторитетных баз данных, которые содержат документы, описывающие идентифицируемые объекты.

3. Авторитетный контроль входных данных (установление связи).

Предлагается технология связывания документов, которую можно использовать для связывания с уже существующими системами идентификаторов, при условии, что в данных системах существуют профили идентифицируемых объектов.

При этом не идет речь о создании собственной системы универсальных идентификаторов. Для технических нужд в процессе работы используются исключительно внутренние идентификаторы документов и объектов.

### 3.4 Формат данных

Данные, используемые в работе, поступают в виде документов в формате CERIF[6]. Данный формат является официальной рекомендацией для членов Европейской комиссии (European Commission). Он определяет набор обязательных и дополнительных полей, которые должны использоваться для описания научных проектов, в том числе название проекта, краткое описание, описание участников, наименование финансирующей организации и т.п. В качестве дополнительной информации могут быть указаны ссылки на другие проекты и на публикации в рамках данного проекта [22].

## 4 Идентификация персон для РНТД

### 4.1 Входные требования к документам

К входным документам предъявляются следующие требования:

• документы должны соответствовать требованиям формата CERIF;

• в упоминании персоны должны быть как минимум указаны фамилия и первый инициал на одном из двух языков (русском или английском).

### 4.2 Краткое описание алгоритма работы

Место системы идентификации персон в РНТД представлено на рисунке 1. Система работает с данными РНТД, формирует авторитетную базу данных с описаниями персон *Persons* и сводную базу данных документов CERIF, для которых установлены связи с соответствующими персонами.

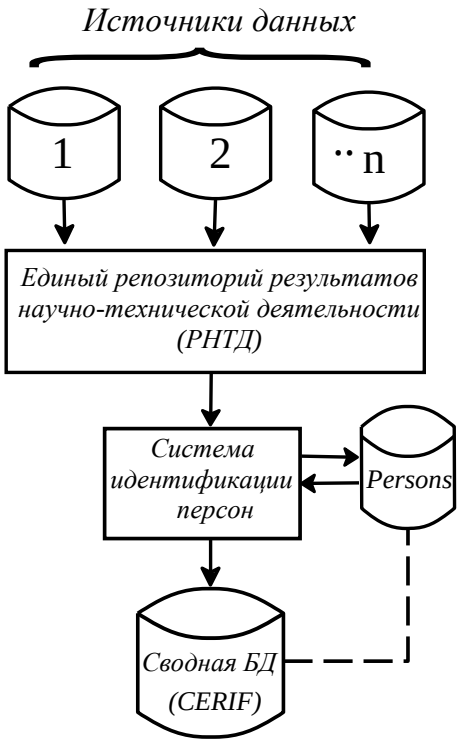



Рис. 1. Взаимодействие системы идентификации персон и РНТД

Основные принципы, которых мы придерживались при разработке модели идентификации сущностей:

• Сохранение исходных данных обеспечивает возможность возвращения входных документов к предыдущему состоянию, до того, как они были связаны. Если будет сделана ошибка при установлении связи, ее легко можно будет исправить. На практике это требование означает, что вместо того чтобы изменять документы в исходной базе данных, создается новая база данных с установленными связями.

• Использование меток содержимого, т.е. лексем, которые несут смысловую нагрузку и описывают данные. Этот принцип позволяет получать данные из системы без изменений (в их первоначальном виде) и, в то же время, при необходимости осуществлять более сложную навигацию по ним. Таким образом, мы дополняем данные и решаем проблему идентификации сущностей.

В процессе обработки входных документов создается ресурс авторитетной базы, который содержит документы с описанием персон. Авторитетные документы наполняются только той информацией, которая есть во входных данных, без привлечения сторонних источников.

В общем виде алгоритм работы программного комплекса выглядит следующим образом. Входной

[6] CERIF – формат обмена научно-исследовательскими данными, разработанный EuroCRIS

документ подвергается формальному контролю. На этом этапе необходимо проверить его корректность с точки зрения соответствия схеме данных CERIF [23]. Также возможна и проверка отдельных полей с помощью словарей допустимых значений.

Далее из входного документа извлекаем определения базовых сущностей CERIF (в данной работе рассматривается сущность типа «Персона»). Извлечение означает, что создается временный авторитетный документ, в который помещается вся информация о сущности, содержащаяся во входном документе CERIF. При этом, как правило, во входном документе упоминается более одной персоны (в проекте участвует сразу несколько исполнителей), временные документы создаются для каждой из них.

Каждый временный авторитетный документ затем сравнивается с теми, что уже находятся в авторитетной базе данных. Если такого документа (или похожего на него) в базе нет, то он загружается в нее и перестает быть «временным». Если для временного документа был найден нечеткий дубликат, то возможны два варианта:

• Во временном документе нет новой информации – авторитетная база данных остается без изменения.

• Временный документ не противоречит найденному в базе данных, но при этом содержит часть неучтенной информации – документ из базы данных дополняется этой новой информацией.

### 4.3 Поиск подобных документов в авторитетной базе данных

Сравнение временного авторитетного документа с каждым из документов, уже содержащихся в базе данных, может оказаться неоправданно трудоемким. В частности, при работе «на лету» может потребоваться сократить количество авторитетных документов, которые будут сопоставляться с временным документом. Существует множество способов ограничить круг документов для сопоставления. Приведем некоторые из них:

1. Метод стандартных блоков выделяет документы в один блок в том случае, если они содержат идентичный блочный ключ [24]. Блочные ключи формируются на основе атрибутов документов, например, первые 4 символа фамилии. Кроме того, блочный ключ может быть и составным, например, атрибут «фамилия» может сочетаться с атрибутом «год рождения». Ключи должны быть выбраны таким образом, чтобы блоки не были ни слишком большими, ни слишком мелкими.

2. Метод ближайших соседей [25] сортирует документы на основе сортирующего ключа и затем двигает окно фиксированного размера последовательно по всем документам. Документы внутри окна составляют пары друг с другом и включаются в список пар-кандидатов. Метод может некорректно работать в том случае, если количество документов с одним значением ключа превышает размер окна, поскольку в такой ситуации будут сравниваться не все нужные документы.

3. Метод Bigram-индексирования [26] предназначен для нечеткого разбиения на блоки. Основная идея заключается в том, что значения блочных ключей конвертируются в лист биграм (подстрок, состоящих из двух символов) и затем из этих биграм формируются списки на основе заданного порога (например, выбираются все документы, в которых встречается 80% биграм).

В рамках данной работы использовался метод Bigram-индексирования на основе фамилии персоны на русском языке. Использование этого метода позволяет найти документы с опечатками в фамилии, что позволяет повысить качество идентификации.

Результаты поиска выдаются в порядке релевантности, то есть в начале списка результатов помещаются документы с точным соответствием (если они есть), а затем все менее и менее похожие (в смысле совпадения по фамилии). Таким образом, ограничение круга документов задается путем установления порога, после которого документы признаются слишком отличающимися и не передаются для более подробного анализа.

В нашей работе такой порог был установлен с помощью расстояния редактирования Левенштейна [27]. Документы с различием в фамилии больше чем в 2 символа считались непохожими исключались из дальнейшего рассмотрения. Оставшиеся документы, являющиеся потенциальными дубликатами рассматриваемого документа, формируют *множество документов для сравнения*.

### 4.4 Описание процедуры сравнения документов

Рассмотрим более подробно процедуру сравнения временного авторитетного документа с документами из множества для сравнения.

Прежде всего производится вычисление строгого соответствия для всех полей документа. В зависимости от результатов сравнения возможны следующие варианты:

1. Если все поля точно равны, делается вывод, что текущий документ является точным дубликатом найденного. Из найденного документа извлекается его идентификатор и возвращается для установления связи. Временный документ не помещается в авторитетную базу данных.

2. Если точного равенства нет, то следует нечеткое сравнение (см. таблицу 1). Допускается расхождение в одном из перечисленных полей (кроме поля с указанием пола):

(а) Если расхождение в одном поле, и не превышает границы, то делается вывод о нечетком дубликате. В найденный авторитетный документ вносится информация о вариантном наименовании, возвращается код найденного документа.

Временный авторитетный документ не помещается в базу данных;

(b) Если не было обнаружено ни точного, ни нечеткого сравнения, переходим к анализу следующего найденного документа.

Таблица 1. Способы сравнения

| Признак | Поле док-та | Способ сравнения | Порог. значение |
|---|---|---|---|
| **Фамилия** (рус. яз.) | 200$a | Расстояние редактирования | 2 |
| **Фамилия** (англ. яз.) | 400$a | | |
| **Имя** (рус. яз.) | 200$g | | |
| **Имя** (англ. яз.) | 400$g | | |
| **Пол** | 120$a | Строгое равенство | 0 |
| **Место работы** (организация) | 601$a | Относительное расстояние редактирования | 30% |

Если временный авторитетный документ прошел процедуру сравнения с каждым документом из множества для сравнения, но при этом не было найдено ни одного строгого или нестрогого дубликата, то он помещается в авторитетную базу данных.

Относительное расстояние редактирования определяется как отношение расстояния редактирования к длине первой из двух сравниваемых строк.

Пороговые значения были установлены эмпирически. В дальнейшем планируется провести более подробное исследование для выбора пороговых значений.

## 5 Оценка качества идентификации

Оценивать качество идентификации персон в рамках данной работы предлагается с помощью широко распространенных показателей: полноты и точности [28].

Показатель полноты можно рассчитать с помощью следующей формулы:

$$Recall = \frac{TruePostive}{TruePostive + FalseNegative}, \quad (1)$$

где *TruePositive* – количество верно установленных связей с созданными авторитетными документами, *FalseNegative* – количество упущенных связей.

Точность оценивается по формуле

$$Precision = \frac{TruePostive}{TruePostive + FalsePostive}, \quad (2)$$

где *FalsePositive* – количество неверно установленных связей.

Для расчета описанных показателей необходима тестовая выборка, на основе которой можно было бы рассчитать количество ошибок и верно установленных связей между документами.

При этом можно оценивать полноту и точность в двух вариантах. Если рассматривать все возможные комбинации документов как потенциальные связи, то в результате получим оценку метода в целом. А если среди связей рассматривать только те, которые были отобраны как потенциальные с помощью процедуры поиска подобных документов, то получим оценку работы механизма сопоставления документов. Второй вариант подходит в том случае, если на этапе сужения круга документов для сравнения не происходит потери нужных связей.

## 6 Описание программного комплекса

### 6.1 Функциональное описание программного комплекса *cflib*

В качестве системы идентификации персон (рис. 1) в данной работе выступает программный комплекс *cflib*, состоящий из следующих модулей:

- *cfchk* – проверка и коррекция входных документов, внедрение временных меток содержимого;
- *cfwrk* – сравнение временного и найденного документов;
- *cfsearch* – поиск в авторитетной базе данных;
- *cfupdate* – дополнение документа из авторитетной базы данных.

Основные модули программного комплекса *cflib* представлены на рисунке 2.

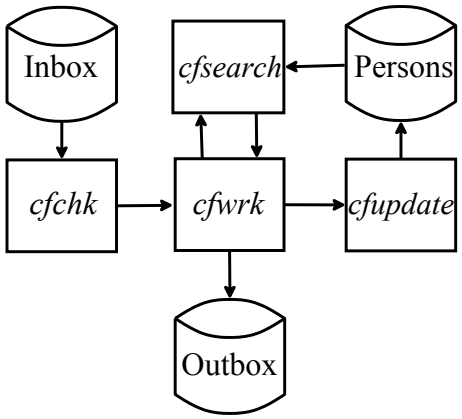



Рис. 2. Основные модули программного комплекса

Модуль *cfchk* кроме проверки входного документа на корректность также осуществляет внедрение меток содержимого, используемых для работы. К таким меткам относятся временные идентификаторы для отдельных персон, наборы биграмм и другая информация, которая понадобится при дальнейшей работе.

При помещении документа в базу данных Persons он индексируется в соответствии с подготовленным профилем индексирования. В этот профиль включено индексирование по биграммам, что позволяет модулю *cfsearch* извлекать подобные

документы для поиска среди них нечетких дубликатов.

В работе модуля *cfwrk* предусмотрены несколько этапов сравнения для выявление нечетких дубликатов. При этом можно изменить функции сравнения и использовать другие методы сравнения, не нарушая логики работы программного комплекса.

Модуль *cfupdate* предназначен для дополнения авторитетного документа отсутствующей информацией или вариантными значениями отдельных полей.

Программный комплекс *cflib* является платформо-независимым и может работать под управлением различных операционных систем или сред, включая Cygwin для MS Windows. Комплекс *cflib* написан на нескольких языках программирования: C, Perl и XSLT

## 7 Заключение

В данной работе приводится описание системы идентификации персон, которая создавалась в процессе разработки Единого репозитория результатов научно-технической деятельности (РНТД) в ИВТ СО РАН.

Особенностью данной системы является то, что в процессе ее работы создается авторитетная база данных с описаниями персон в автоматическом режиме, без участия пользователя. Первое встреченное упоминание ложится в основу авторитетного документа, а последующие могут при необходимости дополнять этот документ. Такой подход был выбран из-за того, что при создании системы в нашем распоряжении не было готовой авторитетной базы данных. Однако допускается и возможность подключения готовой базы данных, если она доступна.

В качестве формата авторитетных данных был выбран формат RUSMARC/Authorities [29], широко распространенный в библиотечном сообществе. Такой выбор позволяет в дальнейшем осуществлять простую интеграцию с библиотечными данными.

Для выявления нечетких дубликатов в упоминаниях персон использовались индексирование по биграммам и расстояние редактирования. Сравнение состоит из нескольких этапов. При необходимости можно предусмотреть и больше вариантов сравнения документов, а также изменить используемые для сравнения методы – логика работы системы от этого не пострадает.

В дальнейшей работе планируется рассмотреть различные методы сравнения документов, исследовать возможность построения обучающих выборок из документов в формате CERIF, а также использовать различные сторонние системы для идентификации персон (в частности, систему Silk).

## Литература


[1] Умаров А.С., Попова Н.В., Зелепухина В.А. Некоторые аспекты создания информационных систем для сбора и хранения научной и наукометрической информации // Прикаспийский журнал: управление и высокие технологии. – 2013. – № 3 (23). – С. 111–118.

[2] Scopus. http://www.scopus.com

[3] Thomson Reuters Web of Science. http://thomsonreuters.com/products_services/ science/science_products/a-z/ web_of_science/

[4] ResearchGate. http://researchgate.net

[5] SciVerse. http://www.info.sciverse.com

[6] CrossRef. http://crossref.org

[7] CRIS concept and CRIS benefits. http://www.eurocris.org/ Index.php?page=concepts_benefits&t=1

[8] Астраханский государственный университет. Результаты научной деятельности. http://science.aspu.ru

[9] Guskov A.E., Zhizhimov O.L., Kikhtenko V., Skachkov D.M., Kosyakov D. RuCRIS: A Pilot CERIF based System to Aggregate Heterogeneous Data of Russian Research Projects // Procedia Computer Science. – 2014. – Vol. 33. – P. 163–167. – ISSN 1877-0509. – http://www.sciencedirect.com/science/article/ pii/S1877050914008175/pdf?md5= d74bdd8e7724f217d214b6aaff40c1eapid= 1-s2.0-S1877050914008175-main.pdf

[10] Паринов С.И., Когаловский М.Р. Технология семантического структурирования контента научных электронных библиотек // Труды XIII Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL-2011, Воронеж, 19–22 окт. 2011 г. – Воронеж: Воронежский государственный ун-т, 2011.

[11] Когаловский М.Р., Паринов С.И. Классификация и использование семантических связей между информационными объектами в научных электронных библиотеках // Информатика и ее применения, 2012. Т. 6, вып. 3. С. 31–41.

[12] Talburt J. Entity resolution and information quality / John R. Talburt. – San Francisco : Morgan Kaufmann/Elsevier, 2011. – 256 p.

[13] Winkler W.E. Overview of record linkage and current research directions [Electronic resource] : tech. report / W.E. Winkler ; U.S. Census Bureau, Stat. res. div. – Washington : [s. n.], 2006. – 44 p. – http://www.census.gov/ srd/papers/pdf/ rrs2006-02.pdf

[14] Elmagarmid A., Ipeirotis P., Verykios V. (2007). Duplicate Record Detection: A survey. IEEE Transactions on Knowledge and Data Engineering 19(1): 1–16.

[15] Bilenko M. Learning to Combine Trained Distance Metrics for Duplicate Detection in Databases / M. Bilenko, R. Mooney. Technical Report AI-02-296, Artificial Intelligence Lab, University of Texas at Austin, 2002.

[16] Sarawagi S. Interactive deduplication using active learning / S. Sarawagi, A. Bhamidipat // Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. – P. 269–278.

[17] Parinov S. Open Repository of Semantic Linkages. In: Proceedings of 11th International Conference on Current Research Information Systems e-Infrastructure for Research and Innovations (CRIS 2012), Prague 2012, http://socionet.ru/publication.xml?h=repec:rus:mqijxk:29.

[18] Бездушный А.Н., Кулагин М.В., Серебряков В.А. и др. Предложения по наборам метаданных для научных информационных ресурсов // Вычислительные технологии. – 2005. – Т. 10. – С. 29–48.

[19] Кулагин М.В., Лопатенко А.С. Научные информационные системы и электронные библиотеки. Потребность в интеграции // Сборник трудов Третьей Всероссийской конференции по электронным библиотекам – RCDL'2001, Петрозаводск, 11–13 сент. 2001 г. – С. 14–19.

[20] Мазов Н.А., Гуреев В.Н. Проблемы идентификации метаданных в наукометрических базах данных Web of Knowledge, Scopus и РИНЦ на примере профилей авторов // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса: 19-я междунар. конф. «Крым 2012» (Судак, 2–10 июня 2012 г.): Труды конф. – М.: Изд-во ГПНТБ России, 2012. – С. 1–4. – http://www.gpntb.ru/win/inter-events/crimea2012/disk/124.pdf

[21] Гуреев В.Н., Мазов Н.А. Идентификация в информационных библиографических системах: проблемы и решения // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса: 21-я междунар. конф. «Крым 2014» (Судак, 7–15 июня 2014 г.): Труды конф. – М.: Изд-во ГПНТБ России, 2014. – С. 1–7. – http://www.gpntb.ru/win/inter-events/crimea2014/disk/066.pdf

[22] CERIF. http://www.eurocris.org/Index.php?page=CERIFreleases&t=1

[23] CERIF in Brief. http://cerifsupport.org/cerif-in-brief/

[24] Jaro M. A. Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Society, 84(406): 414–420, 1989.

[25] Hernandez M. A., Stolfo S. J. Real-world data is dirty: data cleansing and the merge/purge problem. Journal of Data Mining and Knowledge Discovery, 1(2), 1998.

[26] Christen P., Churches T. Febrl: Freely extensible biomedical record linkage Manual, release 0.2.2 edition, November 2003.

[27] Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Докл. Акад. наук СССР. – 1965. – Т. 163, № 4. – С. 845–848.

[28] Manning C., Raghavan P., Schütze H. Introduction to Information Retrieval. – Cambridge University Press, 2008. – ISBN 0-521-86571-9.

[29] Российский коммуникативный формат (RUSMARC) [Электронный ресурс] : [сайт] / Мин-во культуры Рос. Федерации, Рос. библ. ассоц., Нац. служба развития системы форматов RUSMARC. http://www.rusmarc.ru/index.html

## Experience of Person Identification for CRIS-Systems

Anna A. Knyazeva, Igor Y. Turchanovsky, Oleg S. Kolobov, Oleg L. Zhizhimov

The system of persons identification which was created in the process of development of a unified repository of scientific and technical activities (RSTA) in ICT SB RAS is described in this paper. The principles and methods used to create the system as well as its structure are briefly described. An algorithm for establishing an authoritative database with descriptions of persons automatically, without user intervention, is given. Indexing with bigrams and editing distance were used for detecting near-duplicate references to persons.

# Подход к поиску потоков работ по метаданным

© Н.А. Скворцов
Институт проблем информатики РАН
nskv@ipi.ac.ru

**Аннотация**

Работа посвящена методам поиска реализаций потоков работ и их компонентов с целью повторного использования по спецификациям метаданных. Для спецификации потоков работ используются диалекты языка правил RIF, метаданные формулируются как аннотации RIF. Метаданные, необходимые для обеспечения повторного использования потоков работ, применяются в различных задачах, возникающих во время разработки потоков работ. В статье демонстрируются методы спецификации метаданных и семантического поиска потоков работ по ним.

## 1 Введение

Необходимость обработки больших объёмов данных и расширение направлений их обработки при исследованиях в науках с интенсивным использованием данных заставляет подходить к множеству средств обработки данных как к коллекциям научных методов, которые могут быть повторно используемы в различных задачах. Для организации обработки данных становится целесообразно разрабатывать потоки работ, которые представляют собой спецификации порядка обработки данных, обеспечивающего решение научных задач, и использовать существующие деятельности, сервисы, потоки работ.

Спецификации потоков работ в данном исследовании используют языки и технологии, применяемые в рамках Семантического веба. Основными средством спецификации потоков работ являются диалекты языка RIF [2] (Rule Interchange Format). Потоки работ специфицируются в мультидиалектной среде [6]. Деятельности потоков работ могут формулироваться в разных диалектах правил. Концептуальные схемы предметных областей, над которыми разрабатываются спецификации деятельностей, описываются средствами языка онтологий OWL 2.0 [1] и импортируются в спецификации потоков работ. Определенные в концептуальных схемах сущности могут использоваться в качестве предикатов в правилах. Оркестровка потока работ выражается посредством продукционных правил (в диалекте RIF PRD). В продукционных правилах могут использоваться предикаты, определённые в других диалектах при спецификации деятельностей. Для спецификации управляющих конструкций потоков работ определяется пространство имён со специальными предикатами:

- variable-definition и variable-value для организации потоков данных на основе переменных и их значений;
- parameter-definition и parameter-value для организации входных и выходных параметров потоков работ и значений параметров;
- end-of-task – индикатор завершения работы деятельности для организации последовательности, условий, разбиения, соединения и других шаблонов [9] потоков работ с помощью правил.

Например, следующая спецификация определяет шаблон разбиения по конъюнкции, в котором деятельности B и C выполняются одновременно после выполнения деятельности A:

```
If Not(External(wkfl:end-of-task(A)))
Then Do (Act(A)
  Assert(External(wkfl:end-of-task(A))))
If And(Not(External(wkfl:end-of-task(B)))
  External(wkfl:end-of-task(A)))
Then Do (Act(B)
  Assert(External(wkfl:end-of-task(B))))
If And(Not(External(wkfl:end-of-task(C)))
  External(wkfl:end-of-task(A)))
Then Do (Act(C)
  Assert(External(wkfl:end-of-task(C))))
```

Реализации потоков работ могут либо разрабатываться на основе спецификаций RIF при помощи трансляции правил в языки конкретных систем, работающих с определёнными диалектами правил, либо выбираться из существующих релевантных потоков работ, их фрагментов, отдельных деятельностей и сервисов.

Поиск релевантных потоков работ и их фрагментов производится в доступных коллекциях научных методов. Для возможности семантического поиска в таких коллекциях реализации потоков работ, помимо спецификации их структуры, сопровождаются определённым набором метаданных, несущих информацию о связи потоков

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

работ с понятиями предметной области, о качестве и происхождении используемых данных и методов. Состав необходимых метаданных был разработан ранее [13]. Эта информация обеспечивает не только возможность оценки потоков работ и их фрагментов с точки зрения структуры, но и учёт семантики предметной области и требований к качеству и надёжности работы научных методов.

Принципы семантического поиска подходящих реализаций потоков работ и их фрагментов на основе метаданных являются предметом исследования данной статьи. В следующем разделе описаны принципы связывания метаданных со спецификациями потоков работ на правилах. Затем приведён обзор методов поиска потоков работ. Последующие разделы рассматривают сценарии и методы семантического поиска релевантных потоков работ.

## 2 Связывание метаданных с потоками работ

Спецификации RIF несут формальную семантику правил, не позволяющую определять что-либо помимо правил в заданном диалекте. Для связывания со спецификациями дополнительной информации в языке предусмотрен механизм аннотирования. Аннотации могут сопровождать любой класс конструкций RIF в спецификациях правил. Они определяются как фреймы с наборами свойств этих конструкций, которые должны быть сохранены при любых манипуляциях спецификациями, но не добавляют семантики с точки зрения правил. Поэтому при реализации потоков работ спецификации метаданных игнорируются. Тем не менее, они могут обладать семантикой, не зависимой от правил.

Обычно аннотации в RIF определяются в терминах специализированного словаря, специфицирующего набор предопределённых свойств. Состав метаданных в настоящем исследовании не ограничивается набором свойств, а включает в себя более развитые описания. В качестве словарей метаданных используются онтологии предметных областей, а также онтологии, определяющие свойства элементов потоков работ в различных ракурсах рассмотрения, таких как качество и происхождение данных и методов.

Аннотации, которые определяют метаданные, целесообразно связывать со следующими элементами потоков работ, выраженных правилами:

- потоки работ в целом;
- входные и выходные параметры потоков работ;
- деятельности внутри потоков работ;
- входные и выходные параметры деятельностей;
- переменные, определяющие потоки данных;
- отдельные правила и группы правил, определяющие фрагменты потока работ;
- группы правил, определяющие шаблоны потоков работ [9].

Связывание метаданных с потоками работ и поиск релевантных элементов потоков работ далее рассмотрим на примере. В [6] описывается задача составления портфелей ценных бумаг, котировки которых не коррелируют друг с другом, и выбора лучшего из них по определённым критериям.

Для решения данной задачи разрабатываются спецификации потока работ, включающего:

- задачу поиска максимальных портфелей-кандидатов с независимыми друг от друга котировками бумаг;
- оценку бумаг, входящих в портфели, с точки зрения разных критериев, в частности, финансово-экономического и социального;
- оценка портфелей по соответствующим критериям как обобщение оценок бумаг, входящих в них;
- обобщение нескольких критериев оценки портфелей в общую оценку и выбор лучшего портфеля.

Для реализации потока работ используются данные об истории цен на бумаги, принадлежность компаний индексу S&P 500 (индекс оценивается на основе данных о капитализации пятисот крупных американских компаний), оценка соотношения доходности и риска, мониторинг тональности высказываний инвесторов об определённых бумагах. Оценка по разным критериям выполняется в потоке работ параллельными ветвями.

Для описания метаданных в терминах предметной области определяется онтология.

```
Class(Portfolio)
ObjectProperty(includesSecurity)
  ObjectPropertyDomain(includesSecurity Portfolio)
  ObjectPropertyRange(includesSecurity Security)

Class(Security)
ObjectProperty(hasIdentifier)
  FunctionalObjectProperty(hasIdentifier)
  ObjectPropertyDomain(hasIdentifier Security)
  ObjectPropertyRange(hasIdentifier Ticker)
ObjectProperty(listedIn)
  ObjectPropertyDomain(listedIn Security)
  ObjectPropertyRange(listedIn StockMarketIndex)
ObjectProperty(hasRate)
  ObjectPropertyDomain(hasRate Security)
  ObjectPropertyRange(hasRate StockMarketRate)
ObjectProperty(hasMetric)
  ObjectPropertyDomain(hasMetric Security)
  ObjectPropertyRange(hasMetric Metric)
ObjectProperty(correlatesWith)
  ObjectPropertyDomain(correlatesWith Security)
  ObjectPropertyRange(correlatesWith Security)

ObjectProperty(hasMetric)
  ObjectPropertyDomain(hasMetric Security)
  ObjectPropertyRange(hasMetric Metric)
  ObjectPropertyDomain(hasMetric Portfolio)

Class(StockMarketRate)
ObjectProperty(onDate)
```

```
  FunctionalObjectProperty(onDate)
  ObjectPropertyDomain(onDate StockMarketRate)
  ObjectPropertyRange(onDate Date)

Class(Metric)
ObjectProperty(isMetricOfSecurity)
Class(Correlation)
  SubClassOf(Correlation Metric)
  SubClassOf(Correlation ObjectAllValuesFrom
    (isMetricOfSecurity Security))
Class(FinancialMetric)
  SubClassOf(FinancialMetric Metric)
Class(SocialMetric)
  SubClassOf(SocialMetric Metric)
```

Онтология[1] определяет следующие основные понятия:

- Portfolio – портфель, составленный из ценных бумаг определённого списка компаний, имеющий, с ним также могут быть связаны метрики оценки портфеля;
- Security – ценные бумаги компании, участвующие в фондовом рынке, у них есть идентификаторы, они могут принадлежать списку фондового индекса, оцениваются котировками, метриками надёжности, могут иметь зависимость от других бумаг;
- StockMarketRate – котировка бумаги, зависящая от времени;
- Metric – метрика для оценки надёжности ценной бумаги или портфеля; одной из метрик оценки надёжности бумаги является корреляция её котировки с другими бумагами.

Особо отметим, что представленная онтология определяет понятия и связи предметной области фондового рынка в отличие от спецификации концептуальной схемы (названной в [6] онтологией области приложения), определяющей представление данных при решении задачи в потоке работ на правилах, хотя и онтология, и концептуальная схема используют выразительные средства, определяемые языком OWL 2. Описания концептуальной схемы недостаточны для использования в метаданных о предметной областью, так как многие понятия отношения предметной области сведены в ней к примитивным типам данных. Подробнее различия и связи онтологий и концептуальных схем предметных областей обсуждаются в [12].

Одновременно с онтологией предметной области для определения метаданных потоков работ используются другие онтологии, определяющие различные аспекты описываемых элементов потоков работ. В частности, для связывания правил с видами элементов потоков работ, которые определены этими правилами, используется онтология структуры потоков работ[2].

```
Class(Workflow)
ObjectProperty(hasTask)
  ObjectPropertyDomain(hasTask Workflow)
  ObjectPropertyRange(hasTask Task)

Class(Task)
ObjectProperty(hasParameter)
  ObjectPropertyDomain(hasParameter Task)
  ObjectPropertyRange(hasParameter TaskParameter)
  InverseObjectProperties(isParameterOf
    hasParameter)
ObjectProperty(hasInputParameter)
  SubObjectPropertyOf(hasInputParameter
    hasParameter)
  ObjectPropertyDomain(hasInputParameter Task)
  ObjectPropertyRange(hasInputParameter
    InputParameter)
  InverseObjectProperties(isInputParameterOf
    hasInputParameter)
ObjectProperty(hasOutputParameter)
  SubObjectPropertyOf(hasOutputParameter
    hasParameter)
  ObjectPropertyDomain(hasOutputParameter Task)
  ObjectPropertyRange(hasOutputParameter
    OutputParameter)
  InverseObjectProperties(isOutputParameterOf
    hasOutputParameter)
```

В приведённом фрагменте онтологии структуры потока работ определены понятия:

- Workflow – поток работ в целом, состоящий из набора деятельностей;
- Task – деятельность, которая может иметь входные и выходные параметры.

Помимо этого онтология определяет разновидности деятельностей, такие как начало и завершение потока, вызов подпотока, шаблоны управления потоками и другие понятия.

В терминах двух представленных онтологий приведём пример аннотации, определяющей метаданные для выходного параметра деятельсности (спецификация представлена в формате RIF XML):

```
<declare><Var>
  <id>
    <Const>GetPortfolios_Output</Const>
  </id>
  <meta>
  <Frame>
    <object>
      <Const>GetPortfolios_Output</Const>
    </object>
    <slot>
      <Const>rdf:type</Const>
      <Const>wf:OutputParameter</Const>
    </slot>
    <slot>
      <Const>wf:isOutputParameterOf</Const>
      <Const>:GetPortfolios</Const>
    </slot>
    <slot>
      <Const>rdf:type</Const>
      <Const>pont:Portfolio</Const>
    </slot>
  </Frame>
  </meta>
?p
</Var></declare>
```

Данная спецификация метаданных определена для переменной `?p` в правиле, соответствующем деятельности потока работ. В первую очередь, она определяет в текущем пространстве имён уникальный идентификатор данного элемента правила RIF (`GetPortfolios_Output`). С этим идентификатором связываются метаданные в

[1] http://ontology.ipi.ac.ru/ontologies/stockmarket.owl
[2] http://ontology.ipi.ac.ru/ontologies/wf.owl

терминах двух определённых выше онтологий (пространство имён `pont` соответствует онтологии предметной области, а `wf` – онтологии структуры потоков работ). Во-первых, определяется, что элемент с данным идентификатором является выходным параметром (экземпляром класса `OutputParameter`) деятельности, решающей подзадачу поиска портфелей (отношение `isOutputParameterOf` к объекту с идентификатором `GetPortfolios`), а также является экземпляром класса `Portfolio`, то есть возвращаемые деятельностью данные должны являться портфелями. Идентификатор `GetPortfolios`, должен быть определён подобным образом в метаданных, связанных с правилом в целом.

Таким образом, метаописание позволяет связать спецификации правил с предметной областью, в которой решается задача, определить части правил, которые соответствуют элементам потоков работ, а также семантически связать элементы друг с другом с помощью выражений в терминах онтологий.

## 3 Обзор методов, связанных с повторным использованием потоков работ

В большинстве исследований, посвящённых метаданным потоков работ, состав метаданных ограничивается набором предопределённых свойств для работы с простыми сопроводительными данными: именами, вербальными определениями, информацией об авторах, версиях, правах, дате создания и других достаточно ограниченных описаниями [11]. Такие подходы к спецификации метаданных представляются недостаточными для выразительного семантического описания и поиска потоков работ.

Как аннотирование потоков работ метаданными использует простые поля описаний, так же большинство проектов, работающих с потоками работ, ограничиваются методами поиска на основе ключевых слов, относящихся к потокам работ как цельным объектам [5]. Проект wf4ever [10] предоставляет набор средств для поддержки повторного использования, включая аннотирование потоков работ в целом и их компонентов, учитывает в сопровождающих спецификациях происхождение данных, являющихся результатами работы процессов, многоверсионность и другие аспекты. Проект OPM [7] использует развитую модель происхождения данных для выражения семантики воспроизводимости результатов, в том числе, для потоков работ.

В контексте настоящего исследования необходимо упомянуть подходы process mining [4], специализирующиеся, главным образом, на анализе лог-файлов. В исследованиях используются модели процессов, являющиеся спецификациями структуры потоков работ. Записи логов исполняемых деятельностей или происходящих событий сопоставляются моделям процессов. На основе лог-файлов решаются следующие виды задач.

- Под задачей обнаружения потоков работ понимается восстановление фактической структуры потока работ по лог-файлам работы его экземпляра. Таким образом, могут быть вскрыты потоки работ, не имеющие формальных спецификаций модели процесса.
- Задача установления конформности (conformance) потока работ заключается в проверке соответствия модели потока работ данным, получаемым из лог-файлов о работе его реализации.
- Задача усовершенствования модели потока работ отличается от задачи установления конформности тем, что модель не только оценивается на соответствие реальным событиям, но и меняется для более точного соответствия.

Эти исследования рождают множество публикаций с развитием и применением представленных задач. Они полезны для решения задач поиска потоков работ по спецификации их структуры, для описания и дальнейшего повторного использования доступных потоков работ, не имеющих формальной спецификации, но генерирующих лог-файлы во время своей работы, для контроля соответствия реализованных и найденных потоков работ спецификациям.

## 4 Организация поиска релевантных потоков работ по сформулированным требованиям

Благодаря тому, что в используемой в данном исследовании модели потоков работ спецификации правил, выражающие семантику их поведения, независимы от сопровождающих их метаданных, правила и метаданные могут обрабатываться независимыми инструментами. Спецификации правил используются для реализации потоков работ в определённых системах, исполняющих их в соответствии с семантикой используемых диалектов. Для предварительного связывания элементов спецификаций потоков работ должны использоваться метаданные. Для этого реализуется независимая от спецификаций правил возможность поиска потоков работ по метаданным.

Спецификации фреймов, содержащие значения метаданных, преобразуются в триплеты RDF в соответствии с рекомендациями W3C [3], сохраняются в отдельном хранилище RDF и в дальнейшем используются для запросов поиска по метаданным. В частности фрейм RIF с метаданными, соответствующий XML-представлению в приведённом выше примере:

```
GetPortfolios_Output
  [ rdf:type -> wf:OutputParameter,
    wf:isOutputParameterOf -> GetPortfolios,
    rdf:type -> pont:Portfolio ],
```

будет преобразован в триплеты RDF

```
pwf:GetPortfolios_Output
   rdf:type wf:OutputParameter;
   wf:isOutputParameterOf pwf:GetPortfolios;
   rdf:type pont:Portfolio.
```

Таким образом, база триплетов собирает в себе набор метаданных и идентификаторов, по которым можно установить, с какими именно элементами спецификации потоков работ на правилах связаны определённые метаданные. В качестве RDF-словарей может использоваться произвольный набор онтологий, в частности определяющих состав метаданных, разработанный в [13]. Поиск потоков работ и их фрагментов по метаданным организуется с помощью задания запросов на языке SPARQL [8] к базе триплетов, содержащей метаданные.

Запросы на языке SPARQL формулируются в соответствии с требованиями задачи, которая должна быть решена в предметной области, либо с требованиями спецификации потока работ, который необходимо реализовать с помощью повторного использования существующих потоков работ, их фрагментов и доступных сервисов.

## 5 Поиск релевантных потоков работ в целом и их фрагментов

Поиск потоков работ для обеспечения их повторного использования при наличии метаданных, требуемых в [13], производится на основании соответствия выбранных или всех одновременно критериев:

- соответствие потока работ понятиям или выражениям в терминах понятий онтологии предметной области, описывающих зависимости/функции, методы, процессы, могущие применяться в данной предметной области;
- соответствие понятий или выражений в терминах понятий, описывающих входные и/или выходные параметры потоков работ (например, для поиска методов, которые из определённого набора параметров получают требуемый тип результата);
- выполнение требований к качеству входных данных и качеству возвращаемых результатов потока работ в терминах онтологии качества данных (например, требования актуальности);
- требования к происхождению потока работ (например, по автору разработанных реализаций);
- требования к происхождению входных данных (например, определённое оборудование, которым собраны первичные данные наблюдений).

Таким образом, требования к искомым в коллекции научных методов потокам работ могут затрагивать как функциональность реализуемых ими научных методов, так и предусловия и постусловия, выраженные в терминах онтологий, а также требования к надёжности применяемых методов, используемых данных и получаемых результатов.

Информация о происхождении и качестве данных и методов в потоках работ используется для

- спецификации достоверности, полноты, точности требуемых данных и достигаемых результатов
- контроля реальных источников данных и их качества в соответствии с требованиями задачи;
- контроля соответствия требованиям решения задачи используемых открытых реализаций научных методов.

Помимо этого, решение научных задач предметной области может выбираться как фрагментарно из других потоков работ, так и из отдельных фрагментов потоков работ и из существующих сервисов. Необходимый фрагмент обработки данных может оказаться частью реализации потока работ, решающего в целом отличную задачу. Для этого требования в запросах формулируются не к потокам работ в целом, а к параметрам деятельностей в составе потоков работ.

В качестве примера зададим запрос для поиска потоков работ, реализующих метрики оценки надёжности портфелей ценных бумаг.

```
select distinct ?task1 ?task2 where
{
  ?task1 rdf:type pont:Metric .
  ?in1 wf:isInputParameterOf ?task1 .
  ?in1 rdf:type pont:Security .
  ?var wf:isOutputParameterOf ?task1 .
  ?var rdf:type pont:Metric .
  ?in1 pont:hasMetric ?var .
  ?task2 rdf:type pont:Metric .
  ?var wf:isInputParameterOf ?task2
  ?in2 wf:isInputParameterOf ?task2 .
  ?in2 rdf:type pont:Portfolio .
  ?in2 pont:includesSecurity ?in1 .
  ?out2 wf:isOutputParameterOf ?task2 .
  ?out2 rdf:type pont:Metric .
  ?in2 pont:hasMetric ?out2 .
}
```

По условию запроса необходимо найти деятельности, одна из которых принимает на вход объекты ценных бумаг, вычисляет и возвращает для него некоторую метрику, а вторая деятельность принимает на вход результаты первой деятельности, и вычисляет обобщающую метрику для портфеля, содержащего ценные бумаги, для которых вычислена первая метрика.

Представим, что в базе триплетов хранятся метаданные следующих деятельностей:

- getPositiveTweetRatio – вычисляет тональность сообщений о ценной бумаге в Twitter;
- computePortfolioTwitterMetrics – на основе тональности сообщений о ценных бумагах вычисляет тональность отношения к содержащему их портфелю;
- getSecurityFinancialMetrics – вычисляет метрику надёжности ценной бумаги, учитывающую выгоду и риски на основе истории котировок;
- computePortfolioFinancialMetrics – для портфеля в целом, содержащего ценные бумаги, вычисляет обобщённую финансовую метрику.

Например, для одной из деятельностей и её структурных элементов хранятся следующие триплеты:

```
fin:getSecurityFinancialMetrics
   rdf:type wf:Task;
   rdf:type pont:FinancialMetric;
   wf:hasInputParameter fin:finMetricPar;
   wf:hasOutputParameter fin:securityPar;
fin:securityPar
   rdf:type pont:Security;
   rdf:type wf:InputParameter;
fin:finMetric
   rdf:type pont:FinancialMetric;
   rdf:type wf:OutputParameter;
```

При условии адекватного описания метаданными спецификаций деятельностей и их связей друг с другом внутри потоков работ ответ на запрос будет содержать следующие кортежи:

```
<sparql xmlns=http://www.w3.org/2005/sparql-results#>
   <head>
      <variable name="task1"/>
      <variable name="task2"/>
   </head>
<results>
   <result>
      <binding name="task1">

   <uri>http://ontology.ipi.ac.ru/portfolio.rif#
        getPositiveTweetRatio</uri>
      </binding>
      <binding name="task2">

   <uri>http://ontology.ipi.ac.ru/portfolio.rif#
        computePortfolioTwitterMetrics</uri>
      </binding>
   </result>
   <result>
      <binding name="task1">

   <uri>http://ontology.ipi.ac.ru/portfolio.rif#
        getSecurityFinancialMetrics</uri>
      </binding>
      <binding name="task2">

   <uri>http://ontology.ipi.ac.ru/portfolio.rif#
        computePortfolioFinancialMetrics</uri>
      </binding>
   </result>
</results>
</sparql>
```

Таким образом, найдены спецификации деятельностей, которые можно использовать повторно для реализации метрик ценных бумаг и портфелей при решении задачи выбора наилучшего портфеля.

Спецификация потока работ, использующего найденные спецификации, может быть следующей (рис. 1) [6]:

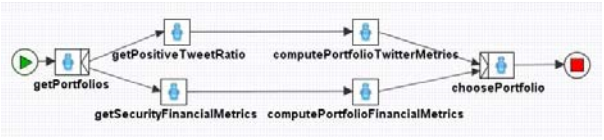



Рис 1. Поток работ для решения задачи выбора лучшего портфеля ценных бумаг

В общем случае выделение релевантных фрагментов является нетривиальной задачей, требующей решения в соответствии со структурой и семантикой выполняемых действий каждым компонентом в составе фрагмента. Эта задача привлекает и метаданные, и сравнение спецификаций шаблонов, и проверку конформности спецификаций и реализаций, и работу экспертов.

Поддержание при спецификациях потоков работ на правилах стиля, при котором управляющая часть потока работ набирается не из произвольных правил, а из наборов правил, реализующих определённые известные шаблоны [9], может упростить проверку конформности. Целесообразно ввести также метаданные, обозначающие те или иные шаблоны в терминах онтологии структуры потоков работ.

## 6 Семантический контроль используемых методов и принятых решений

Метаданные целесообразно использовать не только при поиске потоков работ и их компонентов в коллекциях научных методов, но и для дальнейшей проверки совместимости семантики данных и интероперабельности потоков работ и фрагментов при объединении найденных и выбранных компонентов для реализации решения научных задач.

Для этого необходимо проводить следующие проверки:

- корректность включения в качестве деятельности данного потока работ существующих компонентов в качестве подпроцессов по их входным и выходным параметрам;
- соответствие семантики входных компонентов семантике входных данных и соответствие выходных данных выходным параметрам по понятиям предметной области, требованиям к качеству, происхождению и другим возможным критериям, учтённым с помощью онтологий;
- соответствие семантики данных, проходящих из выхода одного компонента на вход другого.

Эти проверки выполняются по принципу спецификаций пред- и постусловий: постусловие выхода предыдущего компонента должно быть строже предусловия входа последующего компонента. Требования могут включать как выражения в терминах понятий предметной области, так требования качества и происхождения данных.

## 7 Проведение экспериментов и проверка интероперабельности потоков работ на тестовых наборах данных

Требования к релевантности по метаданным могут быть в той или иной степени выразительными, а структурное соответствие само по себе не включает проверку семантики компонентов. К тому же реализации спецификаций могут использовать разные инструменты, и по деталям работы они могут отличаться друг от друга. Поэтому для надёжного повторного использования

реализаций потоков работ необходимо проверять их на определённых наборах тестовых данных.

Тесты включают набор данных и требования к ожидаемым результатам, достаточные для проверки всех возможных особых случаев, могущих возникать в потоке работ. Помимо входных и проверочных выходных данных тесты могут включать метаданные.

При тестировании производится контроль прохождения тестов по определённому пути в потоке работ в зависимости от входных данных. Для этого в состав тестов включаются метаданные происхождения данных. Метаданные происхождения, сгенерированные в результате прогона потока работ на тестовых данных, проверяются на соответствие происхождения данных в составе тестов.

Проверяется соответствие результатов требованиям качества, предоставляемым в спецификациях тестов или специфицирующих выходные параметры потока работ.

Помимо этого, требования тестов могут налагаться и на описания исполняемых сред. Для этого также используются метаданные на основе онтологий описания исполняемых сред [13].

Другой подход тестирования реализаций потоков работ, собранных на основе спецификаций, предполагает генерацию лог-файлов при прохождении тестов, активизирующих все возможные пути в потоке работ. При реализации правил RIF компонентами, использующими различные системы вывода, лог-файлы должны генерироваться каждой из них. Решение задачи установления конформности [4] спецификации потока работ и получившейся реализации позволяет подтвердить их соответствие друг другу.

## 8 Заключение

Работа посвящена организации семантического поиска потоков работ и их фрагментов по метаданным с целью их повторного использования. Она является продолжением исследования, представленного в [13], применяемого к другим техническим условиям. В качестве модели потоков работ [6] используются языки на правилах, что даёт богатые возможности в повышении выразительности спецификаций и в применимых методах анализа потоков работ. В статье разработан подход к представлению и обработке метаданных в данной модели потоков работ. Упор делается на сценариях применения метаданных потоков работ для поиска потоков работ с целью их повторного использования и для проверки их релевантности и интероперабельности.

## Благодарности

Работа выполнена при поддержке грантов РФФИ 13-07-00579, 14-07-00548 и Программы Президиума РАН.

## Литература

[1] OWL 2 Web Ontology Language Document Overview (Second Edition) – W3C, 2011. – URL: http://www.w3.org/TR/owl-overview/

[2] RIF Overview. – W3C, 2013. – URL: http://www.w3.org/TR/rif-overview/

[3] RIF RDF and OWL Compatibility. – W3C, 2013. – URL: http://www.w3.org/TR/rif-rdf-owl/

[4] W.M.P. Van der Aalst. Process mining: Discovery, Conformance and Enhancement of Business Processes. Springer, Heidelberg, 2011.

[5] C.A. Goble, D.C. De Roure. myExperiment: social networking for workflow-using e-scientists // Proceedings of the 2nd workshop on Workflows in support of large-scale science. – ACM, 2007. – С. 1–2.

[6] L. Kalinichenko, S. Stupnikov, A. Vovchenko, D. Kovalev. Multi-dialect Workflows // ADBIS’2014. – 2014. – LNCS 8716. – P. 352–365.

[7] L. Moreau. Provenance-based reproducibility in the semantic web // Web semantics: science, services and agents on the World Wide Web. – 2011. – Vol. 9, No. 2. – P. 202–221.

[8] Polleres A. SPARQL1. 1: New features and friends (OWL2, RIF) // Web Reasoning and Rule Systems. – Springer Berlin Heidelberg, 2010. – С. 23–26.

[9] N. Russell, A.H.M. ter Hofstede, W.M.P. van der Aalst, and N. Mulyar. Workflow Control-Flow Patterns: A Revised View. – BPM Center Report BPM-06-22, BPMcenter.org. – 2006.

[10] S. Sanchez, et al. WF4Ever: Supporting for reuse and reproducibility in experimental science // EGI Technical Forum. – 2012.

[11] C. Tejo-Alonso et al. Metadata for web ontologies and rules: Current practices and perspectives // Metadata and Semantic Research. – Springer Berlin Heidelberg, 2011. – С. 56–67.

[12] А.Е. Вовченко и др. От спецификаций требований к концептуальной схеме // RCDL'2010. – Казань: КФУ, 2010. – С. 375–381.

[13] Н.А. Скворцов, Д.О. Брюхов, Л.А. Калиниченко, Д. Ковалёв, С.А. Ступников. Метаданные о научных методах для обеспечения их повторного использования и воспроизводимости результатов // RCDL’2013. – Ярославль, 2013.

## An Approach to Search of Workflows by Metadata

Nikolay A. Skvortsov

The work is dedicated to methods of search of workflow implementations and their components for reuse by metadata specifications. Workflow specifications are formulated in the RIF language dialects, metadata is represented as RIF annotations. A set of metadata needed for workflow reuse is applied in various tasks during workflow development. The paper demonstrates methods of metadata specifications and semantic search of workflows using them.

# Изучение структуры и динамики сообщества пользователей в массовой многопользовательской онлайн-игре реального времени

© А.В. Сычев
Воронежский государственный университет
Воронеж
sav@sc.vsu.ru

**Аннотация**

Проведен анализ данных из профилей игроков и кланов для массовой многопользовательской игры World of Tanks. Рассмотрены структура сообщества игроков и кланов и факторы, определяющие их эффективность в игре.

## Введение

Массовая многопользовательская онлайн-игра (ММОИ) — это сетевая компьютерная игра, в которой одновременно принимает участие большое количество игроков (обычно десятки тысяч и более, причем, это число может доходить до миллиона и более).

Особенностью ММОИ является ее функционирование исключительно через Интернет. Количество игроков на игровом сервере ММОИ — не фиксированное, игроки могут свободно входить и выходить из игры, при этом игровая вселенная постоянно функционирует. Игровая сессия является непрерывной (может прерываться лишь во время технических работ или неполадок). В рамках ММОИ игроки находятся в едином игровом пространстве, и при этом они взаимодействуют друг с другом, развиваются, соревнуются, объединяются в различные группы [1].

Как отмечается в [2], ММОИ предоставляет принципиально новый способ наблюдения за сотнями тысяч одновременно социально взаимодействующих индивидуумов, вовлеченных в виртуальную экономическую деятельность. При этом, огромный набор данных социально-экономического характера предоставляется из единого источника. Игроки могут генерировать виртуальный доход в форме различных видов экономической деятельности, предусмотренных в игре, и направленных на «выживание» игрока. Исследование проведенное в [2], подтверждает возможность использования игровых онлайн сообществ в качестве модели для широкого класса реальных человеческих сообществ.

В работе [3] отмечается, что повсеместное распространение онлайн сервисов, предоставляет возможности для анализа крупномасштабных архивных данных, содержащих обширную информацию о человеческих взаимодействиях, необходимую для понимания разнообразного и сложного человеческого поведения. В сотрудничестве с одним из глобальных провайдеров ММОИ авторами этой статьи был проведен анализ взаимосвязи между разными типами сетей пользовательского взаимодействия в виртуальном мире.

Результаты исследования в [4] подтверждают наличие закономерностей в распределении доходов между игроками в ММОИ игре *Pardus* (http://www.pardus.at) аналогичных тем, что наблюдаются в современных экономических системах. Там же на примере данных из игры были выявлены причины, объясняющие возникновение «богатых» игроков. Показатель «богатства» игрока зависит как от общего времени его участия в игре, так и от факторов его «социализации».

Задача поиска элит в ММОИ Pardus рассмотрена в статье [5]. Члены элиты часто хорошо связаны между собой, что позволяет им оказывать влияние на многих, а также быстро собирать, обрабатывать и распространять информацию. Как считают авторы, элиты образуются не просто из лиц с большим числом связей, но и из посредников, соединяющих между собой концентраторы социально-сетевых ресурсов и формирующих сплоченную и структурированную элит-подгруппу в виде ядра социальной сети.

В статье [6] был проведен анализ сетевых моделей в рамках ММОИ игры EverQuest II, который показал, что коммуникабельность среди игроков оказалась довольно диффузной, причем значительное число пользователей предпочитают играть в одиночку, несмотря на наличие встроенных

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

механизмов, поощряющих совместную игру. Это исследование также показало, что шаблоны социального взаимодействия в этой игре коллективно формируются как целями и стилями игроков так и под влиянием компьютерного «кода», конструирующего социальную архитектуру игры.

В данной работе после краткого описания игры World of Tanks и набора данных (в разделе 1) приводятся результаты общего анализа профилей игроков и кланов на Глобальной карте (в разделе 2). Для анализа и визуализации данных был использован пакет интеллектуального анализа данных RapidMiner.

# 1 ММОИ World of Tanks

## 1.1 Описание ММОИ World of Tanks

В данной работе был использован набор данных, относящихся к наиболее популярной на сегодня в России ММОИ World of Tanks (WoT). Доступ к этим данным предоставляется компанией-разработчиком игры WOT — Wargaming.net [4].

ММОИ World of Tanks (Мир танков, http://worldoftanks.ru/) — компьютерная игра, клиентская массовая многопользовательская онлайн-игра в реальном времени в жанре аркадного танкового симулятора в контексте Второй мировой войны. Концепция «World of Tanks» базируется на командных танковых сражениях в режиме PvP (игрок против игрока). Онлайн-релиз русской версии игры состоялся 12 августа 2010 г.

По данным на 28.12.2013 г. в игре было зарегистрировано порядка 75 млн аккаунтов по всему миру. 19.01.2014 г. был поставлен рекорд по числу одновременно играющих пользователей на русскоязычном сервере — 1114337 игроков.

Игровой процесс в «World of Tanks» основывается на сражении двух случайно подобранных команд по 15 игроков (режим рандом). Условие победы в битве — полное уничтожение команды противника либо захват его базы. Максимальная продолжительность боя 10–15 минут (в зависимости от режима).

Общение между игроками во время боев и координация действий осуществляются через текстовый чат, либо голосом (между игроками в составе взвода, роты, либо в тренировочных комнатах).

Помимо индивидуально-рандомного участия в бою возможен также вход в бой взводами, состоящими из двух или трех знакомых между собой игроков. Благодаря слаженным действиям участников, взвод способен оказать большое влияние на исход боя. В режиме ротных боев команда (рота) создается командиром, а все игроки соединены внутриигровой голосовой связью. Набор бойцов в команду осуществляется посредством балльной системы. Каждая команда после боя получает опыт и кредиты в зависимости от нанесенного ею урона и других игровых достижений. Кроме того, победившая команда помимо заработанного опыта и кредитов получает довольно значимую долю заработанных средств врага. Также в игре поддерживаются кланы игроков и межклановые бои за территории на глобальной карте (режим «Мировая война»). Игрок, состоящий в клане, получает доступ к отдельному внутриигровому каналу текстового общения между членами клана. Для режима «Мировая война» создана глобальная карта, разделённая на небольшие зоны влияния, за контроль над которыми и происходят сражения между кланами. Чем больше территорий будет находиться под контролем клана, тем больше игровых преимуществ он получает. За владение территорией в казну клана поступает внутриигровая валюта — «золото», обычно покупаемая за реальные деньги.

Игровая валюта в «World of Tanks» представлена в виде игровых «кредитов» и «золота». «Кредиты» начисляются за конкретные достижения игроков в бою, причем вознаграждаются как победа команды в целом, так и индивидуальные достижения игроков. «Золото» может быть приобретено за реальные деньги. Некоторые возможности в игре становятся доступными только за игровое «золото». Участвуя в сражениях, игрок зарабатывает «кредиты», тренирует экипаж и накапливает очки опыта для получения возможности приобретать новые узлы и танки. Имеющаяся в игре рейтинговая система отображает статистику побед и поражений, а также фиксирует достижения отдельного игрока.

## 1.2 Набор данных

Для проведения данного исследования была сформирована коллекция данных, содержащая информацию о 104900 кланах и 1032284 участниках кланов. Всего в коллекции доступна информация о порядка 9 миллионах игроках русскоязычного сервера. Для каждого игрока собрана игровая статистика (16 показателей) и информация о его достижениях — медалях, знаках отличия и знаках классности (73 показателя).

# 2 Исследование

## 2.1 Анализ данных из профилей игроков

Для проведения исследования были выбраны (рассчитаны) следующие показатели из профилей игроков:

1) *LClan* — метка членства игрока в клане;

2) *battles* — общее количество проведенных боев игрока;

3) *max_xp* — максимальное количество опыта, полученное игроком за один бой;

4) *xp* — общее количество опыта, полученное игроком за все проведенные бои;

5) *battle_avg_xp* — количество опыта, полученное игроком в среднем за один бой;

6) *hits_percents* — средний процент попаданий игрока во вражескую технику за бой;

7) *AVGMLvl* — среднее значение знака классности (по всем танкам игрока);

8) *NTanks* — общее количество танков в ангаре у игрока;

9) *AVGLvl* — средний уровень танка в ангаре у игрока (от 1 до 10);

10) *crdate* — дата создания аккаунта игрока;

11) *upddate* — дата последнего обновления аккаунта игрока.

В пакете RapidMiner была рассчитана корреляционная матрица, представленная в таблице 1.

Развитие игрока в рамках данной игры заключается в расширении его игровых возможностей и повышении своего персонального (либо группового) рейтинга. Для решения обеих задач игроку необходимо совершенствовать умения и навыки своих танковых экипажей, повышать тактико-технические возможности своих танков путем модернизации их оборудования, установки на них дополнительных модулей и снаряжения. Кроме того, очевидно, должны развиваться игровые навыки и опыт самого игрока.

*Таблица 1.* Корреляционная матрица для атрибутов профилей игроков WOT

| | Attributes | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1)** | **LClan** | **1,00** | 0,43 | 0,44 | 0,43 | 0,45 | 0,35 | 0,36 | 0,45 | 0,43 | -0,07 | 0,39 |
| **2)** | **battles** | 0,43 | **1,00** | 0,78 | 0,96 | 0,79 | 0,64 | 0,62 | 0,86 | 0,81 | -0,20 | 0,67 |
| **3)** | **max_xp** | 0,44 | 0,78 | **1,00** | 0,73 | 0,94 | 0,84 | 0,78 | 0,86 | 0,95 | -0,17 | 0,73 |
| **4)** | **xp** | 0,43 | 0,96 | 0,73 | **1,00** | 0,79 | 0,58 | 0,56 | 0,80 | 0,76 | -0,21 | 0,59 |
| **5)** | **battle_avg_xp** | 0,45 | 0,79 | 0,94 | 0,79 | **1,00** | 0,83 | 0,74 | 0,82 | 0,93 | -0,20 | 0,69 |
| **6)** | **hits_percents** | 0,35 | 0,64 | 0,84 | 0,58 | 0,83 | **1,00** | 0,72 | 0,71 | 0,84 | -0,12 | 0,62 |
| **7)** | **AVGMLvl** | 0,36 | 0,62 | 0,78 | 0,56 | 0,74 | 0,72 | **1,00** | 0,71 | 0,77 | 0,17 | 0,68 |
| **8)** | **NTanks** | 0,45 | 0,86 | 0,86 | 0,80 | 0,82 | 0,71 | 0,71 | **1,00** | 0,86 | -0,14 | 0,74 |
| **9)** | **AVGLvl** | 0,43 | 0,81 | 0,95 | 0,76 | 0,93 | 0,84 | 0,77 | 0,86 | **1,00** | -0,19 | 0,73 |
| **10)** | **crdate** | -0,07 | -0,20 | -0,17 | -0,21 | -0,20 | -0,12 | 0,17 | -0,14 | -0,19 | **1,00** | -0,03 |
| **11)** | **upddate** | 0,39 | 0,67 | 0,73 | 0,59 | 0,69 | 0,62 | 0,68 | 0,74 | 0,73 | -0,03 | **1,00** |

Развитие танковых экипажей и повышение тактико-технических характеристик танков требует больших затрат очков опыта и «кредитов». По итогам боя игроку автоматически начисляются виртуальные очки опыта и «кредиты», которые игрок может тратить по своему усмотрению. Также особые достижения игрока в бою отмечаются в виде «неэкономических» знаков — медалей, знаков отличия и знаков классности.

Как следует из таблицы 1, количество очков опыта (показатели *xp*, *max_xp* и *battle_avg_xp*) в существенной степени коррелирует с общим количеством боев (показатель *battles*), проведенных игроком. Косвенное влияние показатель *battles* также оказывает на общее количество танков в ангаре — *NTanks* и их средний уровень *AVGLvl* (в интервале от 1 до 10). Чем выше уровень танка, тем больше очков опыта и «кредитов» может заработать игрок за один бой.

Таким образом, первое очевидное направление «экономической стратегии» игрока — это увеличение общего количества проведенных индивидуально-рандомных боев.

На рисунке 1 показаны гистограммы распределения значений некоторых из показателей профиля игроков. Представлены значения как для игроков, являющихся членами игровых кланов WOT (более темным оттенком), так и не являющихся таковыми (более светлым оттенком). По оси частот была использована логарифмическая шкала. На гистограмме включен режим прозрачности.

При сравнении гистограмм 1a) и 1b) можно увидеть, что характер распределения общего количества опыта *xp* и количества проведенных боев *battles* у игроков очень похож. Для других показателей скорость изменения частоты в распределении заметно отличается.

На рисунке 2 также представлен пример диаграммы рассеивания, построенной в трехмерном пространстве атрибутов профилей игроков. В качестве координат *X* и *Y* были выбраны атрибуты *battles* и *xp*. Выбор третьей координаты принципиально не изменяет вид диаграммы ввиду высокой степени корреляции других атрибутов с атрибутом *battles*. Визуальный анализ этой диаграммы а также гистограмм на рисунке 1 не позволяет выделить четко локализованные кластеры игроков (по крайне мере в области начальных и средних значений атрибутов *battles* и *xp*).

Другая альтернатива «экономической стратегии» игрока в WOT — это использование различных форм координации с другими игроками своей команды во время боя для достижения большей слаженности действий команды. В принципе, в режиме рандомного боя все игроки имеют возможность размещать текстовые сообщения в общем боевом чате, но изначально случайный подбор игроков в команде и общедоступность сообщений в чате (в том числе для игроков команды противника) затрудняют эффективное использование чата.

Заметно большая слаженность достигается в рамках таких боевых подразделений как взвод и рота. Для коммуникации в бою между игроками

таких подразделений используется голосовой чат, защищенный от других игроков (как союзников, так и противников). Умелое использование этой возможности (конечно, при наличии заинтересованности в кооперации усилий в бою) позволяет игрокам этих подразделений существенно повлиять на исход боя, повысить свои достижения и в конечном итоге заработать больше очков опыта и «кредитов». Таким образом, WOT стимулирует игроков к развитию умения работать в команде.

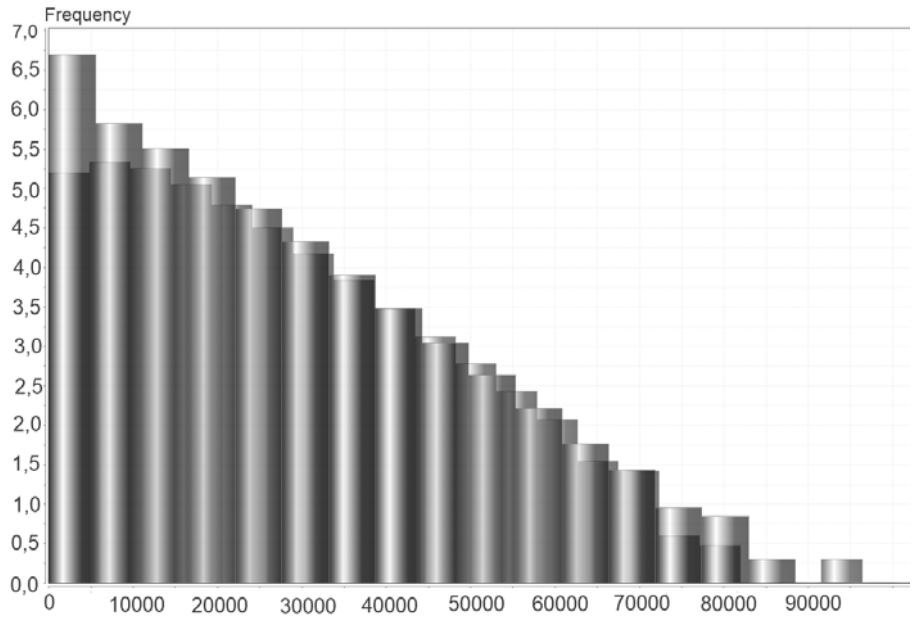
Структура взвода (2-3 человека) не предполагает наличия формальной роли командира. Такая роль появляется в ротных боях. Сама рота создается командиром и набор бойцов в команду осуществляется посредством балльной системы. Победившая в бою команда помимо собственного заработанного опыта и «кредитов» получает значимую долю заработанных средств врага, в результате чего выплаты за выигрыш довольно значительны, при этом проигравшая команда уходит в большой минус.

Поскольку конкретный взвод в отличие от клана не является устойчиво существующей группой в рамках WOT, то для определения игроков, имеющих опыт боев в составе взвода был использован атрибут профиля *brothers_in_arm*, отражающий особые боевые достижения игрока в бою в составе взвода.

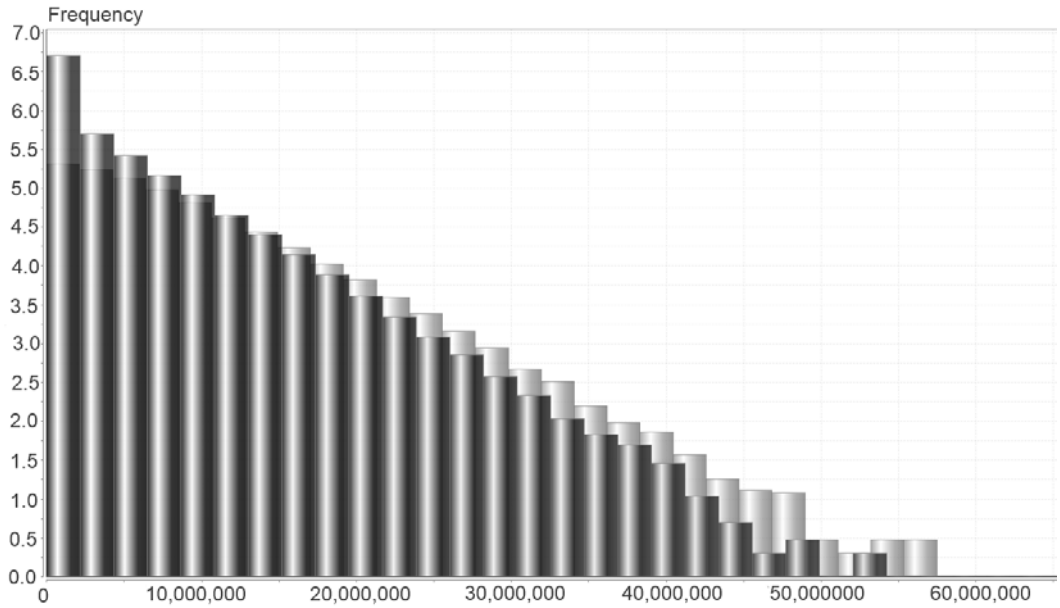
На гистограммах на рисунке 1 в режиме наложения представлены распределения атрибутов для игроков состоящих (темный оттенок) и не состоящих (светлый оттенок) в кланах.

Видно, что по некоторым из атрибутов распределение довольно заметно отличается в области низких значений для членов кланов в сравнении с игроками, не состоящими в кланах.

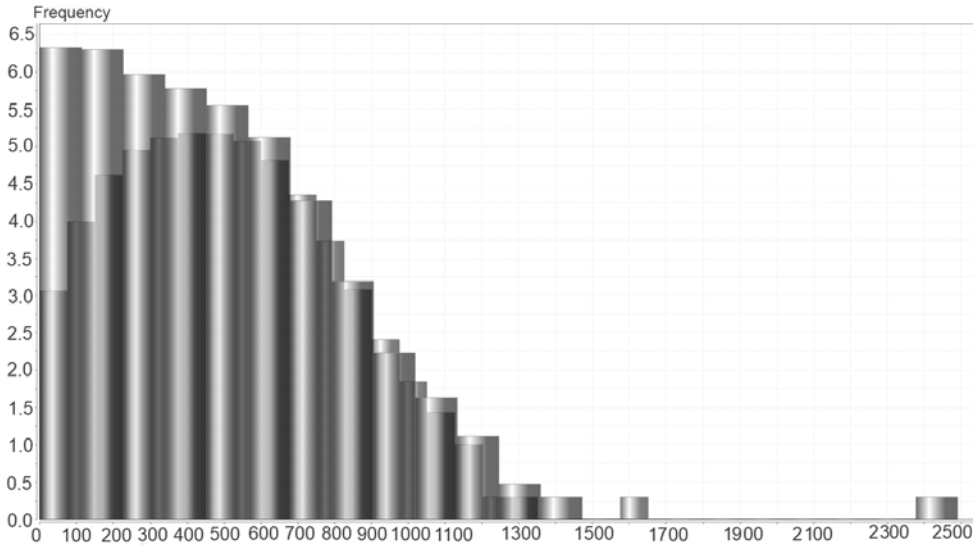
a)



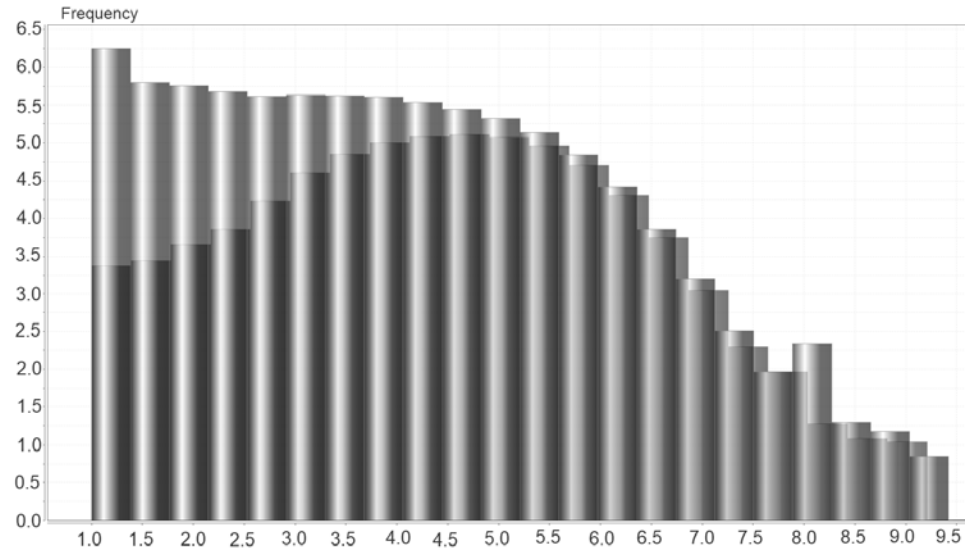

b)



c)




d)




e)

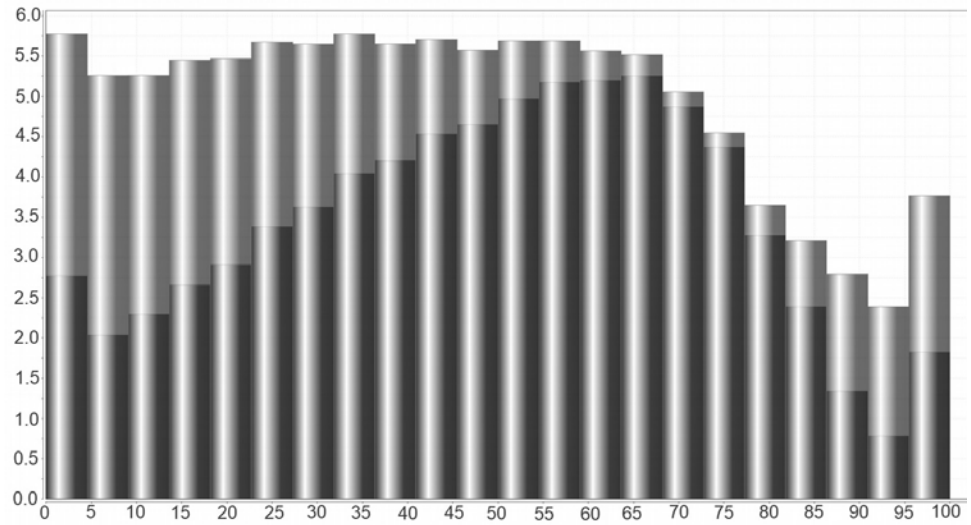



*Рисунок 1*. Гистограмма распределения значений показателей профиля игрока для атрибутов:
a) *battles*; b) *xp*; c) *battle_avg_xp*; d) *AVGLvl*; e) *hits_percents.*

Более светлый уровень шкалы серого цвета (на переднем плане) отражает частоту для клановых игроков. На гистограммах включен режим прозрачности

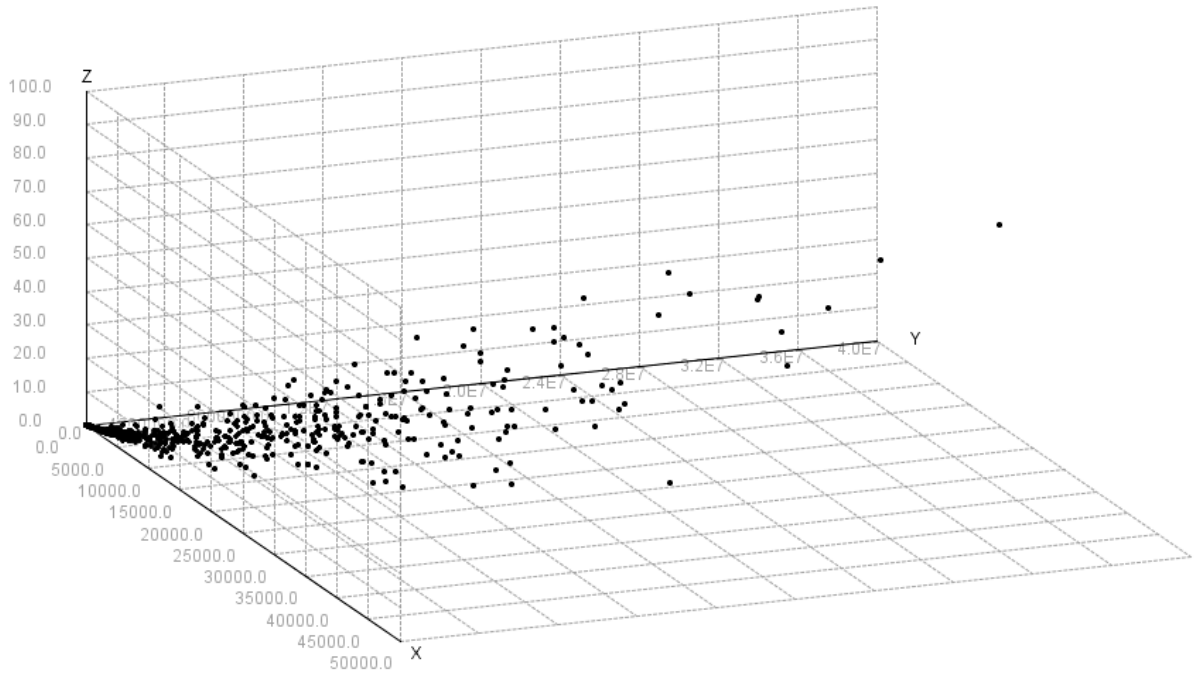



*Рисунок 2*. 3D диаграмма рассеивания для профилей игроков.
Для лучшей наглядности использовалась ограниченная случайная выборка из полной коллекции профилей

Данный вывод также подтверждается диаграммой рассеяния на рисунке 3. Соответствующий уровень шкалы серого цвета у точки на гистограмме показывает участие игрока во взводе (атрибут *brothers_in_arm* фиксирует достижения игрока в составе взвода). Ось *X* соответствует признаку членства игрока в клане ('0' или '1').

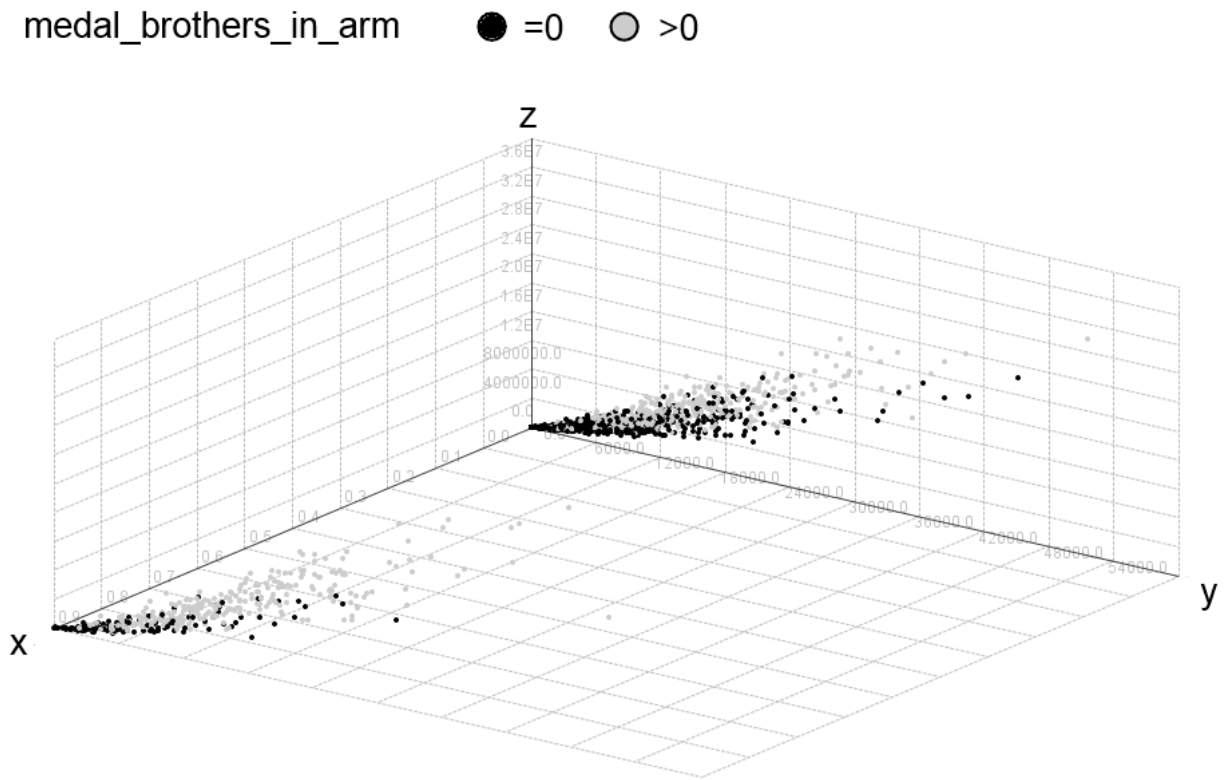



*Рисунок 3.* 3D диаграмма рассеивания для профилей игроков

Количество профилей игроков, имеющих ненулевое значение атрибута *medal_brothers_in_arms* составило около 16% от всех имеющихся в коллекции данных, а доля игроков, состоящих в кланах составила около 9%. Оценка корреляции между двумя данными атрибутами имеет величину примерно 0.27. При этом доля взводных игроков заметно выше доли невзводных среди членов кланов, в то время как для неклановых игроков доля взводных в разы меньше. Гистограммы распределений значений атрибутов, отражающих игровые достижения у взводных и клановых игроков, помимо того, что отличаются формой, имеют также заметный сдвиг по оси атрибутов в сторону больших значений (как, например на рисунках 1c — 1e). Причем у взводных игроков этот сдвиг выражен в большей степени.

С помощью пакета RapidMiner был построен бинарный классификатор в форме дерева принятия решений для профилей игроков, определяющий принадлежность игрока к клану на основании значений других его атрибутов. Классификатор представлен на рисунке 4. В качестве критерия расщепления был выбран *gini_index*.

Видно, что вся совокупность профилей по атрибуту *max_xp* делится на два подмножества в соотношении 1 : 2. Игроки, имеющие значение атрибута *max_xp* (максимальный опыт за один бой) менее 1870.5, образуют достаточно однородную группу неклановых бойцов (ошибка составляет менее 1%). То есть данное значение *max_xp* можно условно считать пороговым для возможности вступления игрока в какой-либо из кланов. Профили игроков, попавшие в левую ветку, поддаются классификации значительно хуже и требуют учета большего числа признаков.

### 2.2 Кланы на глобальной карте

Глобальная карта — это виртуальное поле боя, за основу которого взята карта мира. Для удобства карта разделена на несколько частей (например, Северная Европа, Западная Африка и т.п.). Каждая из этих частей, в свою очередь, делится на провинции. Для того чтобы воевать на Глобальной карте, игроки объединяются в кланы.

Сражаясь на Глобальной карте, кланы преследуют «экономические» цели: каждая провинция ежедневно приносит своему обладателю определенную сумму игрового золота, которая может варьироваться от нескольких десятков до нескольких тысяч, в зависимости от географического положения и исторической ценности провинции.

Весь процесс «мировой войны» делится на два уровня: глобальный (то, что происходит непосредственно на Глобальной карте) и тактический (то, что происходит в клиенте игры). Непосредственные сражения рот из кланов проходят в клиенте игры. По-сути, клановый бой почти не отличается от ротного боя абсолютного формата.

На рисунке 5 показана динамика создания игровых кланов по месяцам и их характеристик — суммарного числа игроков в созданных за месяц кланах и среднего за месяц размера клана. Рассматривался период с декабря 2010 по январь 2014 года. На рисунке заметная сезонная составляющая: пики — в начале года и спады — в середине года.

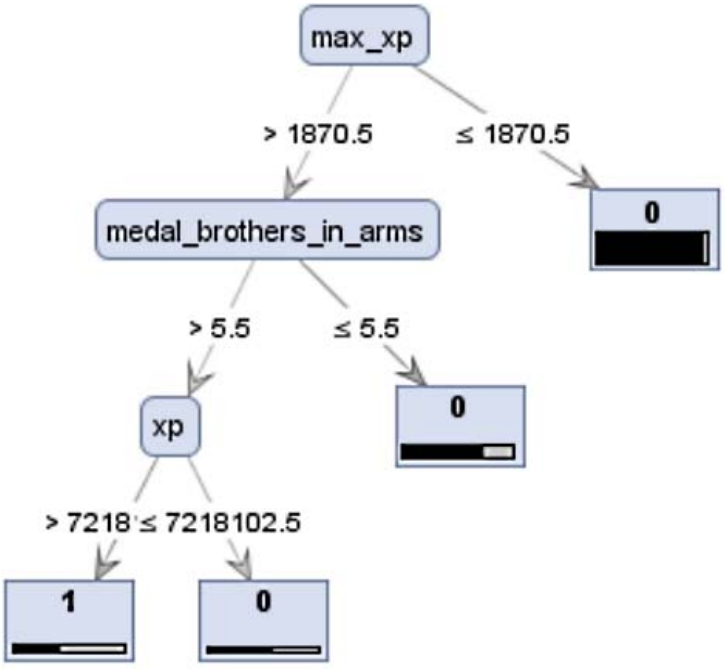



*Рисунок 4.* Дерево принятия решений бинарного классификатора игроков

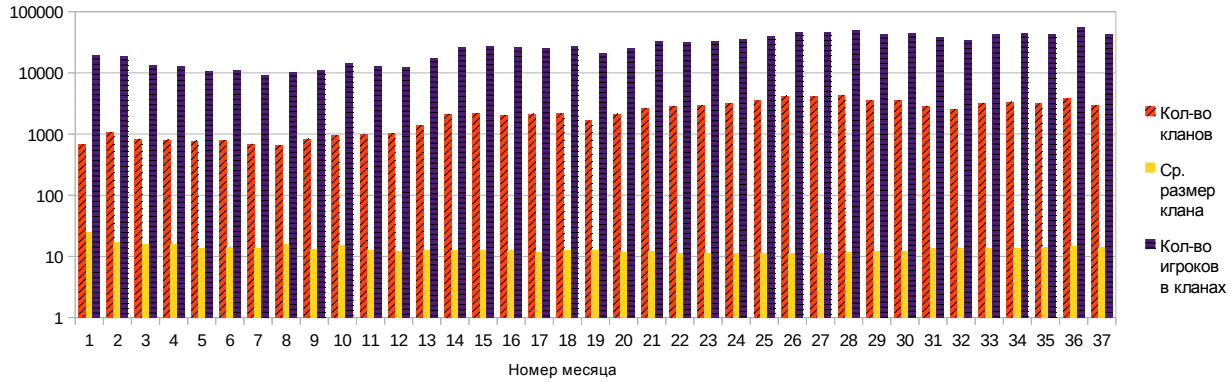



*Рисунок 5*. Динамика создания кланов по месяцам

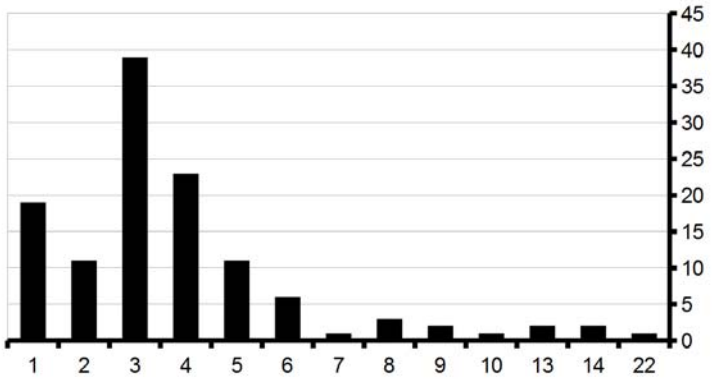



*Рисунок 6*. Гистограмма распределения числа провинций по кланам

На глобальной карте 479 провинций распределились между 122 кланами. Гистограмма распределения числа провинций представлена на рисунке 6. Вертикальная ось соответствует количеству кланов. На гистограмме заметна сильная неравномерность в распределении провинций.

Таким образом, из почти 105 тысяч представленных в коллекции данных кланов, ненулевое значения атрибута *victory_points* имеют всего 985 кланов, а провинциями на Глобальной карте владеют всего лишь 122 из них, что составляет примерно 0.9% и 0.1% соответственно. В кланах, владеющих провинциями, состоит примерно 11 тыс. игроков. Данные цифры показывают высокий уровень конкуренции между кланами на Глобальной карте.

Для исследования кланов были выбраны (рассчитаны) следующие атрибуты:

1) *members_count* — количество игроков в клане;

2) *victory_points* — очки победы клана;

3) *provinces_count* — количество провинций, которыми владеет клан;

4) *combats_count* — количество сражений, проведенных кланом на глобальной карте;

5) *Income* — общий доход от провинций, которыми владеет клан;

6) *WinRate* — процент побед клана на глобальной карте;

7) *AvgGlobRating* — средний глобальный рейтинг игроков клана;

8) *AvgXPAll* — средний общий опыт у игроков клана (по всем видам боев);

9) *AvgXPClan* — средний общий опыт у игроков клана в клановых боях;

10) *AvgBatAll* — среднее количество боев (всех видов) у игроков клана;

11) *AvgBatClan* — среднее количество клановых боев у игроков клана;

12) *AvgWinRateAll* — средний процент побед у игроков клана (по всем видам боев);

13) *created_at* — время создания клана.

Для анализа связи между этими атрибутами была рассчитана корреляционная матрица, приведенная в таблице 2.

Эта матрица подтверждает прямую связь между атрибутами *victory_points*, *provinces_count*, *Income* и *WinRate*.

*Таблица 2*. Корреляционная матрица для атрибутов кланов с Глобальной карты WOT

| Атрибут | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1) members_count** | 1 | 0,11 | 0,01 | 0,06 | 0,06 | 0,13 | -0,05 | 0,26 | 0,17 | 0,33 | 0,23 | 0,36 | 0,06 | 0,02 |
| **2) victory_points** | 0,11 | 1 | 0,22 | 0,77 | 0,07 | 0,65 | 0,61 | 0,46 | 0,46 | 0,28 | 0,27 | 0,17 | 0,42 | 0,00 |
| **3) wins_count** | 0,01 | 0,22 | 1 | 0,01 | 0,96 | 0,07 | 0,34 | 0,43 | 0,48 | 0,40 | 0,39 | 0,32 | 0,35 | 0,14 |
| **4) provinces_count** | 0,06 | 0,77 | 0,01 | 1 | -0,12 | 0,73 | 0,49 | 0,26 | 0,20 | 0,10 | 0,06 | 0,03 | 0,23 | 0,02 |
| **5) combats_count** | 0,06 | 0,07 | 0,96 | -0,12 | 1 | -0,04 | 0,11 | 0,33 | 0,40 | 0,39 | 0,36 | 0,34 | 0,27 | 0,13 |
| **6) Income** | 0,13 | 0,65 | 0,07 | 0,73 | -0,04 | 1 | 0,43 | 0,33 | 0,32 | 0,26 | 0,19 | 0,18 | 0,27 | -0,06 |
| **7) WinRate** | -0,05 | 0,61 | 0,34 | 0,49 | 0,11 | 0,43 | 1 | 0,51 | 0,47 | 0,21 | 0,25 | 0,08 | 0,42 | 0,11 |
| **8) AvgGlobRating** | 0,26 | 0,46 | 0,43 | 0,26 | 0,33 | 0,33 | 0,51 | 1 | 0,85 | 0,45 | 0,68 | 0,30 | 0,92 | -0,02 |
| **9) AvgXPAll** | 0,17 | 0,46 | 0,48 | 0,20 | 0,40 | 0,32 | 0,47 | 0,85 | 1 | 0,68 | 0,91 | 0,56 | 0,73 | 0,03 |
| **10) AvgXPClan** | 0,33 | 0,28 | 0,40 | 0,10 | 0,39 | 0,26 | 0,21 | 0,45 | 0,68 | 1 | 0,73 | 0,97 | 0,30 | 0,04 |
| **11) AvgBatAll** | 0,23 | 0,27 | 0,39 | 0,06 | 0,36 | 0,19 | 0,25 | 0,68 | 0,91 | 0,73 | 1 | 0,69 | 0,55 | 0,03 |
| **12) AvgBatClan** | 0,36 | 0,17 | 0,32 | 0,03 | 0,34 | 0,18 | 0,08 | 0,30 | 0,56 | 0,97 | 0,69 | 1 | 0,17 | 0,03 |
| **13) AvgWinRateAll** | 0,06 | 0,42 | 0,35 | 0,23 | 0,27 | 0,27 | 0,42 | 0,92 | 0,73 | 0,30 | 0,55 | 0,17 | 1 | -0,09 |
| **14) created_at** | 0,02 | 0,00 | 0,14 | 0,02 | 0,13 | -0,06 | 0,11 | -0,02 | 0,03 | 0,04 | 0,03 | 0,03 | -0,09 | 1 |

На диаграмме рассеивания (рисунок 7), построенной в пространстве атрибутов *victory_points* (*X*), *provinces_count* (*Y*), *Income* (*Z*), видна принципиальная неоднорость в распределении кланов, владеющих регионами на Глобальной карте.

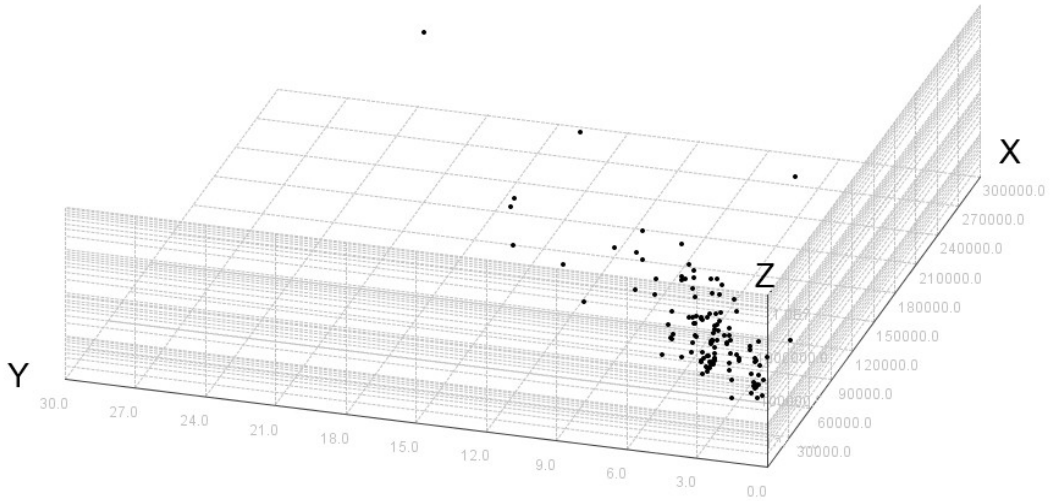



*Рисунок 7.* 3D-диаграмма рассеивания для кланов с Глобальной карты

## Заключение

Наличие виртуальной «экономической» составляющей в игре WOT стимулирует игроков к установлению связей друг с другом и формированию сообществ в форме взводов, рот и кланов с целью кооперации их усилий для достижения наилучших игровых результатов. Наличие высоких игровых результатов у отдельных пользователей также способствует росту их персонального рейтинга, который может «монетизироваться» как в рамках самой игры, так и за пределами непосредственно игрового процесса.

На приведенных в работе гистограммах и диаграммах рассеяния отчетливо присутствует неоднородность в распределении характеристик профилей игроков и кланов, что в общем соответствует выводам, полученным в других работах.

Результаты исследования могут быть практически использованы для мониторинга за развитием игрового сообщества WOT разработчиками игры. Путем анализа аномалий в распределениях числовых значений атрибутов профилей можно выявлять игровых ботов, которые заметно ухудшают качество игры для реальных игроков.

Также модели процессов образования и развития «социальных» структур в виртуальном мире ММОИ могут способствовать лучшему пониманию социальных и экономических процессов в реальном мире.

В данной работе представлены лишь некоторые из результатов исследования. Огромный объем и

многомерность доступных данных открывают широчайшее поле для исследований. Ограниченный объем тезисов позволяет представить лишь некоторые из полученных автором результатов.

Следует также отметить, что набор данных, предоставляемых через Wargaming.net Public API, не содержит информации о социально-сетевых аспектах игры WOT, например, таких как списки друзей в профилях игроков, информацию об учетных записях игроков в официальном форуме WOT и их участии в тематических разделах форума. На основе данной информации возможно было бы дополнительно провести исследование социально-сетевых связей игроков.

## Литература


[1] MMOG: Massively Multiplayer Online Game (MMO). http://www.gamedev.ru/gamedesign/terms/MMOG

[2] M. Szell, S. Thurner. Measuring social dynamics in a massive multiplayer online game // Social Networks. — 2010. — Vol. 32, Iss. 4. — P. 313–329.

[3] S. Son, A. R. Kang, H.-c. Kim, T. Kwon, J. Park et al. Analysis of Context Dependence in Social Interaction Networks of a Massively Multiplayer Online Role-Playing Game. — 2012. PLoS ONE 7(4): e33918. doi:10.1371/journal.pone.0033918

[4] B. Fuchs, S. Thurner. Behavioral and Network Origins of Wealth Inequality: Insights from a Virtual World. — 2014. arXiv preprint. http://arxiv.org/abs/1403.6342

[5] B. Corominas-Murtra, B. Fuchs, S. Thurner. Detection of the elite structure in a virtual multiplex social system by means of a generalized K-core. — arXiv:1309.6740. http://arxiv.org/abs/1309.6740

[6] C. Shen. Network patterns and social architecture in Massively Multiplayer Online Games: Mapping the social world of EverQuest II // New Media & Society. — 2014. — 16(4). — P. 672–691.

[7] Кабинет разработчика — единая точка доступа к пользовательским данным Wargaming.net. https://ru.wargaming.net/developers/


## Study on the Structure and Dynamics of the Gamer Community in the Massively Multiplayer Online Game World of Tanks


Alexander V. Sychev



The results on the analysis of the profiles data for players and clans in the massively multiplayer game World of Tanks is presented. The structure of players and clans community, and the factors determining their efficiency in the game are considered.

# Инфраструктура электронного научного журнала и облачные сервисы поддержки жизненного цикла электронных публикаций

© А.М. Елизаров © Д.С. Зуев © Е.К. Липачёв
Институт математики и механики им. Н.И. Лобачевского
Казанского (Приволжского) федерального университета
amelizarov@gmail.com dzuev11@gmail.com elipachev@gmail.com

**Аннотация**

Представлены облачные сервисы, необходимые для поддержки жизненного цикла электронных научных публикаций в информационных системах управления электронными научными журналами. Приведены примеры расширения функционала базовых сервисов этих систем путем разработки дополнительных модулей. Исследование основано на анализе открытых информационных систем, проведенном с учетом особенностей процесса электронного книгоиздания, и опыте авторов по созданию программной платформы управления научными журналами.

*Работа поддержана РФФИ (проекты № 12-07-00667 и № 12-07-97018-р_поволжье).*

## 1 Введение

Развитие ИКТ в целом и веб-технологий в частности послужило стимулом для переориентации всех типов коммуникации, в том числе научной, в сторону виртуализации. Как известно, в информационном обществе ценность информации и нематериальных ресурсов становится все более ощутимой и подчас существенно более весомой, чем материальных активов.

Проникновение интернета во все сферы жизнедеятельности человека влияет на его поведение в целом, а также на качество и скорость выполнения научных исследований. Для большинства пользователей доступ в интернет стал каждодневной потребностью, а возможность получения разносторонней информации – необходимой для формирования достоверной картины мира.

Узконаправленные информационные порталы на текущий момент времени мало жизнеспособны – все крупные игроки рынка информационных услуг так или иначе расширяют спектр сервисов, предоставляемых пользователю. Например, Google сегодня является не только поисковой машиной, но предоставляет ряд дополнительных услуг, включая почту, социальную сеть, сервисы совместной работы, и даже имеет собственную операционную систему. То же самое можно сказать и про лидеров российского информационного рынка – компании Яндекс, Mail.Ru и др. Соответственно мы наблюдаем тенденцию расширения спектра предоставляемых сервисов, а сами сайты, нацеленные на какую-то одну тематику, становятся универсальными веб-порталами.

Названная тенденция выгодна как поставщикам услуг, так и самим пользователям. Для поставщика услуг владение широкой пользовательской базой является существенным преимуществом на рынке и помогает выжить в конкурентной среде. Пользователю, несомненно, удобнее получать все интересующие его услуги в одном месте.

Описанная ситуация характерна и для научного сообщества: ведущие мировые научные библиотеки сегодня осваивают новые функции, связанные с оцифровкой бумажного фонда и хранением электронной информации, интеграцией электронных ресурсов и обеспечением эффективной навигации в них; участвуют в формировании системы научной коммуникации и, используя сетевую инфраструктуру, налаживают новую систему сервисов интеграции научной информации (см., например, [1, 2]). Одновременно происходит активное формирование электронных библиотек (ЭБ) – распределенных информационных систем, позволяющих надежно сохранять и эффективно использовать коллекции электронных документов через глобальные сети передачи данных. ЭБ создаются в университетах и исследовательских организациях и часто являются междисциплинарными проектами. Появление новых электронных библиотек, увеличение числа хранимых в них документов, расширение набора и повышение качества предоставляемых ими сервисов способствуют развитию науки, облегчая ученым доступ

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

к источникам информации, а также предоставляя им эффективное средство распространения научных результатов и взаимодействия на основе сетевых коммуникаций.

Помимо создания ЭБ осуществляется перевод процессов издания научных журналов в электронную форму, создан целый ряд информационных систем, автоматизирующих соответствующие процессы (их обзор приведен в [3, 4]). Осуществляется переход от традиционного издательского процесса к электронному не только на этапах верстки выпусков журналов и публикации научных статей, но и на этапах их рецензирования. Перевод редакционных процессов в электронную форму и размещение журналов во Сети нацелены не только на облегчение/удешевление работ по изданию научных журналов, но и на расширение целевой аудитории, повышение доступности журналов для научного сообщества. Поэтому современный электронный научный журнал помимо собственно публикаций должен предоставлять с помощью своего веб-портала ряд дополнительных сервисов.

Цели настоящей работы – выявление и анализ дополнительных сервисов, которые должны предоставлять пользователям информационные системы электронных научных журналов. Исследование основано на анализе открытых журнальных систем, проведенном с учетом особенностей процесса издания отечественных научных журналов [4], и нашем опыте создания платформы научных журналов [5].

## 2 Программные платформы управления электронными научными журналами

Практически все ведущие научные издательства внедряют системы автоматического управления рабочими процессами, в числе которых – наиболее сложные и длительные по времени редакционные процессы, обеспечивающие независимое научное рецензирование. Как отмечено, например, в [6], создаваемые платформы управления электронными научными журналами пока реализуют стандартные процессы издания журналов и соответствующие стандартные алгоритмы работы. Одновременно ведется работа по автоматизации основных рабочих процессов, реализуемых редколлегиями научных журналов, в частности, на основе технологий Cloud computing [5, 7].

Для поддержки жизненного цикла как отдельных научных статей, так и научных журналов в целом целесообразно использовать в качестве ядра системы управления электронными научными журналами программную платформу Open Journal Systems – OJS (см. [3, 4]). В [4] предложена архитектура универсальной платформы управления электронными научными журналами, которая содержит три уровня – физический, базовый и уровень сервисов.

Физический уровень характеризует аппаратную составляющую системы, обеспечивающую функционирование верхних уровней, и содержит системное и прикладное программное обеспечение. Эти компоненты предполагают техническую поддержку с использованием технологий виртуализации и облачных вычислений.

Базовый уровень реализует основные сервисы управления электронными научными журналами, в том числе, регистрацию авторов и пользователей, прием и первичную обработку статей, включая автоматизацию проверки соблюдения правил редакции и рецензирования, контроль соблюдения сроков рассмотрения статей, назначение рецензентов и рассылку уведомлений. Базовый уровень включает также сервисы удаленного взаимодействия и совместной работы, поиска в электронном хранилище и автоматического извлечения метаданных, структурирования входящей информации, управления пользователями и ролями, платного доступа к контенту.

На уровне сервисов размещены дополнительные надстройки и функции, учитывающие специфику предметной области научного журнала. Например, для математических журналов востребованы сервисы конвертации в специализированные форматы ($T_EX$, MathML и др.). Здесь реализуется front-end системы и происходит взаимодействие с конечным пользователем.

Взаимодействие с системой управления электронными научными журналами может быть организовано либо через собственный веб-портал, либо через специальные программные адаптеры с сайта конкретного журнала, размещающего свой контент в хранилище системы.

При первом способе взаимодействия зарегистрированный пользователь получает доступ ко всем журналам, размещенным в системе, а веб-портал служит единой точкой входа. Такой способ наиболее удобен для новых журналов, не имевших собственных сайтов в Сети.

Для журналов, уже имеющих историю и поддерживающих собственные сайты, более приемлемым, на наш взгляд, является второй способ взаимодействия. В частности, это позволяет сохранить привычный адрес сайта журнала и его «историю» в интернете, при этом максимально автоматизировав редакционные процессы.

Внедрение информационной системы управления бизнес-процессами научного журнала позволяет, прежде всего, автоматизировать наиболее трудоемкие рабочие процессы, а портальное решение дает возможность интегрировать журнал в мировое информационное научное пространство. Например, согласно статистике проекта PKP (https://pkp.sfu.ca/ojs/ojs-usage/ojs-stats/) – разработчика OJS, в 2013 году система OJS использовалась более чем в 6800 активных журналах, и это количество постоянно росло с момента начала ее внедрения (при этом учитывались только журналы, опубликовавшие не менее 10 статей в течение года).

Вместе с тем, для развития электронного научного журнала необходимо расширение функциональности его базовой информационной системы, что связано как с особенностями предметной облас-

ти этого журнала, так и со сложившимися традициями работы его редакции и редколлегии, что достигается разработкой специализированных модулей. Например, для журналов физико-математического направления необходима программная поддержка процесса обработки электронных документов, созданных в TEX-нотации. Для системы OJS такая поддержка реализована в виде специализированного плагина [8].

## 3 Базовый набор сервисов программной платформы управления электронным научным журналом

Функциональность современных информационных систем управления научными журналами должна содержать ряд обязательных и опциональных сервисов.

К обязательным можно отнести функции, регулирующие процесс рецензирования и обеспечивающие коллективное редактирование электронных документов. Также важны такие редакционные сервисы, как классификация, выделение метаданных, публикация, долгосрочное хранение, конвертирование в различные форматы и распространение, статистика использования, объединение в коллекцию, контроль доступа, подписка, рассылка уведомлений.

Вместе с тем, современные информационные системы управления электронными научными публикациями не ограничиваются сервисами удаленного представления статей в научный журнал и их дальнейшей обработки для окончательной публикации, а обеспечивают доступ к сформированному контенту и расширенный поиск (по автору, названию статьи, ключевым словам и др.) в соответствующих электронных коллекциях, т. е. в полном объеме реализуют функциональные возможности, присущие электронным библиотекам.

Все вышеуказанные сервисы фактически являются сервисами, присущими любой информационной системе управления журналом, и их реализация на портале журнала безусловно необходима, однако вовсе не может быть достаточной для устойчивого развития издания в современном информационном обществе.

К дополнительным функциям, расширяющим спектр предоставляемых услуг электронного журнала, можно отнести следующие функции:

- автоматизация формирования коллекций документов и конвертации статей: выделение метаданных, автоматическая разметка статей; примерами служат методы формирования математических электронных коллекций [9];
- учет специфики обрабатываемых информационных ресурсов, например, расширенный поиск, в частности, по фрагментам формул в математических коллекциях [9];
- оплата услуг, например, OJS имеет возможность работы с электронным кошельком PayPal (www.paypal.com);
- информетрический анализ, например, в системе OJS реализована поддержка сервиса Article-Level Metrics (http://article-level-metrics.plos.org).
- поддержка научных конференций, например, система www.easychair.org и система автоматизации конференций Open Conference System, созданная в рамках Public Knowledge Project;
- онлайн-общение (вебинары и видеоконференции, в том числе для распределенных редколлегий);
- поиск и сбор OAI-метаданных, например, система индексирования метаданных Open Harvester Systems
- проверка загружаемых ресурсов на плагиат.

Отметим, что помимо дополнительных функций важным является удобство пользования порталом платформы, в частности, необходима локализация всех сервисов платформы на русский язык. Перечень дополнительных модулей и функций на текущий момент не является законченным – с развитием Сети появляются новые технологии, которые могут быть внедрены в кратчайшее время. Поэтому этот перечень будет постоянно корректироваться.

## Заключение

С 2013 г. в Республике Татарстан на основе архитектуры, описанной выше, создана и развивается система управления электронными научными журналами. На текущий момент времени система реализована технически, создан ее веб-портал (www.science.tatarstan.ru), и ряд научных журналов переведен под ее управление. Система проходит тестирование с целью ее дальнейшей интеграции в единую научно-образовательную среду. Выделен набор дополнительных модулей, функций и сервисов, который реализован на портале. По окончании тестирования будут сделаны выводы о достаточности сформированного перечня функций, их полезности и применимости.

## Литература


[1] Hawkins Kevin S. A model for integrating the publication and preservation of journal articles // CEUR Workshop Proc. Selected Papers of the 15th All-Russian Scientific Conference "Digital libraries: Advanced Methods and Technologies, Digital Collections", Yaroslavl, Russia, October 14–17, 2013. – V. 1108. – P. 112-116. – http://ceur-ws.org/ Vol-1108/paper14.pdf.

[2] Hawkins Kevin S. A model for integrating the publication and preservation of journal articles // Russian Digital Libraries. – 2014. – V. 17, No 2. – http://www.elbib.ru/index.phtml?page=elbib/eng/journal/2014/part2/H.

[3] Elizarov A.M., Zuev D.S., Lipachev E.K. Open Scientific E-journals Management Systems and Digital Libraries Technology // CEUR Workshop

Proc. Selected Papers of the 15th All-Russian Scientific Conference "Digital libraries: Advanced Methods and Technologies, Digital Collections", Yaroslavl, Russia, October 14–17, 2013. – V. 1108. – P. 102–111. – http://ceur-ws.org/Vol-1108/paper13. pdf.

[4] Елизаров А.М., Зуев Д.С., Липачёв Е.К. Информационные системы автоматизации цикла подготовки электронных научных журналов (Electronic Scientific Journal Management Systems) // Научно-техническая информация. Серия 1. – 2014. – № 3. – С. 31–38 (англ. пер.: Scientific and Technical Information Processing, 2014, Vol. 41, No. 1, P. 66–72).

[5] Ахметов Д.Ю., Герасимов А.Н., Грачев А.О., Елизаров А.М., Липачёв Е.К. Облачная платформа поддержки электронных научных изданий // Учёные записки Института социальных и гуманитарных знаний. – 2014. – №1 (12), Ч. 1. – С. 13–19. – http://www.isgz.ru/images/Alexey/Chirko/ek%202014_i.pdf.

[6] Гусев А.Л. Анализ рынка услуг издательских платформ по управлению деятельностью распределенных коллегий электронных изданий // International Scientific Journal for Alternative Energy and Ecology. – 2014. – № 04/1 (123). – С. 82–86.

[7] Ахметов Д.Ю., Елизаров А.М., Липачёв Е.К. Система автоматизации редакционных процессов на платформе электронных научных журналов // Учёные записки Института социальных и гуманитарных знаний. – 2014. – №1 (12), Ч. 2. – С. 228–233. – http://www.isgz.ru/images/Alexey/Chirko/ek%202014_ii.pdf.

[8] Ахметов Д.Ю., Елизаров А.М., Липачёв Е.К. Автоматизация процесса первичной обработки математической статьи в информационной системе электронного научного журнала // Тр. Математического центра имени Н.И. Лобачевского. Материалы Двенадцатой молодежной науч. шк.-конф. «Лобачевские чтения–2013». – Казань: Изд-во Казан. матем. об-ва, 2013. – Т. 47. – С. 6-10.

[9] Biryal’tsev E., Elizarov A., Zhil’tsov N., Lipachev E., Nevzorova O., Solov’ev V. Methods for analyzing semantic data of electronic collections in mathematics // Automatic Documentation and Mathematical Linguistics. – 2014. – V. 48, No 2. – P. 81–85.

## Infrastructure of Electronic Scientific Journal and Cloud Services Supporting Lifecycle of Electronic Publications

Alexander Elizarov, Denis Zuev, Eugene Lipachev

Cloud services required for support of the life cycle of electronic scientific publications in the information systems focused on management of electronic journals are presented. Examples of the functional expansion of basic services of these systems through the development of additional modules are given. The study is based on the analysis of open information systems conducted by taking into account features of the process of electronic publishing as well as on the authors' experience to create a software platform for management of the scientific journals.

# Investigation As a Member of Research Discourse*

© Vasily Bunakov
Scientific Computing Department, Science and Technology Facilities Council,
Harwell OX11 0QX, United Kingdom
vasily.bunakov@stfc.ac.uk

## Abstract

Investigations are specific intellectual entities that circulate in large research facilities with shared access by multiple research teams; investigations have some common features with research papers (publications) and can be included in citation networks. We consider different approaches to modelling the relations between research papers and investigations and discuss opportunities for matching these two members of common research discourse. The analysis undertaken can be of interest for research centres that consider information services based on data and publications contextualization.

## 1 Introduction

The journal articles, e-prints, reports and other similar artefacts that irrespective of their physical manifestation can be seen as derived from their paper-based "document" ancestors are the well-established means of research communication and a popular aide for tracking the state and the trends of research discourse. The "papers" have clear identity, allow review (of different kinds) and participate in citation networks; this supports performing the aforementioned functions of the quality research communication and measurable research tracking; this also makes "papers" valuable intellectual entities worth capturing in library catalogues, and worth sharing via advanced information services.

We suggest that other type of intellectual entities, *investigations*, have essential features similar to the document-like entities hence are the natural candidates to supplement "papers" as valuable members of research discourse. We consider the types of relation between the document-like and investigation entities, and take a look at the simultaneous circulation of them in our own research domain of experimental science utilizing large research facilities: neutron sources, synchrotrons, and powerful lasers shared by multiple researchers (visiting scientists).

**Proceedings of the 16th All-Russian Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" — RCDL-2014, Dubna, Russia, October 13–16, 2014.**

***** This work is related to the projects of PaNdata collaboration www.pan-data.eu supported by the EU 7th Framework Programme for Research and Technological Development. The author would like to thank his colleagues in PaNdata for their input for this paper although the views expressed are the views of the author and not necessarily of the collaboration.

## 2 Facilities research lifecycle and data modelling

### 2.1 Facilities science landscape

Research facilities can be thought of as well-equipped hubs where research teams or individual researchers come to perform their experiments on their own samples. The research facility core is typically represented by a unique scientific instrument: a particle accelerator, a neutron source, a powerful laser, a telescope, or a supercomputer that allows detailed simulation of natural phenomena, or by a few such instruments that offer researchers different research techniques. The examples include European Synchrotron Radiation Facility (www.esrf.eu), neutron source in The Institut Laue-Langevin (www.ill.eu), Siberian Synchrotron and Terahertz Radiation Centre (http://ssrc.inp.nsk.su/CKP/eng/) or the future Extreme Light Infrastructure (www.eli-beams.eu).

Research conducted in facilities bears characteristics of "big science" such as a long-term capital investment, permanent support staff, scalable computing infrastructure; and "bench science" with individual scientists and small research teams that may have specific and short-time research goals. The user community of European facilities counts tens of thousands scientists who pursue different applications: crystallography reveals the structures of proteins important for the development of new drugs; neutron scattering identifies stresses within engineering components such as turbine blades, and tomography can image microscopic details of biological tissues ([1]).

A business model for user research on large facilities that emerged a few decades ago has been influenced by the advances in instrumentation and data analysis that are now more automated and more user friendly than in early days of facilities a few decades ago. This has led, among other effects, to a lesser significance of the instrumentation "gurus" ([2]), and to the emergence of specific services for research and industry that allow users sending their samples for remote investigation according to one of the service plans ([3]).

Yet the facilities business model has proved to be effective and is a foundation for a specific research lifecycle, and for specific information modelling and information services in support of it.

### 2.2 Generic research lifecycle

Despite the variety of facilities instruments and experimental techniques, the following distinct stages are typical across facilities and thus represent a generic facilities lifecycle:

- **Research Proposal:** the facilities are often oversubscribed so the researcher (investigator) should justify the value of her research and the suitability of a particular experimental technique
- **Approval Process:** multilateral assessment by the facility, including risk assessment (as the experiment may involve hazardous materials or techniques)
- **Experiment Scheduling:** allocation of the time slot within a facility operating cycle, and registration of all visitor scientists
- **Series of Experiments (that altogether constitute Investigation with the proclaimed goals):** the user will bring samples, and sometimes an additional equipment to the facility, calibrate the experimental environment and actually take measurements
- **Data Archiving:** facilities offer high-throughput data collection and archiving services; archiving of raw data collected in the facility data storage is often a policy requirement
- **Data Analysis:** it can be done through multi-layer computing environment where some tools are offered by facilities, and others applied by scientist individually
- **Results Publication:** journal articles and alike; facilities often require the visitor scientist to report back on any publications derived from the experiments.

This generic lifecycle is illustrated by Figure 1.

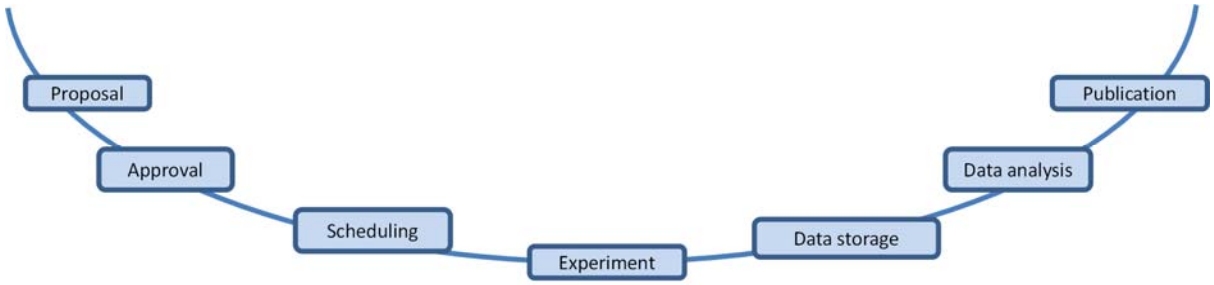



*Figure 1. Facilities research lifecycle*

### 2.3 Data modelling effort so far

Facilities collect raw experimental data in a variety of formats yet there is a movement towards unification best represented by NeXuS standard and community around it (www.nexusformat.org). There are of course data checks and data replication services, as well as some recent attempts to form and curate archival packages according to OAIS reference model ([4]).

The aforementioned generic lifecycle gave birth to the rich CSMD metadata model ([5], [6]) which is implemented, with some modifications, in the popular ICAT software platform ([7]).

Some facilities started assigning persistent identifiers to datasets ([8]) and there is a recent effort of having persistent identifiers for other aspects of facilities research such as instruments or experimental techniques ([9]).

The promotion of the research idea through the facilities lifecycle has inspired the concept of Research Objects for facilities science ([10]) that acquire more and more detail whilst the investigation proceeds from its conceptual stage through the experiment to the research paper and associated artefacts.

An interesting recent development is the intention of some facilities to start publishing the descriptions of the approved research proposals (grants) – that are the "cores" to the future investigation entities – on the national research portals, e.g. ISIS neutron and muon source (www.isis.stfc.ac.uk) intends to publish the descriptions of all approved proposals on the UK common gateway to publicly funded research (http://gtr.rcuk.ac.uk/). The internal representation format for these entities is going to be CERIF (see under www.eurocris.org) that is widely used in the European grant information systems.

## 3 Research data in research discourse

### 3.1 The modes and purposes of sharing research data

The earlier mentioned NeXuS format, Research Objects and persistent identifiers for data present three different modes of sharing research data.

NeXuS file includes both data and data context (metadata) and thus offers research result as a "package" that can be interpreted by other researchers – or the same research team in future – with the help of format-compatible software. It is a responsibility of the "package" creator to embed all essential information in there; the boundaries of information context are very well defined (it is literally one data file).

Research Objects suggest the enrichment of information according to a specific model while the intellectual entity moves through the research lifecycle;

this implies that there is a "creator" to the model and the "curator" of intellectual entity on each phase of lifecycle; the boundaries of intellectual entity are more flexible (it may be an aggregation of various components) but are still well-defined.

The supply of nothing more but persistent identifiers for data, perhaps associated with some moderate contextual description (metadata), implies the paradigm of "open world" where intellectual entities can be deliberately constructed by various agents, hence there are no clear (predefined) boundaries to the entities, and virtually everyone can be considered a data "curator".

Sharing data or information, however, is not the end in itself and can be considered as a means to empower research discourse, to supply some intellectual entities into it. So quite often, when people speak of "research data" they actually mean intellectual entities where data may be just a component, or something associated with a "quantum" of research discourse.

This can be illustrated by observations over DataCite (www.datacite.org) – a platform that proclaimed goal is supplying data with dereferencaeble persistent identifiers (well-formed DOIs). The data centres who actually use DataCite in fact tend to assign DOIs not to datasets but to "quantums" of research discourse, e.g. to doctoral theses (that may of course contain some data but is not the data per se). In case of facilities science, we observe that DataCite DOIs are in fact dereferenceable to the landing Web pages that contain descriptions of *investigations* which are, as we explained it earlier, the series of experiments performed with a certain research goal on the assigned instrument within a dedicated timeslot.

So when a researcher cites "data" via DataCite DOI, she in fact quite often cites an intellectual entity – which can be a paper or something else, e.g. event (such as an earthquake) in geophysics, or investigation in the case of facilities science. [1] This attitude towards "data" DOIs assignment is only natural as what researchers tend to cite may not be "data" per se but certain identifiable elements of research discourse.

### 3.2 The place of investigation and the place of data in facilities research discourse

Investigation as an intellectual entity bears some features that are common with traditional research paper. Indeed, an investigation proposal is peer-reviewed; investigation can be cited from papers by the well-formed DOI and from other investigations, too, as when a researcher submits proposal, she refers to the relevant past publications and past investigations.[2]

The Figure 2 illustrates provenance relations between investigations and research papers that are a foundation for appropriate "citations".

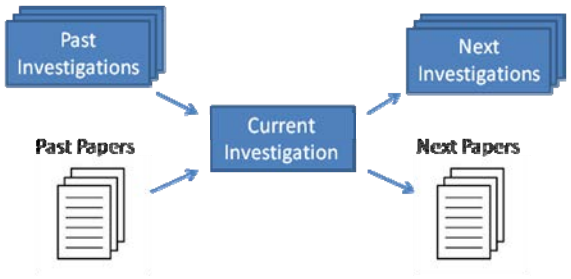



*Figure 2. Research provenance chain*

Similarities between investigation and research paper as intellectual entities are summarized in the Table 1.

*Table 1. Common features of investigations and research papers.*

| Feature / aspect | Publication (research paper) | Investigation |
|---|---|---|
| Is an intellectual entity | Yes | Yes |
| Is a subject of peer review | Yes | Yes (via proposal approval) |
| Can cite all significant intellectual entities of research discourse | Yes | Yes |
| Citation chains exist (steps of discourse observed | Yes | Yes |
| Universal identifiers available | Yes | Yes |

Looking into what intellectual entities can refer to what other intellectual entities (with the inclusion of datasets and software – which may or may not bear a clear identity) suggests the asymmetry in the direction of references so that e.g. a research paper can cite a dataset but not vice versa:

*Table 2. Cross-references of intellectual entities*

| References ("from" row "to" column) | Paper | Investi-gation | Dataset | Software |
|---|---|---|---|---|
| Paper | Yes | Yes | Yes | Yes |
| Investigation | Yes | Yes | Yes | Yes |
| Dataset | No | Yes | Yes | Yes (e.g. simulation) |
| Software | Yes (e.g. to paper about algorithm) | No | Yes | Yes |

[1] Examples of dereferenceable "data" DOIs that in fact resolve in investigation or research paper descriptions:
http://dx.doi.org/10.5286/ISIS.E.24066298
http://dx.doi.org/10.5167/UZH-27029

[2] Looking into the ICAT database for ISIS facility indicates the existence of investigation "chains" when the next investigation refers to the previous one, with as many as four investigations in a row undertaken in the last 10 years.

In fact, research discourse in facilities science splits into the two distinctive layers that can be called "research per se" and "data management"; this is illustrated by Figure 3.


*Figure 3. Directions of typical references and two layers of research discourse*

The two layers only loosely interact with each other and the bottom one can be considered a service layer in support of the top one, despite recent attempts to promote a view that information departments can play a role of data curation units, hence expanding their remit from the mere support of information technology to catering for richer tastes of researchers interested in semantic representation of information and in its sensible reuse ([11]).

### 3.3 Problems, challenges and opportunities

The above analysis contributes to modelling of research discourse in facilities science with the suggestion that data and software should play a modest (supportive) role compared to research papers and well-defined investigations. Different information models that can be applied to the same facilities research discourse. One of them is the model based on Research Objects ([10]) that suggest the "enrichment" of the core Investigation entity while it moves down the facilities research lifecycle illustrated by Figure 1 – turning into a rich aggregation of data, data context (metadata), and software. Another view is seeing research discourse as "grid" composed of provenance chains similar to that in Figure 2; the Research Activity model ([12]) offers a basic semantic means to support this view.

Irrespective of what of the two models we adhere to, they are likely to use the same techniques, e.g. for matching research papers with investigations.

One problem here is that, despite it is a requirement of facilities to submit the "input" to the investigation proposal and then the "output" of it in terms of research papers that led to the idea of the experiment, or have been resulted from it – there is no good curation of these bibliographic records, or a clear requirement for their format. On the other hand, when the institutional library eventually and independently collects the facility output in the form of research papers, they do it in a systematic way with good coverage and according to the best cataloguing practice but there is no record of the investigation that the paper has been resulted from as there is no requirement to capture it in the bibliographic record, also the investigation is often mentioned only implicitly in the paper. So if we want more context for the research papers and for the investigations, there is a task of matching bibliographic records coming from facilities User Office (the unit that looks after investigations lifecycle) and those in the institutional library catalogues.

To estimate the viability of automated techniques, we tried to match the bibliographic records for the papers that were the "input" to the investigations performed on ARGUS muon spectrometer.[3] We managed to visually identify the small number of the well-formed bibliographic records in the institutional repository that for sure match the corresponding poorly-formed ARGUS bibliographic records. We then applied different modifications to the ARGUS records in combination with measuring the Levenstein distance ([13]) between them and those in the library catalogue.

The first experiments suggest that bibliographic records from two systems: ePubs which is the institutional papers repository and ISIS ICAT which is the data catalogue supported by ISIS neutron and muon facility, can be successfully matched if we measure Levenstein distance between modified bibliographic records. A particular pretty simple technique could be the extraction and normalization of the numeric components from the bibliographic record (volume, pages and year), measuring distances between such normalized extracts – in effect, between two strings with only numbers in them – then playing with the threshold (the particular Levenstein distance) that allows to distinguish between matches and non-matches. This technique was tried out via bespoke Java software module and is illustrated by Table 3.

The technique tuning, including the measurements of precision and recall, should be done with the larger numbers of bibliographic records; there is about a thousand records in ICAT data catalogue that have bibliographic components – candidates for matching them with bibliographic records in ePubs papers repository. Yet it has to be understood that mere matching bibliographic records is just the first step in what we aspire to: a reasonably automated technique for linking investigations to research papers in situations where there are no bibliographic records catalogued for investigations, only investigations textual descriptions and other metadata.

There are more than ten thousand papers in ePubs repository that are marked up by the librarians as having relation to ISIS neutron and muon facility with no indication which investigation (series of experiments) or instrumental work they actually relate to. For the majority of these papers, there are no corresponding bibliographic records in the facility investigations catalogue; hence other techniques are required to match the papers to investigations. We consider decomposition of, on one hand, the bibliographic records from ePubs institutional repository and, on the other hand, the investigation descriptions from the ISIS investigations database into the corresponding elements, then looking into distances between elements with the further aggregation of them into sensible metrics. The analysis of bibliographic records and investigation descriptions suggests the following elements as the candidates for mutual mapping:

[3] http://www.isis.stfc.ac.uk/instruments/argus/argus6461.html

*Table3. Matching bibliographic records in ICAT data catalogue and ePubs papers repository (ARGUS case)*

| ICAT Reference | ePubs reference | Levenstein distance between full bibliographic references | Levenstein distance between "numeric" parts | Levenstein distance between "numeric" parts with the year normalized and the last page removed |
|---|---|---|---|---|
| Pratt et al, Phys. Rev. Lett. 96, 247203 (2006) | Phys Rev Lett 96 247203 (2006) | 17 | 0 | 0 |
| Lancaster et al, Phys. Rev B73, 020410(R) (2005) | Phys Rev B 73 020410 (2006) | 24 | 1 | 1 |
| Blundell and Pratt, J. Phys.: Condens. Matter 16, R771 (2004) | J Phys Condens Matter 16 R771-R828 (2004) | 30 | 3 | 0 |
| M.T.F.Telling and S.H.Kilcoyne, Electron transfer in dextran, J. Phys.: Condens. Matter 19 No 2 (17 January 2007) | J Phys Condens Matter 19 2 026221 (2007) | 81 | 6 | 6 |
| J Tomkinson and M.T.F Telling, Ammonium ions in alkali metal halide crystals: Tunnelling and spin relaxation, PCCP 2006 8 38 4434 | Phys Chem Chem Phys 8 4434-4440 (2006) | 113 | 12 | 5 |

The mentioned massive of records in the ISIS ICAT data catalogue (about a thousand of them) – for which the association with ePubs papers catalogue can be established via the earlier outlined bibliographic records matching technique – can be used for the validation of automated matching between investigation metadata records and (more than ten thousand) bibliographic records for all ISIS instruments. Then validation by the researchers themselves will be required, as well as some technical means in support of that validation – such as online polls.

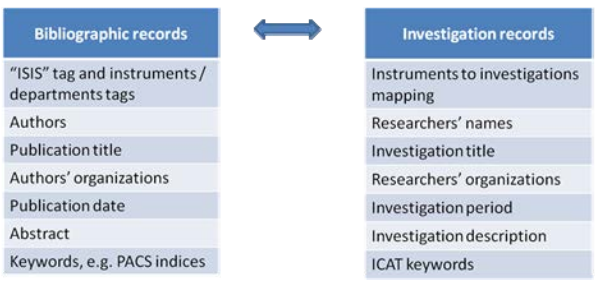



*Figure 4. Mapping research papers bibliography to investigations metadata*

Another opportunity for the validation of the investigations-to-publications matching technique will be looking into descriptions of research proposals (grants) in the research information portals. For ISIS facility, it will be Gateway to Research portal (http://gtr.rcuk.ac.uk/) that is about to start collecting investigation proposals in a systematic manner so that sometime after the investigations are over, they will be supplemented by the submission of research papers resulted from them. It will be possible then to use the newer investigations accompanied by papers resulted from them (as submitted by the researchers themselves) for the calibration of the automated matching technique that can be applied to the large corpus of past investigations and research papers.

Validated via two independent sources of bibliography: ePubs institutional repository and (forthcoming) records in the Gateway to Research portal, the automated matching technique may become a useful tool for research contextualization and for enrichment of the existing records in publications and data catalogues.

Apart from matching research papers with investigations, an interesting theme for further research could be looking into the cases of "indirect citations" when (see Figure 2) one research paper does not directly cite another one but there is an identifiable connection from one to another through the intermediary investigation; or the similar consideration from the investigations network perspective where one investigation does not explicitly refer to another but they are in fact connected through the intermediary research paper(s). Discovering these sorts of "indirect citations" may contribute to the development of alternative metrics for measuring research output, in addition to traditional paper citation metrics.

## 4 Conclusion

Our analysis indicates that Investigation in facilities science is an intellectual entity that has a clear identity, is involved in structured information exchange and bears some essential features similar to traditional research papers. There are various opportunities for the information modelling and for the formation of links between investigations and other intellectual entities, namely research papers that can be either an input to the investigation, or an outcome of it.

This study can be considered an analysis and a roadmap that precede the scalable experiments on the information contextualization in the domain of facilities science. It is also a call for information practitioners to share their views on the research information contextualization and on the role of various intellectual entities in their research domains, as the popular notion of "data" and its widely accepted importance may sometimes misrepresent the actual content of research discourse where other domain-specific intellectual entities could be more appropriate for sensible information management and for measuring research output.

## References


[1] Vasily Bunakov, Brian Matthews and Catherine Jones. Towards the Interoperable Data Environment for Facilities Science. A chapter in “Collaborative Knowledge in Scientific Research Networks” (AKATM book series). In press by IGI Global.

[2] J. Mesot. A need to rethink the business model of user labs? Neutron News, 2012, 23 (4), 2–3.

[3] S.J. Coles and P.A. Gale. Changing and Challenging Times for Service Crystallography. Chemical Science, 2012, 3 (3), 683–689.

[4] Reference Model for an Open Archival Information System. CCSDS 650.0-M-2 (Magenta Book) Issue 2, June 2012. http://public.ccsds.org/publications/archive/650x0m2.pdf

[5] B. Matthews et al. Using a Core Scientific Metadata Model in Large-Scale Facilities. 5th International Digital Curation Conference, London, UK.

[6] B. Matthews et al. Model of the data continuum in Photon and Neutron Facilities. PaNdata ODI, Deliverable D6.1. http://pan-data.eu/sites/pan-data.eu/files/PaNdataODI-D6.1.pdf

[7] D. Flannery et al. ICAT: Integrating Data Infrastructure for Facilities Based Science. In e-Science: Fifth IEEE International Conference on e-Science.

[8] Michael Wilson. Meeting a scientific facility provider's duty to maximise the value of data. Talk in DataCite Summer Meeting, Digital Research Data in Practice (DataCite2012), Copenhagen, Denmark. http://purl.org/net/epubs/work/62852

[9] PaNKOS: Proton and Neutron Knowledge Organisation System. www.purl.org/pankos

[10] B. Matthews et al. Investigations as research objects within facilities science. In 1st Workshop on Linking and Contextualizing Publications and Datasets, Malta, September 26th, 2013. http://purl.org/net/epubs/work/11912059

[11] Vasily Bunakov and Brian Matthews. Data Curation Framework for Facilities Science. In 2nd International Conference on Data Technologies and Applications, Reykjavik, Iceland, 29-31 Jul 2013, (2013): 211–216. http://purl.org/net/epubs/work/10938269

[12] Vasily Bunakov. Core semantic model for generic research activity. In 15th All-Russian Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections", Yaroslavl, Russia, 14-17 Oct 2013, CEUR Workshop Proceedings (ISSN 1613-0073) 1108 (2013): 79–84. http://ceur-ws.org/Vol-1108/paper10.pdf

[13] Владимир Левенштейн. Двоичные коды с исправлением выпадений, вставок и замещений символов. Доклады Академии Наук СССР (in Russian) 163 (4): 845–8, 1965. (Appeared in English as: Vladimir Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10 (8): 707–710, 1966.)

# Проблемы качества представления измерительных данных в научных е-публикациях

© В.В. Ежела
Сотрудничество КОМПАС
Институт физики высоких энергий
Протвино
ezhela@ihep.ru

## Аннотация

Обоснована необходимость и возможность изменения структуры научных е-публикаций и процедур предпубликационного рецензирования измерительного числового материала, представляемого к публикации. Использование устаревших издательских стандартов по представлению многомерных измерительных данных в вычислительной среде приводит к порче дорогостоящих измерительных данных. Недостаточность собственных профессиональных центров систематизации и оценки критических данных и знаний в России вынуждает наших доверчивых специалистов и метрологов использовать некорректные «оцененные» е-данные, зачастую поддержанные и/или рекомендованные авторитетными международными организациями. Такие данные не должны использоваться в модернизации Метрологии в РФ без жесткого критического анализа и сертификации, как это предписывается последними концепциями развития метрологической системы РФ.

В докладе также обсуждаются возможные пути выхода из этой парадоксальной ситуации:

(**i**) Общепринятая в науке система обеспечения качества публикуемых многомерных измерительных данных должна быть дополнена требованиями обязательного «числового рецензирования» материала;

(**ii**) При использовании заимствованных е-данных входной контроль их качества обязателен;

(**iii**) Метрологическая система РФ должна быть дополнена системой центров поверки е-данных. Для любых заимствованных данных, используемых в Науке, Технологии и в Образовании (НТО) РФ, сертификация обязательна.

(**iv**) В университетах РФ необходимо создание лабораторий систематизации и анализа данных по тематикам университетов, и своевременного производства актуального константного обеспечения НТО в РФ под надзором ГНМЦД(ГСССД).

На нескольких типичных примерах публикаций ведущих зарубежных центров систематизации и оценки экспериментальных данных и их отражения в российской метрологии показаны ситуации, породившие эти предложения.

*Работа выполнена при финансовой поддержке РФФИ, грант 14-07-000950. Автор благодарен членам сотрудничества КОМПАС, участникам семинара ИФВЭ, семинара Кафедры общей физики МФТИ, сотрудникам кафедры ФТИ ФИВТ МФТИ и участникам конференций СРТ за полезные обсуждения затронутых в докладе проблем. Особая благодарность организаторам СРТ за возможность привлечения внимания студентов и молодых специалистов к обсуждаемым проблемам.*

## 1 Введение

С переходом к электронным текстам и гипертекстам в обобществлении научных результатов появилась возможность избавиться от издательских ограничений на представление измерительных данных в бумажных научных публикациях. Более того, традиционный способ гарантии качества — квалификационное рецензирование материалов, представленных к опубликованию, также может стать более глубоким. Появляется возможность оперативной поверки корректности представления численного материала

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

(numerical peer review). Например, американское физическое общество пошло по пути разделения публикации на описательную часть (традиционную в бумажной технологии) и фактографическую часть, в которой численный материал, составляющий результат исследования, представлен в виде, удобном для дальнейшего использования в вычислительных средах [1].

Многие крупные исследовательские лаборатории создают собственные лабораторные хранилища веб-доступных числовых данных по гиперссылкам из текстов описательных статей, опубликованных в журналах. Однако структуры данных и стандарты представления измерительных данных все еще остаются прежними (сформированными в условиях бумажного обмена научными результатами).

Планетарная метрология сильно запаздывает с обобщением предложений исследователей по модернизации руководств по численному выражению результатов измерений (особенно многомерных) и с разработкой стандартов числового представления результатов измерений, наблюдений и оценок в вычислительной среде.

Это запаздывание приводит к возникновению научных статей-ребусов в большинстве быстро развивающихся областей естествознания, к обширной синонимии в метаданных для описания измерительных данных и даже к некорректным численным данным (частично из-за следования давно устаревшим стандартам).

Все больше экспертов, работающих с измерительными данными, обращают внимание научного сообщества на проблемы качества численных е-знаний в научных коммуникациях [6, 11, 12, 13, 14, 15] и необходимость регулярной систематизации возникающих уточненных данных, подстройки стандартов и их синхронизации со стандартами из смежных областей. Наиболее выпукло эта озабоченность выражена в недавней совместной работе звездных астрономов и ИТ-экспертов [16].

Одной из острых проблем научно-технического сообщества является проблема доступа к новым измерительным и наблюдательным данным, добываемым учеными и инженерами. В последнее десятилетие эксперты призывают к перестройке практики обмена дорогостоящими научными данными и их сохранения для грядущих поколений. Однако усилия в основном направлены на проблемы стандартизации описательной стороны наборов измерительных данных. Интенсивно обсуждается стандартизация метаданных («оформление упаковок данных») и методов обеспечения удобного и оперативного доступа к данным для совместного использования разнородных данных в наукоёмких приложениях.

Исчерпывающий обзор современного состояния идей и предложений по сохранению научных измерительных и наблюдательных данных представлен в недавнем документе КОДАТА [18], цитата следует:

"*…Based on a review of emerging practices and analysis of existing literature on citation practices, we* (*WG: www.codata.org, www.icsti.org,* www.icsu.org) *have identifie the following set of "first principles" for data citation:*

1. ***Status of Data:*** *Data citations should be accorded the same importance in the scholarly record as the citation of other objects.*
2. ***Attribution:*** *Citations should facilitate giving scholarly credit and legal attribution to all parties responsible for those data.*
3. ***Persistence:*** *Citations should be as durable as the cited objects.*
4. ***Access:*** *Citations should facilitate access both to the data themselves and to such associated metadata and documentation as are necessary for both humans and machines to make informed use of the referenced data.*
5. ***Discovery:*** *Citations should support the discovery of data and their documentation.*
6. ***Provenance:*** *Citations should facilitate the establishment of provenance of data.*
7. ***Granularity:*** *Citations should support the finest-grained description necessary to identify the data.*
8. ***Verifiability:*** *Citations should contain information sufficient to identify the data unambiguously.*
9. ***Metadata Standards:*** *Citations should employ widely accepted metadata standards.*
10. ***Flexibility:*** *Citation methods should be sufficiently flexible to accommodate the variant practices among communities but should not communities…*".

Далее доклад организован по подразделам:

**2. Проблемы международной рекомендательной метрологии**

(или почему документы ИСО — продукты рабочей группы WG1(JCGM) пока не могут служить основой для разработки или гармонизации российских стандартов).

**3. Проблемы оценки Фундаментальных Физических Констант (ФФК).** История, современное состояние, необходимость регулярной и независимой, но синхронной переоценки ФФК в тесном сотрудничестве физиков-систематиков, ИТ-экспертов и метрологов России с физиками-систематиками и метрологами США, ЕС, Японии и Китая в согласовании новых версий переоценок ФФК.

**4. Предложение по реализации концепции развития метрологической системы РФ** Создание в университетах России центров данных и лабораторий систематизации данных и их критическому анализу по тематикам университетов, производству и своевременной сертификации современного константного обеспечения НТО России в сотрудничествах с центрами данных ГСССД РФ и с зарубежными центрами данных.

В докладе обращается внимание экспертов на не менее важную **проблему обеспечения качества численного содержания наборов измерительных данных**, потому что никому не нужны хорошо организованные и легкодоступные, но недостоверные измерительные (оцененные) данные.

Для полноты представления тем доклада приведем выдержку из последней «Концепции развития национальной системы стандартизации Российской Федерации на период до 2020 года» [17], цитата следует:

«…***3. Разработка национальных стандартов в приоритетных отраслях экономики***

*Разработка национальных стандартов в приоритетных отраслях экономики должна осуществляться на основе общепринятых международных принципов стандартизации.*

*При этом необходимо обеспечить:*

*ежегодное обновление от 10 процентов до 15 процентов фонда стандартов в секторах экономики с высоким потенциалом развития;*

*гармонизацию национальных стандартов с международными стандартами;*

*сокращение сроков разработки и обновления национальных стандартов, в том числе исходя из обязательств, принятых Российской Федерацией при вступлении во Всемирную торговую организацию;*

*создание механизма постоянного обновления национальных стандартов на базе передовых международных и региональных стандартов, обеспечение разработки национальных стандартов на базе проектов международных стандартов (до их окончательного принятия) с учетом требований законодательства Российской Федерации;*

*разработку документов по стандартизации в целях соблюдения требований, не относящихся к техническому регулированию, и их гармонизацию с аналогичными требованиями, установленными в государствах – членах Таможенного союза и Европейского союза, с учетом требований законодательства Российской Федерации;* …».

Мы показываем, что отнесение некоторых международных стандартов к разряду «передовых» и «гармонизация» национальных стандартов с международными стандартами может оказаться неподъемной (в разумное время) задачей. Основная причина — отсутствие некоторых видов деятельности в наших метрологических организациях и службах. Без специалистов, работающих в областях сопряжения фундаментальной науки, информатики и метрологии, невозможно своевременное восприятие и освоение новых направлений в планетарной метрологии.

## 2 Проблемы международной рекомендательной метрологии

Первый документ международной организации КОДАТА, призывающий к объединению национальных усилий в стандартизации измерительных данных и их сохранению на машинных носителях появился в конце 1973 года [2]. В 1977 году к этому призыву присоединилась Международная Палата Мер и Весов (CIPM) и начала работу по формированию технического комитета и рабочих групп для разработки рекомендательных документов по формализации идеи «Единства Измерений».

К 1995 году призыв КОДАТА частично реализовался в документе ISO: Guide to the Expression of Uncertainty in Measurement (GUM-1995) с весьма ограниченной областью приложения. Документ применим только к измерениям одной величины. В случае косвенных измерений одной величины, зависящей от нескольких случайных величин, оценки которых получены в статистически независимых измерениях, рекомендации GUM-1995 применимы только в случаях, когда допустим линейный закон переноса неопределенностей. К сожалению, после «косметической» правки руководства GUM-1995 и его переиздания в 2008 году, область применимости GUM-2008, осталась прежней. Для косвенных измерений величин-функций, нелинейно зависящих от случайных величин-аргументов, рекомендации GUM-2008 некорректны, и их не следует использовать как основу для совершенствования стандартов. Более детальная критика некоторых «вредоносных» положений GUM-2008 приведена в работе [6].

Несмотря на эти ограничения и некоторые недоработки, продуктивная идея о необходимости планетарного единства измерений была сразу воспринята многими метрологическими системами в разных странах. Документ GUM-1995/2008 стал рекомендательным руководством к формированию национальных стандартов и во многом способствует становлению планетарного единства измерений (см. например документы: [7, 8, 9, 10]. К сожалению, российские метрологии не заметили указанных некорректностей и явно нелепых рекомендаций при переводе руководства GUM на русский язык.

Сопровождением и дальнейшей разработкой концептуального базиса идеи «единства измерений», применимого для большинства измерительных ситуаций, занимается рабочая группа объединенного комитета по развитию GUM –WG1 of Joint Committee for Guides on Metrology (JCGM) [19]. Уже разработаны первые приложения к GUM для расширения области применимости GUM: GUM-S1 [4] и GUM-S2 [5]. Попытки использовать их в реальных задачах показали необходимость пересмотра избранного подхода к формализации идеи «единства измерений» и уже появилось соответствующее официальное заявление рабочей группы [20].

Рабочая группа WG1 — это интернациональный коллектив, образованный в 1997 году для продолжительной совместной работы. К сожалению, в этой рабочей группе не работали и не работают эксперты из метрологического сообщества РФ. А эта группа разрабатывает и сопровождает руководства, фокусирующие усилия сообществ, работающих в созидании и использовании измерительных данных, на проблемы достижения планетарного единства измерений.

Участие российских специалистов в работе этой группы представляется весьма актуальным для реализации нашей концепции метрологического развития: «…*обеспечение разработки национальных стандартов на базе проектов международных стандартов (до их окончательного принятия) с учетом требований законодательства Российской Федерации*…».

Воплощение концепции развития метрологической системы РФ в реально работающую метрологическую систему и ее гармонизация с «мировой Метрологией», по-видимому, возможны, но пути реализации концепции не подкреплены соответствующими «дорожными картами» и пока остаются весьма туманными. **Однако, реализация концепции невозможна без тесного сотрудничества ученых-систематиков, ИТ-экспертов и метрологов РФ с зарубежными экспертами в формировании базы рекомендаций и синхронизированных документов на национальных языках для подстройки национальных стандартов.**

## 3 Проблемы оценки Фундаментальных Физических Констант (ФФК)

Исторически сложилось так, что “регулярная” переоценка ФФК проводится только в центре данных FCDC(NIST) [21] с 1973 года, под надзором и при участии группы международных экспертов “CODATA Task Group on Fundamental Constants” [22], в которой принимают участие и российские специалисты из рабочей группы РНК КОДАТА при президиуме РАН по “Фундаментальным физическим константам и единицам измерений физических величин” [23]. С результатами нескольких последних последовательных переоценок ФФК и их качеством можно ознакомиться по опубликованным отчетам в научных обзорных журналах. Эти отчеты свободно доступны на сайте FCDC(NIST) (см. [26], [27], [28], [29] и комментарии к ним). Здесь для примера мы приводим таблицу выборки констант, две из которых базовые $\boldsymbol{h}$ и $\boldsymbol{\alpha(0)}$ (крупные символы в таблице).

| ***CODATA*: 1986 (1987)** | **Symbol** | **Unit** | **Value(Uncertainty)·Scale** | **Correlations** | | |
|---|---|---|---|---|---|---|
| **Elementary charge** | $\boldsymbol{e}$ | **C** | $\mathbf{1.602\ 177\ 33\ (49)\cdot 10^{-19}}$ | $\boldsymbol{e}$ | $\boldsymbol{h}$ | $\boldsymbol{m_e}$ |
| **Planck constant** | $\boldsymbol{h}$ | **J s** | $\mathbf{6.626\ 075\ 5\ (40)\cdot 10^{-34}}$ | **0.997** | | |
| **Electron mass** | $\boldsymbol{me}$ | **kg** | $\mathbf{9.109\ 389\ 7\ (54)\cdot 10^{-31}}$ | **0.975** | **0.989** | |
| **1/(Fine struct. const.)** | $\mathbf{1/\alpha(0)}$ | | **137.035 989 5 (61)** | **−0.226** | **−0.154** | **−0.005** |

Такие же выборки были сделаны с сайта FCDC(NIST) в разные годы и для каждой из них были вычислены собственные числа матриц корреляций неопределенностей.

Результаты вычислений приведены в следующей таблице, из которой видно, что результаты всех последовательных переоценок несостоятельны, так как матрицы корреляций неопределенностей имеют отрицательные собственные значения. Двойные даты в первой колонке этой таблицы дают указание на (год переоценки и выкладки численных результатов на сайте FCDC : год публикации описательной части в обзорных журналах).

**Собственные числа матриц корреляций неопределенностей констант {*e, h, $m_e$, 1/α*(0)}, оценки которых были рекомендованы КОДАТА**

**1986: 1987**
**{**2.99891, 1.00084, 0.000420779, **–0.000172106}**

**1998: 2000**
**{**2.99029, 1.01003**, –0.000441572**, 0.000123580**}**

**2002: 2005**
**{**2.99802, 1.00173, 0.000434393, **–0.000183906}**

**2006: 2008**
**{**2.99942, 1.00006, 0.000719993, **–0.000202165}**

**2010: 2012**
**{**2.99983, 1.00022, **–0.0000451921**, **$-5.92939\cdot10^{-6}$}**

Причин такого систематического конфуза несколько. Частично они разобраны в работах [24], [25], [6]. Основной причиной, по-видимому, является неоправданное использование линейных приближений в «теоретических» соотношениях связи физических наблюдаемых величин (observational equations) с базовыми ФФК, косвенно измеряемыми (согласуемыми) по методу наименьших квадратов (МНК). Это приводит к возникновению смещенных оценок базовых ФФК и плохо обусловленных матриц ковариаций их неопределенностей. Возможные смещения не обсуждаются в публикациях FCDC(NIST) и, по-видимому, не оцениваются для обоснования применения «линейного» МНК при согласовании и «линейного» переноса неопределенностей от базовых (adjusted) к выведенным (derived) ФФК при их вычислениях.

Одним из важных критериев состоятельности результата измерения является полнота представления численного материала, необходимого для проведения численного рецензирования и для предоставления возможности использования результатов измерения в высокоточных приложениях. В последней переоценке 2010 года из **67** базовых ФФК (см [29], с. 1579, TABLE.XXX и с. 1580, TABLE.XXXII) **30** ФФК — это аддитивные поправки $\delta_{(.)}(\ldots)$ к теоретическим формулам для учета их несовершенства (например, оценки вкладов высших порядков теории возмущений в КЭД).

К сожалению, значения этих поправок и коэффициентов корреляций их с ФФК, полученные при согласовании ФФК, ни в печатных версиях, ни на сайте FCDC(NIST) не представлены. Этот численный материал по $\delta_{(.)}(\ldots)$ не отражен и в наиболее полной, по численному материалу, версии CODATA-02 [27]. Сокрытие этих данных препятствует корректному использованию ФФК в высокоточных вычислениях в атомной спектроскопии. Свидетельство о недопустимости пренебрежения информацией о корреляциях ФФК в вычислениях атомных спектров приведено в публикации [30] с участием экспертов из FCDC(NIST), цитата следует:

*"…The energy level $E_i$ of state i can be written as a function of the fundamental constants and an additional adjusted constant $\delta_i$ which takes into account the uncertainty in the theory [27, 30, 31].*

*For example, for the case in which i is a state of hydrogen, we have*

$$E_i = H_i\,(R, A_r(e),\ A_r(p), R_p) + \delta_i \qquad (1)$$

*where the constants that appear as arguments of the function $H_i$ are listed in Table II. Because the values of the constants in Eq. (1), including $\delta_i$, result from a least squares adjustment, they are correlated, particularly those for R, and $R_p$, which have a correlation coefficient of* **0.996**.

*The uncertainty of the calculated value for the 1S-2S frequency in hydrogen is increased by a factor of about* **500** *if such correlations are neglected…."*

Следует отметить, что значения ФФК используются в бесчисленных приложениях с разными требованиями к полноте и достоверности значений ФФК. Большинство приложений не требуют исчерпывающего численного представления результатов измерений ФФК. В многих прикладных справочниках, в учебниках и на сайтах вообще не упоминается наличие корреляций неопределенностей ФФК, свободно доступных на сайте FCDC.

Однако, сокрытие уже имеющихся знаний, добытых длительным кропотливым трудом, — это грубое нарушение научной этики (см. десять заповедей авторам измерительных данных от исследователей живых систем в [11]).

Для приложений, требующих высокоточных вычислений с несколькими ФФК совместно, использование несостоятельных ФФК недопустимо, а если высокоточное приложение еще и чревато повышенной опасностью в технических реализациях, то использование несостоятельных ФФК, рекомендованных как актуальные и надежные преступно.

Вызывает удивление документ ГСССД 237-2008, «ФУНДАМЕНТАЛЬНЫЕ ФИЗИЧЕСКИЕ КОНСТАНТЫ», МОСКВА, СТАНДАРТИНФОРМ, 2009, под рубрикой «ТАБЛИЦЫ СТАНДАРТНЫХ СПРАВОЧНЫХ ДАННЫХ», в котором полиграфически блестяще напечатаны «копипастянутые» численные значения констант, выложенных на сайте FCDC в 2007 году и опубликованных [28] в 2008 году. Сокрытие информации об оцененной в FCDC коррелированности неопределенностей ФФК в этом документе, по-видимому, означает признание экспертами ГСССД факта несостоятельности этой переоценки ФФК для использования их в высокоточных вычислениях неопределенностей функций, зависящих от нескольких ФФК. Но производители таблиц ГСССД 237-2008 не предостерегают потенциальных пользователей об этой особенности таблиц и их неполноте (отсутствию значений **30** настраиваемых поправочных констант и их корреляций). Остались без должного внимания предостережения из оригинального текста на с. 635, [28]: цитата следует:

*«…Since the calculations are carried out with more significant figures than are displayed in the text to avoid rounding errors, data with more digits are available on the FCDC website for possible independent analysis…*с. 635.*»*

Кроме того, остается неопределенность с синхронизацией оригинала и его освоения в нашей метрологической системе: очередная переоценка ФФК была произведена в FCDC в 2010 году, результаты выставлены на сайте FCDC в начале 2011 года и опубликованы в 2012 году [29].

Похоже, эксперты ГСССД не успевают осваивать последовательные концепции [17] и, как следствие, не успевают осваивать и сертифицировать новые переоценки ФФК к использованию в НТО РФ. Частично это запаздывание, по-видимому, связано с непониманием причин отсутствия 30 базовых ФФК в таблицах согласованных ФФК в FCDC и появлением в результатах согласования некорректных матриц ковариаций и корреляций.

**Российские потребители ФФК от ГСССД вынуждены использовать «одряхлевшие» данные 8-летней давности**.

Таким образом, становится очевидной острая необходимость глубокого и строгого численного предпубликационного рецензирования результатов измерений, представляемых для опубликования в журналах и/или на сайтах е-журналов. Также очевидна необходимость тщательного численного контроля качества заимствуемых научно-технических ресурсов. Однако, без «единства

численного выражения и представления измерений в е-среде» оперативное численное рецензирование невозможно. Жесткий стандарт на численное выражение и представление результатов измерений равно как и разделение публикаций на описательную и численную части в е-науке неизбежны. Хочется верить, что эксперты из JCGM учтут эту неизбежность при пересмотре и развитии GUM и его расширений.

Другой, столь же очевидный вывод из анализа современной практики обмена научными достижениями — это необходимость национальных, но сотрудничающих центров данных в созидании научных информационных ресурсов, критических для развития региональных и планетарной НТО. Для обеспечения жизнеспособности национальных центров данных нужны: целевое финансовое обеспечение их основной работы; подготовленные ученые-систематики; программы их воспроизводства в вузах РФ. В нашем центре данных такая программа была сформулирована (в прошлые 90-е годы в рамках ОСССД, распадающегося ГКАЭ СССР) в положении о Центре Данных Физики Частиц в ИФВЭ.

В последней концепции развития Метрологической Системы РФ [17] необходимость программ воспроизводства квалифицированных систематиков и метрологов для работ в центрах данных сформулирована достаточно полно и внятно, но в рекомендательном стиле, без учета реальной ситуации в НТО РФ, цитата следует:

«***11. Совершенствование системы подготовки специалистов и экспертов в области стандартизации***

*…С учетом динамичного развития стандартизации следует обеспечить как подготовку специалистов в области стандартизации в высших и средних специальных учебных заведениях, так и периодическое повышение квалификации работающих специалистов.*

*Для решения этих задач необходимо:*

*актуализировать или ввести в образовательных учреждениях высшего и среднего профессионального образования инженерного и экономического профиля дисциплины по стандартизации по соответствующим направлениям подготовки;*

*развивать возможности получения обучающимися в образовательных учреждениях высшего профессионального образования и среднего профессионального образования дополнительного образования в области стандартизации параллельно с освоением ими основной профессиональной образовательной программы;*

*обеспечить в соответствии с требованиями федеральных государственных образовательных стандартов высшего профессионального образования привлечение практикующих специалистов в области стандартизации к формированию соответствующих компетенций в рамках подготовки бакалавров, специалистов и магистров; …».*

**В РФ исчезающе мало профессиональных центров данных по фундаментальным направлениям естествознания. Нет или почти нет профессиональных открытых центров данных и по приоритетным направлениям развития технологий.** Комитеты есть, комиссии есть, рабочие группы есть, концептуальная ГСССД есть, а вот центров данных для формирования надежных научных фактографических ресурсов поддержки инновационного развития НТО в РФ нет.

**Куда, для чего и как готовить специалистов? Вот в чем вопрос!**

## 4 Предложение по реализации концепции развития метрологической системы РФ

В каждом вузе России необходимо создать лаборатории систематизации данных и их критическому анализу по тематикам вуза, производству и своевременной сертификации актуального константного обеспечения НТО вуза в сотрудничестве с центрами данных ГСССД РФ и с зарубежными центрами данных. Для этого, по-видимому, нужна ФЦП по реализации концепции развития метрологической системы РФ, если такой еще нет, а если есть, то подправить ее с учетом некоторых предложений этого доклада, если эксперты ФЦП сочтут их полезными.

В лабораториях:

– Сначала создавать профильные системы поверки качества научных знаний в национальных и зарубежных информ-ресурсах (необходимых вузам для собственных научных исследований, в подготовке молодой научной смены и при перепрофилировании научных кадров на «прорывные» проекты);

– Организовывать сотрудничества с национальными и зарубежными центрами данных по совместному сопровождению и синхронизации «сертифицированных знаний»;

– Преобразовывать лаборатории систематизации в вузовские профильные центры данных, а их системы поверки в базы генерации профильных системных знаний;

– Обучать и тренировать профильных специалистов-систематиков в вузовских центрах данных.

Можно ожидать, что такие, непрерывно действующие структуры, приживутся в наших вузах и будут способствовать своевременному повышению уровня профессиональной информированности обучающих и обучаемых, готовить достойную и ответственную смену для корпоративных центров данных и для центров данных ГСССД по специальностям: физик-систематик, химик-систематик, нано-систематик, био-систематик, инфо-систематик, когни-систематик, …

## Литература


[1] EPAPS http://www.aip.org/publishing/authors/supporting-data

[2] D. Gavrin, T. Golashvili, H.V. Kehiaian, N. Kurti, E. F. Westrum Jr. [CODATA Task Group on Publication of Data in the Primary Literature)] “Guide for the Presentation in the Primary Literature of Numerical Data Derived from Experiments”, CODATA Bulletin **9,** 1973.

[3] BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, and OIML “Guide to the Expression of Uncertainty in Measurement”, 1995 ISO/IEC Guide 98:1995, ISBN 92-67-10188-9, Second Edition. ISOGUM, JCGM 100:2008, Evaluation of measurement data – “Guide to the expression of uncertainty in Measurement”, 2008. http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf

[4] JCGM 101:2008, Evaluation of measurement data – Supplement 1 to the – “Guide to the expression of uncertainty in measurement” – Propagation of distributions using a Monte Carlo method, 2008 http://www.bipm.org/utils/common/documents/jcgm/JCGM_101_2008_E.pdf

[5] JCGM 102:2011, Evaluation of measurement data – Supplement 2 to the – “Guide to the expression of uncertainty in measurement” – Extension to any number of output quantities, 2011 http://www.bipm.org/utils/common/documents/jcgm/JCGM_102_2011_E.pdf

[6] Vladimir V. Ezhela “Multimeasurand ISO GUM supplement is Urgent”, CODATA DSJ, **6**, S676-S789, 2007; [Errata: CODATA DSJ, **7,** E2-E2] , 2007.

[7] Barry N. Taylor and Chris E. Kuyatt, “Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results” NIST Technical Note 1297, 1994 http://www.nist.gov/pml/pubs/tn1297/

[8] (ВНИИМ GUM) Руководство по выражению неопределенности измерения. Перевод ISO GUM, ВНИИМ, 1999.

[9] РМГ 43 2001, Применение “Руководства по выражению неопределенности измерения”, 2003.

[10] NASA-HDBK-8739.19-3,“Measurement Uncertainty Analysis Principles and Methods”, 2010 https://standards.nasa.gov/documents/detail/3315776

[11] NAP publication, “SHARING PUBLICATION-RELATED DATA AND MATERIALS: RESPONSIBILITIES OF AUTHORSHIP IN THE LIFE SCIENCES”, Washington, D. C., 2003 http://www.nap.edu/catalog.php?record_id=10613

[12] Nicole M. Radziwill “Foundation for Quality Management of Scientific Data Products”, Quality Management Journal, **13**(2), 7, 2006.

[13] David R. Lide, “Data quality - more important than ever in the Internet age”, CODATA DSJ, **6,** 154-155, 2007.

[14] Ray P. Norris, **“**How to Make the Dream Come True: The Astronomer’s Data Manifesto”, CODATA DSJ, **6,** S116-S124, 2007.

[15] Shuichi Iwata, “SCIENTIFIC "AGENDA" OF DATA SCIENCE”, CODATA DSJ, **7,** 54-56, 2008.

[16] О.Ю. Малков и др. Задачи и данные в области звездной астрономии // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XV Всероссийской научной конференции RCDL-2013, Ярославль, Россия, 14–17 окт. 2013 г. – Ярославль: ЯрГУ, 2013. – С. 38–47. – http://rcdl.ru/doc/2013/paper/pd_6.pdf;

[17] Правительство РФ, Распоряжение от 28.02.**2006**, № 266-р, М. Фрадков, Правительство РФ, Распоряжение от 12.10.**2010**, № 1760-р, В. Путин, Правительство РФ, Распоряжение от 24.09.**2012**, № 1762-р, Д. Медведев.

[18] CODATA-ICSTI Task Group on Data Citation Standards and Practices “OUT OF SITE OUT OF MIND: The Current State of Practice, Policy, and Technology for the Citation of Data”, CODATA DSJ, **12,** CIDCR1-CIDCR75б, **2**013.

[19] JCGM, http://www.iso.org/sites/JCGM/JCGM-introduction.htm

[20] Walter Bich et al.,[ JCGM.WG1], “Revision of the ‘Guide to the Expression of Uncertainty in Measurement””, Metrologia **49**, 702–705, 2012.

[21] FCDC (NIST)-Fundamental Constants Data Center http://www.nist.gov/pml/div684/fcdc/

[22] CIPM, CODATA FPC Working group http://www.bipm.org/extra/codata/members.html

[23] РНК РАН http://codata.gcras.ru/ РНК РАН Рабочая группа по ФФК http://www.gao.spb.ru/russian/psas/kodata/index.html

[24] V.V. Ezhela, Y.V. Kuyanov, V.N. Larin and A.S. Siver, “The Inconstancy of the Fundamental Physical Constants: Computational Status”, 2004 web.ihep.su/library/ps/2004-36.pdf

[25] В.В. Ежела. О корректном числовом представлении результатов косвенных измерений. 2006 http://web.ihep.su/library/pubs/prep2006/ps/2006-28.pdf

[26] P.J. Mohr and B.N. Taylor “CODATA Recommended Values of the Fundamental Physical Constants: 1998”, Rev. Mod. Phys. **72**, 351, 2000 CODATA-98.

[27] P.J. Mohr and B.N. Taylor, “CODATA Recommended Values of the Fundamental Physical Constants: 2002”, Rev. Mod. Phys. **77**, 1. 2005 CODATA-02.

[28] P.J. Mohr, B.N. Taylor, and D.B. Newell, “CODATA Recommended Values of the Fundamental Physical Constants: 2006”, Rev. Mod. Phys. **80**, 633,2008. CODATA-06.

[29] P.J. Mohr, B.N. Taylor, and D.B. Newell “CODATA Recommended Values of the

Fundamental Physical Constants: 2010", Rev. Mod. Phys. **84**, 1527, 2012. CODATA-10.

[30] U.D. Jentschura, S. Kotochigova, Eric-Olivier Le Bigot, P.J. Mohr, and B.N. Taylor, "Precise energies of highly excited hydrogen and deuterium" PRL **95**, 163003, 2005.

## On Problems with Measured Data Quality in Scientific e-publishing

Vladimir. V. Ezhela

It is argued that the structural changes in scientific e-publishing of the research results based on measured data are unavoidable, and even are on the marsch already. Namely: narrative part and factual numerical parts of the research report should be separated somehow. The factual part should be strongly connected by metadata and hyperrefs with the narrative part, but it should be in a computer usable form to provide possibility for a special **numerical peer review** being as deep-and-fast as possible for quality assuarance of the reported measured data. It seems that the absence of the selfconsistent numerical standards for multidimensional measured data expression and presentation in computer media is the main reason for the appearance of unintended corrupted published data. It is hoped that JCGM experts will fill this harmful missing in Metrology soon.

# Структуры заимствований в диссертациях по историческим наукам

© П.В. Ботов © Д.В. Вьючнов © Н.С. Суровенко
© А.С. Хританков © С.В. Царьков © Ю.В. Чехович
ЗАО «Анти-Плагиат»
Москва

khritankov@antiplagiat.ru

**Аннотация**

В работе описано исследование структуры взаимных заимствований текстовых фрагментов в диссертациях кандидатов и докторов наук по историческим специальностям рубрикатора ВАК (07.xx.xx). С помощью алгоритмических, статистических методов и методов анализа графов и сетей были обнаружены группы сильно связанных по заимствованиям между собой диссертаций, обнаружены «скомпилированные» работы и указаны предполагаемые источники таких компиляций.

## 1 Введение

В данной статье представлены результаты исследования диссертаций на соискание степеней кандидатов и докторов наук по историческим наукам (коды специальностей ВАК: 07.xx.xx), проведенного по заказу Российской Государственной Библиотеки с использованием Электронной библиотеки диссертаций РГБ (ЭБД РГБ), системы «Антиплагиат» и специального программного обеспечения обработки данных и машинного обучения.

ЭБД РГБ [7] содержит библиографические описания и полные тексты авторефератов и диссертаций по различным специальностям ВАК, полученные путем сканирования текстовых документов.

Система «Антиплагиат» [1, 4, 6, 20] позволяет проводить для текста проверяемого документа и произвольной коллекции источников сравнительный анализ. Результатом такого анализа является список всех значимых фрагментов проверяемого документа, совпадающих полностью или частично с фрагментами в коллекции источников. Совпадения фрагментов текстов документа и источников обозначаются как «заимствования». При этом практически совпадения могут иметь различную интерпретацию: цитирование источника, цитирование третьего неизвестного текста в обеих работах, академический плагиат, использование общеупотребимых словосочетаний, случайное совпадение и т.д. Результат работы системы обычно анализируется экспертом, который и принимает решение о том, как квалифицировать обнаруженные системой заимствования и об академической ценности работы в целом [21]. Работа эксперта требует значительных затрат времени для квалифицированного анализа объемной диссертации – от нескольких часов до нескольких дней на одну работу. С учетом того, что в России ежегодно защищается около 25 тысяч диссертаций, проверка всего потока работ оказывается практически неподъемной задачей.

Основной целью проведенного исследования, таким образом, стала проверка технической возможности глубокого автоматического анализа заимствований в больших текстовых коллекциях для формирования «грубого фильтра» работ для последующего экспертного анализа. Такой фильтр позволил бы выделять часть работ, проведение экспертного анализа которых необходимо. В настоящем исследовании авторы главным образом сосредоточились на выборе процедур предобработки исходных данных, постобработки результатов и настройках параметров системы, с целью автоматизации и уточнения результатов последующей экспертной обработки.

Инициатором и заказчиком исследования выступила РГБ. Основные направления исследования были сформулированы в виде нескольких гипотез. В данной статье представлены результаты по гипотезам и исследовательским вопросам, приведенным в разделе 2.

Для корректного учета заимствований необходимо было исключить из состава обнаруженных совпадений корректно оформленные цитаты (см. раздел 3) и технические заимствования – общие фрагменты диссертаций вследствие использования общего формата, шаблона и правил

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

оформления, а также списка литературы (см. раздел 4).

После предварительной обработки, возможно проведение более глубокого анализа и проверка гипотез (см. раздел 5).

## 2 Гипотезы и цели исследования

В ходе исследования предполагалось проверить следующие гипотезы и дать ответы на вопросы:

− определить возможность проведения глубокого анализа заимствований в объемных текстовых коллекциях на наличие некорректных заимствований;

− оценить долю работ с существенными заимствованиями текста из других диссертаций;

− понять, является ли подготовка таких работ частью процессов систематической компиляции, либо это единичные не связанные случаи.

## 3 Выделение корректно оформленных цитат

В тексте диссертации автор может дословно цитировать фрагменты других произведений. Цитаты оформляются в соответствии с правилами русского языка [15], библиографические ссылки к ним – согласно стандарту [16]. Так как цитата дословно повторяет часть другого текста, она может быть распознана поисковыми модулями системы «Антиплагиат» как заимствованный блок, поэтому нужно выделять корректно оформленные цитаты и исключать их из блоков заимствований.

Для выделения цитат предлагается подход, основанный на применении методов машинного обучения и состоящий из трех этапов:

1. Выделение текстовых блоков-кандидатов при помощи эвристик.

2. Расчет значений признаков для блоков-кандидатов.

3. Бинарная классификация блоков-кандидатов по принадлежности к классу корректно оформленных цитат.

На первом этапе текстовые блоки выделяются согласно правилам русского языка [15]. Практически во всех случаях цитируемый текст должен быть заключен в кавычки. Исключением из этого правила являются стихотворения, которые можно цитировать без кавычек в случае сохранения авторских переносов строк. Так как цитирование стихов не свойственно диссертациям по историческим наукам, то для повышения точности распознавания и снижения сложности системы в качестве блока-кандидата выделяется текст, заключенный в кавычки. При этом учитывается, что одни блоки могут быть вложены в другие.

На втором этапе происходит расчет значений признаков блоков-кандидатов. Признаки построены на основе правил оформления цитат и библиографических ссылок. Например, реализован признак, что если после текста цитаты в пределах одного предложения встретилось слово, написанное слитно с числом, или число следует сразу после закрывающей кавычки в блоке-кандидате, то значение признака равно 1, иначе 0.

Таких признаков было построено более 60, однако в результате отбора, о котором будет рассказано ниже, было оставлено только 23.

На третьем этапе к рассчитанным значениям признаков блоков применяется обученная модель дерева решений, выполняющая бинарную классификацию, является ли блок корректно оформленной цитатой или нет.

Для построения и настройки модели были вручную размечены тексты диссертаций по историческим наукам. Для этого была разработана программа разметки корректно оформленных цитат среди блоков текстов с графическим интерфейсом. Всего исходные данные составили 24479 блоков, в которых 4277 корректно оформленных цитат. Набор данных был разделен на обучающие данные из 16320 блоков (из которых 2848 корректно оформленных цитат) и тестовые из 8159 блоков (из которых 1429 цитат).

Далее, на обучающих данных с помощью программы Weka [17] были проанализированы признаки и с применением критерия «Gain Ratio» [18] отобрано 23 признака для классификации блоков.

Для построения дерева решений был использован алгоритм C4.5 [18]. Модель дерева решений использована потому, что ее можно интерпретировать в виде правил «если – то», понятных даже не специалисту в области машинного обучения. Глубина дерева была ограничена значением 7. Оценка качества проводилась по двум критериям: точность и полнота.

Точность – это доля верно выделенных моделью корректно оформленных цитат среди всех выделенных моделью текстовых блоков.

Полнота – это доля верно выделенных моделью корректно оформленных цитат среди всех корректно оформленных цитат.

В результате для использованной в работе модели на обучающей выборке точность составила 96,8%, полнота – 73,5%, на тестовой выборке точность составила 95,8%, полнота – 43,8%.

## 4 Предварительная обработка данных

Система «Антиплагиат» анализирует тексты документов, строит по ним инвертированный индекс групп последовательно идущих слов (n-грамм) [19] и сравнивает документы попарно после нахождения потенциально совпадающих блоков в индексе.

На вход были поданы тексты диссертаций коллекции ЭБД РГБ по историческим наукам 07.xx.xx, всего более 14 тыс. кандидатских и

докторских диссертаций, защищенных преимущественно в 1999–2012 гг. (рис. 1). Атрибуты библиографического описания диссертаций также получены из ЭБД РГБ. Были исключены 51 документ с ошибками выделения текста и 114 документов размером менее 15 тысяч символов. Бимодальное распределение документов по годам соответствует содержанию ЭБД РГБ и, по видимому, является следствием порядка оцифровки документов в РГБ.

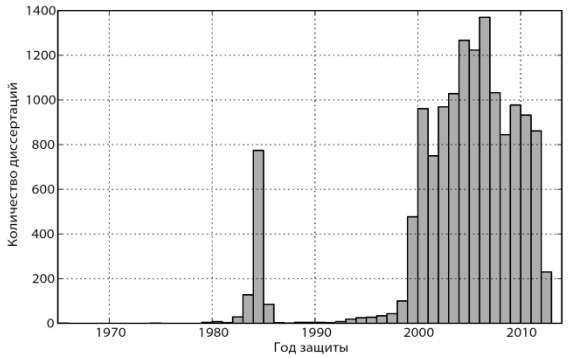



Рис. 1. Количество диссертаций по годам защиты

При поиске заимствований между документами одной коллекции возникает проблема установления направления заимствования и формирования набора источников. В данном исследовании проблема была решена следующим образом. Для каждой диссертации отбиралось 100 источников с наибольшим количеством заимствований из них в данной диссертации. Минимальный размер блока заимствования варьировался от трёх до семи слов в зависимости от контекста. Направление заимствования устанавливалось эвристически по году защиты диссертации. Полагалось, что источником заимствования является диссертация, год защиты который предшествует году зашиты рассматриваемой диссертации.

Вычисления блоков заимствований проводились на сервере с восемью виртуальными ядрами Xeon 1,6 ГГц, 6 ГБ ОЗУ в течение четырех дней. Было проведено три итерации вычислений блоков с различными параметрами. Полное время проведения вычисления блоков с учетом пауз между итерации составило две недели. Общий несжатый объем блоков заимствований в XML формате составил около 4 ГБ.

Полученные блоки заимствования были дополнительно обработаны: выполнено объединение блоков, исключение корректных цитирований, повторное объединение, фильтрация по размеру блока.

Алгоритм объединения блоков составлял из двух блоков, разделенных менее чем 30 символами, один блок, включающий оригинальные блоки и символы между ними (рис. 4).

После объединения блоков из них были исключены корректно оформленные цитаты, сформированы новые блоки, которые были повторно объединены тем же алгоритмом.

Предварительный анализ расположения и размера блоков заимствований (рис. 2) показал, что большая часть совпадающих блоков находится в титульном листе и, по-видимому, области библиографии диссертации. Предполагая, что эти блоки связаны с общим форматом титульного листа и сходными источниками в списке литературы, исключены блоки, находящиеся в первых 1000 символов и последних 10% текста диссертации.

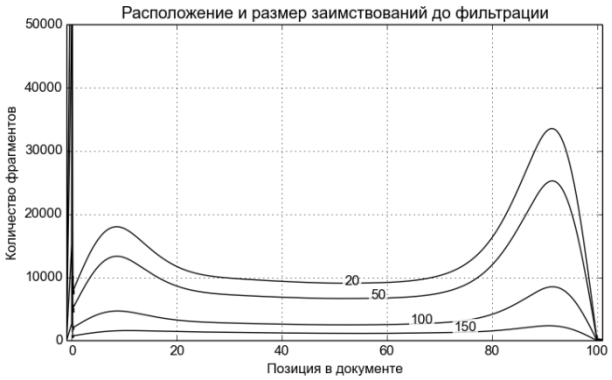



Рис. 2. Размер и позиция блоков до предварительной обработки. Изоденсы обозначают размер блоков, значения выбраны экспертно

По результатам анализа распределения блоков по размеру в разных частях документа, были исключены блоки размером менее 250 символов как незначительные заимствования, по большей части относящиеся к введению и библиографии. В дальнейшем при построении графа заимствований были исключены блоки размером менее 750 символов, в результате пропадает зависимость между размером блока и его положением в документе.

В результате были построены распределение блоков по размеру и положению в документе (рис. 3), направленный граф заимствований, составлен список диссертаций с наибольшей долей заимствованного текста.

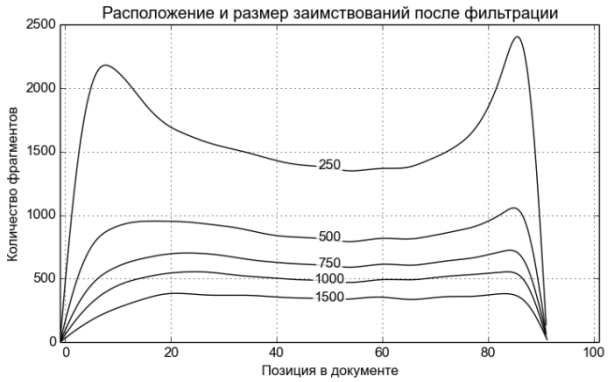



Рис. 3. Размер и позиция блоков после фильтрации, исключения цитат и объединения блоков. Изоденсы обозначают размер блоков, значения выбраны экспертно

В текстах диссертаций были замечены и исследованы аномалии – чаще всего связанные с ошибками оцифровки или обработки документов.

В частности, около 50 документов состояло из склеенных в одном тексте нескольких диссертаций, которые также встречались отдельно.

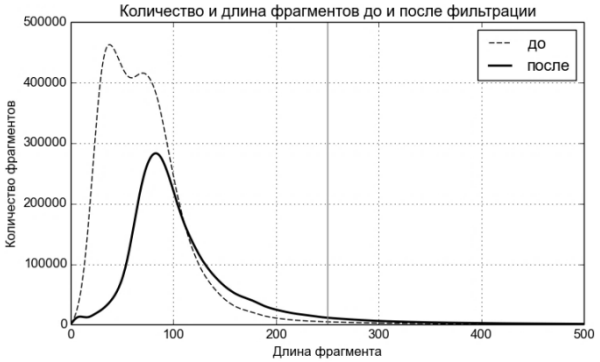



Рис. 4. К описанию алгоритма слияния блоков

## 5 Выделение групп диссертаций

Анализ групп и сообществ диссертаций позволяет установить «контекст» заимствований между ними, выделить скрытые внутренние структуры заимствований. Для проведения такого анализа заимствования между диссертациями в данной работе был построен граф, в котором в качестве вершин были диссертации, а ребра определялись заимствованиями из этих работ. Вес ребра рассчитывался как количество совпадающего текста в символах.

Для анализа графов и сетей используются специализированные алгоритмы объединения вершин графа в кластеры, называемые сообществами (community). В работе [2] предложен быстрый алгоритм поиска сообществ в графах, основанный на максимизации внутреннего критерия качества – модульности (modularity):

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j),$$

где $A_{ij}$ – вес дуги между $i$ и $j$, $k_i = \sum_j A_{ij}$ – сумма весов дуг, связанных с вершиной $i$, $c_i$ – сообщество, к которому принадлежит вершина $i$, $\delta$-функция $\delta(u,v)$ равна 1, если $u = v$, и 0 иначе, и $m = \frac{1}{2}\sum_{ij} A_{ij}$ .

Алгоритм выделения сообществ [2] состоит из итеративно повторяющихся двух шагов.

На первом шаге каждая вершина графа приписывается к своему уникальному сообществу. Затем для каждой вершины $i$ рассматривается возможность её переноса в сообщество вершины $j$, до которой из $i$ есть ребро, при условии, что модульность увеличивается. Процесс повторяется, пока модульность не достигнет локального максимума.

На втором шаге из полученных сообществ получают вершины для нового графа, веса ребер которого определяются суммой весов ребер вершин, входящих в сообщество. Таким образом, первый шаг можно заново выполнить для нового графа.

Итерации продолжаются до тех пор, пока с новой итерацией не перестанет изменяться состав сообществ.

Всего в исходном графе получилось порядка 13 000 вершин и 164 000 ребер. В исходном графе, при отсутствии фильтрации, присутствовала гигантская компонента (giant component) размером в 12000 вершин, что указывало на наличие большого числа «шумовых» ребер. Предполагая, что шумовые ребра имеют небольшой вес, можно подобрать пороговое значение, отсекающее большинство таких ребер. С другой стороны, завышение порога отсечения могло привести к удалению значимых связей между вершинами, образующих сообщества и искажении структуры сообществ в графе. Поэтому необходимо было подобрать порог минимального допустимого веса ребра для выделения сообществ.

В эксперименте были проанализированы зависимости следующих параметров от порога отсечения: количество выделяемых сообществ, количество слабо связанных компонент в графе, максимальный размер связанного компонента (рис. 5–6).

При увеличении порога количество сообществ и связанных компонент возрастало за счет «развала» гигантской связанной компоненты (см. рис. 5), достигло максимума, а затем начало убывать. Эта точка максимума и определила искомый порог отсечения, так как дальнейшее его увеличение приводило к удалению значимых связей между вершинами и уменьшению количества сообществ.

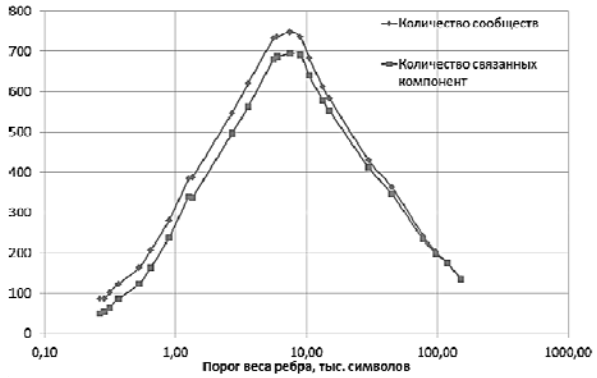



Рис. 5. Зависимость количества связанных компонент и количества сообществ от порога веса ребра

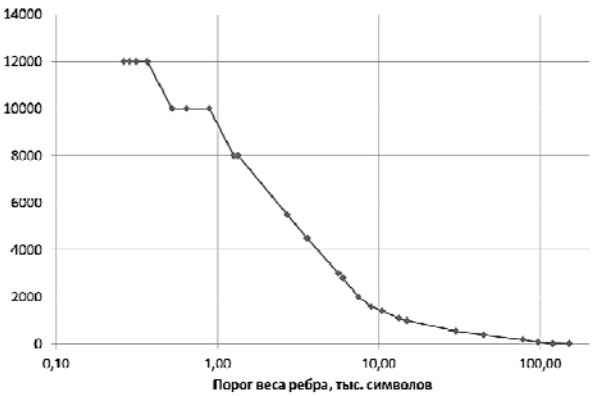



Рис. 6. Зависимость максимального размера связанного компонента от порога веса ребра

В результате порог веса ребра выбран равным 0,05, что соответствует суммарному заимствованию в 7500 символов между диссертациями. При данном пороге в графе выделяется 748 сообществ.

Полученные сообщества характеризуются более высоким уровнем заимствования среди диссертаций сообщества, чем из диссертаций вне сообщества. Пример сообщества и заимствований между диссертациями показан на рис. 7.

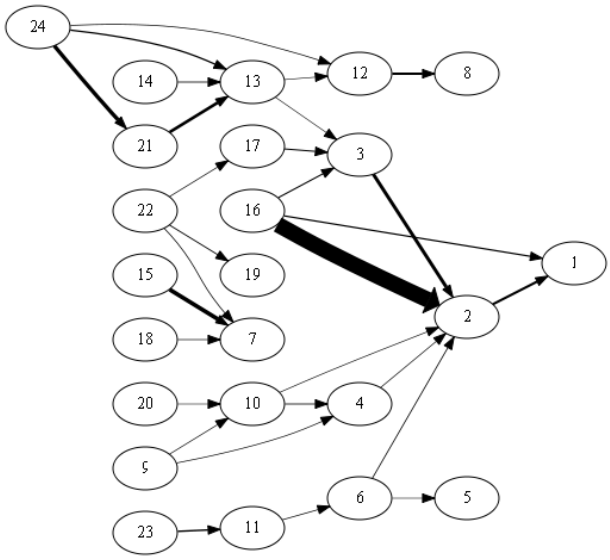



Рис. 7. Пример найденного сообщества. Диссертации представлены вершинами графа и пронумерованы, заимствования показаны ребрами, толщина ребра пропорциональна объему заимствования

В сообществах диссертации могут выполнять две функции: являться источниками для заимствований и получателями заимствований из других источников. На рис. 7 диссертации 24, 16, 22 можно назвать популярными источниками в данном сообществе. Диссертации 2, 3, 7, 13 – получатели заимствований. Заметим, что 2, 3 и 13 при этом так же используются в качестве источников для заимствования другими диссертациями. Жирная стрелка между работами 2 и 16 указывает на большой объем заимствованного текста.

Источники и получатели заимствований можно найти в большинстве сообществ. В таких сообществах существенны заимствования текста между диссертациями, что указывает на наличие коллективов, занимающихся подготовкой диссертаций путем компиляции из других работ. Отнесение источников заимствования к сообществу позволяет увидеть сообщество в целом и не указывает на автора источника как участника коллектива.

Если все сообщества диссертаций расположить на диаграмме с зависимостью полного объема заимствования от среднего их объема по заимствованиям внутри сообщества (рис. 8), то среди них можно выделить три вида. Небольшие сообщества диссертаций с высоким средним объемом заимствований, по-видимому, скомпилированных в индивидуальном порядке из небольшого числа работ назовём «индивидуальными предпринимателями». Большие сообщества с умеренным средним размером заимствований – «фабрики диссертаций», а также «странные сообщества», которые не получается однозначно отнести к предыдущим двум видам. Диссертации из сообществ, не относящихся к указанным, полагаются подготовленными научными группами, не основанными на систематических заимствованиях текстов диссертаций.

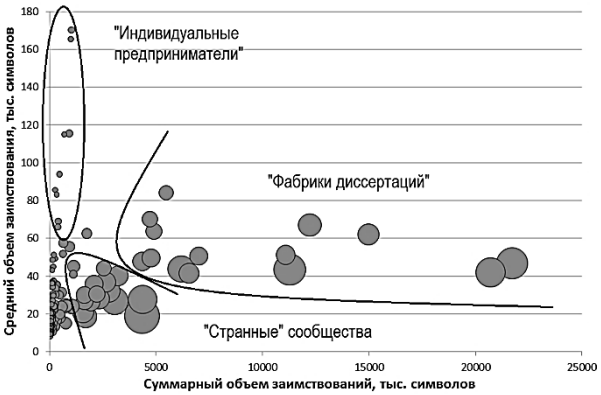



Рис. 8. Сообщества диссертаций по среднему объему заимствования (по вертикали) и суммарному объему (по горизонтали) с условной классификацией по видам. Площадь метки соответствует размеру сообществ, на диаграмме – от 4 до 169 диссертаций

При анализе заимствований в диссертациях, вследствие использования только ЭБД РГБ в качестве источника данных, не учитывались заимствования из других источников, статей, журналов. Такого рода заимствования, в исследуемом графе заимствований, могут косвенно проявляться как заимствования между диссертациями, если в них имеется общий текст из стороннего источника.

## 6 Сходные исследования

Диссертации, защищаемые в области наук, в целом отражают структуру и состояние исследований в своей области, и представляют отдельный интерес как объект научного исследования. Исследования диссертаций и научных работ, связей между ними проводились ранее в других областях [8–13]. В работах [8, 9] проведено исследование диссертаций и авторефератов с целью выявления научных школ, связей между научными руководителями и диссертантами, использованы методы анализа текстов. В исследовании авторефератов докторских диссертаций [10] проведен анализ качества подготовки диссертаций за 2008–2011 годы по материалам, опубликованным на сайте ВАК.

Проведенное исследование отличается использованием данных ЭБД РГБ [7], полных текстов диссертаций, рассмотрением диссертаций по историческим наукам и механизмом установления связей между диссертациями – по текстовым заимствованиям, и методами анализа

полученного графа. Причем наличие текстовых заимствований, с нашей точки зрения, указывает на общность в подготовке текстов диссертаций.

Определение общности научных работ по текстовым заимствованиям – достаточно распространенный метод [1, 5], однако известны и другие подходы, основанные на методах анализа текстов [13] и рассмотрении совместного библиографического цитирования между документами [14].

## 7 Заключение

Насколько известно авторам, проведенное исследование по определению структур заимствований в диссертациях является первым в своем роде. Исследованные гипотезы и вопросы ранее не выдвигались. Поэтому так же важно, что были отработаны методы исследования.

Проведенное исследование продемонстрировало техническую возможность проведения анализа заимствований в крупных текстовых коллекциях с применением системы «Антиплагиат» в совокупности с методами анализа данных для фильтрации потока диссертационных работ и выделения документов, для которых необходим последующий экспертный анализ.

Было обнаружено, что большинство проверенных диссертаций не имеют значимых заимствований. Однако не менее 500 работ имеют существенный объем более 33% общих текстовых фрагментов с другими диссертациями, что может указывать либо на наличие общих источников заимствования, либо на прямое заимствование.

В построенном графе заимствований обнаружены коллективы и «сообщества» диссертаций, по-видимому, связанные с процессом их подготовки. Сообщества с большим объемом заимствований между диссертациями отнесены к коллективам, в которых налажен процесс подготовки текстов диссертаций путем компиляции из готовых источников.

Результаты исследований были предоставлены на рассмотрение экспертам РГБ и получили положительную оценку. В дальнейшем планируется проведение подобных исследований и в других областях науки.

## Литература


[1] Авдеева Н.В., Ботов П.В., Букаев А.С., Вислый А.И., Груздев И.А., Житлухин Д.А., Романов М.Ю., Чехович Ю.В. Внедрение системы «Антиплагиат» в Российской государственной библиотеке // Интеллектуализация обработки информации: 8-я международная конференция. Республика Кипр, г. Пафос, 17–24 окт. 2010 г.: сб. докл. – М.: МАКС пресс, 2010. – С. 499–503.

[2] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre. Fast unfolding of communities in large networks // Journal of Statistical Mechanics: Theory and Experiment 2008(10):P10008 (2008).

[3] R. Lambiotte, J.C. Delvenne, M. Barahona. Laplacian dynamics and multiscale modular structure in networks // Arxiv preprint arXiv:0812.1770 (2008).

[4] ЗАО Анти-Плагиат, Система «Антиплагиат». http://www.antiplagiat.ru

[5] iParadigms, LLC. Turnitin. Plagiarism prevention engine. Available online at: http://www.turnitin.com

[6] Шарапов Р.В., Шарапова Е.В. Система проверки текстов на заимствования из других источников // Труды 13-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL’2011. –Воронеж, 2011. –С.121–126.

[7] Лавренова О.А. Развитие проекта библиотеки электронных диссертаций и авторефератов в открытом доступе // Образовательные технологии и общество (Educational Technology & Society). – Казань: Изд-во Казанский государственный технологический университет. – 2006. – Т. 9, № 3. – С. 335–341.

[8] Ю.В. Леонова, А.М. Федотов. Извлечение знаний и фактов из текстов диссертаций и авторефератов для изучения связей научных сообществ // Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL-2013, Ярославль, Россия, 14–17 октября 2013 г. – Ярославль: ЯрГУ, 2013. – С. 135–144.

[9] Леонова Ю.В., Добрынин А.А., Веснин А.Ю. Построение графа диссертаций // XIV Российская конференция с участием иностранных ученых «Распределенные информационные и вычислительные ресурсы» (DICR-2012): программа конференции и тезисы докладов (Новосибирск, Россия, 26–30 нояб. 2012). – Новосибирск: ИВТ СО РАН, 2012. – С. 17. – ISBN 978-5-905569-05-0.

[10] Донецкая С.С. Статистическое исследование структуры и качества подготовки докторских диссертаций в России // Вопросы статистики. – 2012. – № 12. – С. 71–76.

[11] Бескаравайная Е.В., Митрошин И.А. Анализ базы данных диссертаций ПНЦ РАН // Информационное обеспечение науки. Новые технологии: сб. науч. тр. / Н.Е. Каленов (ред.). – М.: Научный Мир, 2011. – С. 124–133.

[12] Ю.Н. Климов. Количественно-информационный анализ потока публикаций по библиотекам и библиотековедению на основе поиска по ключевым словам в базе данных Science-Direct // Межотраслевая информационная служба. – 2011. – № 3. С. 51–58.

[13] В.Н. Захаров, А. А. Хорошилов. Автоматическая оценка подобия тематического содержания текстов на основе сравнения их формализованных смысловых описаний // Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15–18 окт. 2012 г. – С. 189–195.

[14] Bela Gipp and Joeran Beel, 2009 "Citation Proximity Analysis (CPA) – A new approach for identifying related work based on Co-Citation Analysis" in Birger Larsen and Jacqueline Leta, editors, Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09), volume 2, pages 571–575, Rio de Janeiro (Brazil), July 2009.

[15] Розенталь Д.Э., Джанджакова Е.В., Кабанова Н.П. Справочник по правописанию, произношению, литературному редактированию. – Издание второе, исправленное. – М.: ЧеРо, 1998. – 400 с.

[16] ГОСТ Р 7.0.5–2008 Библиографическая ссылка, общие требования и правила составления.

[17] University of Waitako. Weka Toolkit. http://www.cs.waikato.ac.nz/~ml/weka/

[18] J. Ross Quinlan. C4.5: Programs for Machine learning. Morgan Kaufmann Publishers 1993.

[19] К.Д. Маннинг, П. Рагхаван, Х. Шютце. Введение в информационный поиск. : Пер. с англ. – М.: ООО «И.Д. Вильямс», 2011. – 528 с.

[20] Авдеева Н.В., Никулина О.В., Сологубов А.М. Система «Антиплагиат.РГБ» и недобросовестные авторы диссертаций: кто победит? // Научная периодика: проблемы и решения. – 2012. – №5(11). – С. 11–16.

[21] Авдеева Н.В., Лобанова Г.А. Классификация фрагментов текста при экспертизе диссертаций на предмет заимствований (плагиата) // «Информационные ресурсы России»: науч.-практ. журн. – М.: ФГБУ «Российское энергетическое агентство» Минэнерго России. – 2014. – № 11. – С. 2–6.

## Structures of Text Paraphrasing and Plagiarism in Dissertations on Historical Sciences

P.V. Botov, Y.V. Chehovich, A.S. Khritankov, N.S. Surovenko, S.V. Tsarkov, D.V. Viuchnov

We report on the research of structures in graphs of text paraphrasing and plagiarism in Ph.D. dissertations on historical sciences in Russia (07.xx.xx, according to HAC classification). Using algorithmic, statistical and network analysis methods we discovered groups of highly related dissertations, which intensely borrowed from each other, which we call “science shops”, found so-called “compiled” works and probable sources of such compilations.

# Метод обнаружения дубликатов в потоке текстовых документов

© А.М. Андреев © Д.В. Березкин © И.А. Козлов © К.В. Симаков
МГТУ им. Н.Э. Баумана,
Москва
arkandreev@gmail.com dmitryb2007@yandex.ru kozlovilya89@gmail.com skv@ixlab.ru

**Аннотация**

Работа посвящена решению задачи устранения дублирующихся документов из потока текстовых сообщений. Приведена многокритериальная модель документа, предложен метод обнаружения дубликатов на основе бинарной классификации с помощью метода опорных векторов. Основной акцент сделан на обеспечении применимости метода для обработки документов из разных предметных областей. Предложен способ снижения вычислительной сложности метода посредством предварительной фильтрации кандидатов.

## 1 Введение

В настоящее время во многих предметных областях существует потребность в формировании больших текстовых коллекций. При этом производится сбор текстовой информации из открытых Интернет-источников, а также специализированных ресурсов. Основной областью использования создаваемых таким образом хранилищ документов является интеллектуальная обработка текстов, которую, как правило, можно отнести к классу Text Mining.

С ростом количества разнообразных источников данных в сети Интернет (новостные сайты, блоги, социальные сети) всё более серьезной проблемой становится дублирование информации. Сообщения, публикуемые одним источником, зачастую многократно перепечатываются другими (в исходном виде или с небольшими изменениями). В результате, при выполнении автоматического сбора документов из многочисленных источников в формируемой текстовой коллекции накапливаются идентичные или близкие по содержанию документы – дубликаты. В некоторых задачах наличие дубликатов должно учитываться – например, при определении значимости сообщений [14]. Но в большинстве случаев попадание таких документов в коллекцию снижает её качество [12, 16].

В данной статье рассматривается решение задачи обнаружения дубликатов в потоке текстовых сообщений. Особое внимание уделяется обеспечению возможности использования разработанного метода для обработки документов из различных предметных областей.

## 2 Постановка задачи

### 2.1 Функционирование системы сбора и обработки новостной информации

В работе [9] авторами было предложено решение задачи качественного автоматического сбора новостных данных из Интернет-источников, предполагающего извлечение с веб-страницы текста новости, а также сопутствующих метаданных, включающих название, дату публикации, автора новости и др. При этом осуществляется контроль корректности извлекаемой информации, то есть проверка соответствия текстов загружаемых документов исходным текстам новостей на сайте.

Текстовые данные, извлеченные с веб-сайтов, подвергаются обработке различными методами интеллектуального анализа [7-8], такими как автоматическая классификация и кластеризация документов, извлечение знаний и фактов из естественно-языковых текстов, выявление трендов и прогноз развития ситуаций.

Для эффективного применения перечисленных методов обработки текстовых данных необходимо обеспечить качество анализируемой коллекции документов. Помимо вышеупомянутой корректности каждого конкретного текста, качество коллекции подразумевает требование оригинальности составляющих её новостей. Присутствие в обрабатываемом наборе одинаковых или очень близких по содержанию документов может отрицательно сказаться на качестве обработки. Это касается работы модулей, выполняющих статистический анализ документов, например, модуля анализа трендов. Его работа основана на выявлении в коллекции новостей, относящихся к анализируемой ситуации, и определении зависимости частоты встречаемости таких документов от времени. Появление дублей приведет

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

к многократному учету модулем идентичных новостей, что повлечет за собой некорректный вид построенной зависимости.

Для обеспечения оригинальности документов, составляющих текстовую коллекцию, в систему сбора необходимо встроить подсистему, задачей которой является оперативное обнаружение и удаление из коллекции нечетких дубликатов (рис. 1). Она должна анализировать каждый загружаемый документ и принимать решение о его оригинальности. Для этого необходимо сравнить его с загруженными ранее новостями и определить, является ли он нечетким дубликатом одной из них. При обнаружении дубля он должен быть удален до этапа загрузки данных в базу данных.

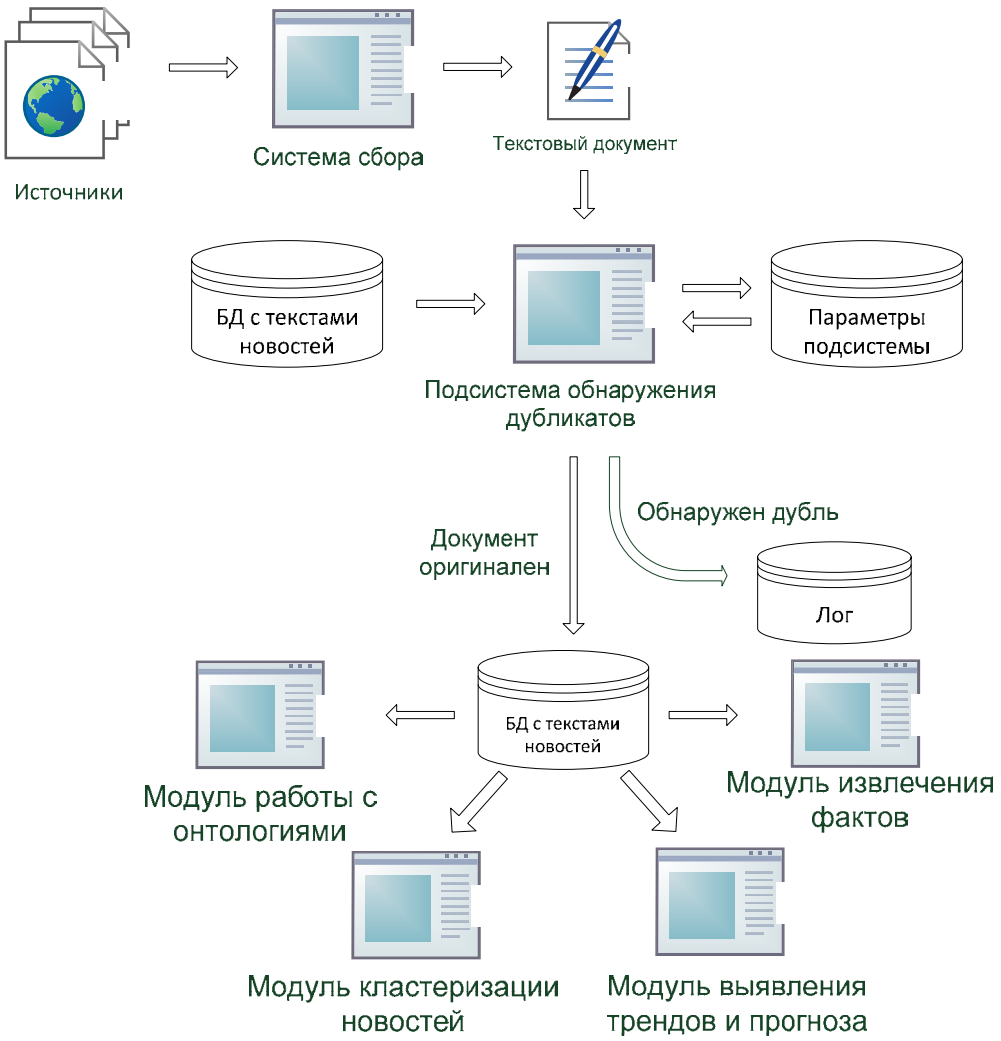



Рис. 1. Место подсистемы обнаружения дубликатов в системе автоматизированного сбора и анализа новостной информации

### 2.2 Особенности решаемой задачи

Если проблема обнаружения и удаления полных дублей тривиальна, то при необходимости распознавать нечеткие дубликаты (то есть, документы, имеющие различный текст, но близкие по содержанию) возникают значительные сложности.

В связи с явлением частичного дублирования в работе [10] предлагается понятие информативной необходимости элемента высказывания: «Всякий частично-дублетный элемент, если он отвечает коммуникативной задаче высказывания, признается информативно-необходимым в той же мере, что и прочие, недублетные элементы речи: такой элемент информативно необходим постольку, поскольку вносит свой уникальный, неповторимый в других элементах, вклад в суммарную информацию, передаваемую высказыванием». Таким образом, задача обнаружения нечетких дубликатов состоит в распознавании и удалении сообщений, не являющихся информативно-необходимыми.

Установить информативную необходимость и ценность некоторого высказывания возможно только с учетом соответствующего ситуативного контекста. Так, при ручной проверке документов эксперт, определяя наличие или отсутствие дублирования, принимает решение с учетом предметной области и характера документов, составляющих ана-

лизируемую коллекцию. Например, при работе с юридическими документами особое внимание должно уделяться метаданным – для текстов такого типа два документа с практически идентичным содержанием, но различающимися названиями и датами публикации не могут считаться дублями. Другой подход используется при анализе экономических новостей, например, сводок о состоянии фондового рынка – дубликатами не должны признаваться документы, имеющие одно и то же название и одинаковый текст, но различающиеся числовыми данными (значениями курсов валют).

Таким образом, в разных случаях эксперт сравнивает документы с точки зрения различных критериев (близость текстового содержания, сходство названий, разница во времени публикации), то есть использует различные модели распознавания дубликатов в зависимости от предметной области и контекста коммуникационных сообщений. Поэтому не представляется возможным выработать единую систему правил для обнаружения нечеткого дублирования сразу для всех случаев. Следовательно, разрабатываемая подсистема должна иметь возможность гибкой настройки на разные предметные области, что позволит ей моделировать деятельность эксперта по распознаванию дублей с использованием различных моделей. Поскольку система сбора выполняет извлечение документов из множества источников, которые относятся к различным предметным областям, необходимо обеспечить возможность работы с несколькими моделями распознавания дубликатов одновременно.

Еще одной проблемой, с которой приходится столкнуться при решении задачи устранения дубликатов, является большой объем обрабатываемых данных. Наличие в коллекции сотен тысяч документов делает весьма трудоемким анализ каждого нового сообщения путем сравнения его с каждым из ранее загруженных. Эта проблема может быть решена с помощью приближенных методов, но их применение ведет к снижению качества обнаружения дубликатов, то есть уменьшению точности и полноты [12]. При разработке предлагаемого подхода решалась задача совмещения высокого качества и низкой вычислительной сложности проверки документов.

## 3 Обзор методов обнаружения дублей

### 3.1 Методы, основанные на использовании шинглов

Одним из наиболее популярных методов, используемых при поиске нечетких дубликатов вебдокументов, является алгоритм шинглов [2, 5]. Он основан на представлении документа в виде множества всевозможных последовательностей фиксированной длины *k*, состоящих из соседних слов. Такие последовательности называются «шинглами». Два документа считаются похожими, если их множества шинглов значительно пересекаются. Количество шинглов примерно равно длине документа в словах, поэтому в целях повышения эффективности авторы оригинального алгоритма предложили несколько способов сэмплирования множества.

Дальнейшим развитием этого метода стал алгоритм «супершинглов» [4]. Его идея состоит в применении к элементам множества шинглов различных хэш-функций и выборе для каждой из них шингла, минимизирующего её значение. Из выбранных шинглов формируются группы, именуемые «супершинглами». Два документа считаются похожими, если мера сходства их наборов «супершинглов» не меньше заданного значения.

### 3.2 Сигнатурные методы

Другим распространенным классом приближенных подходов к поиску нечетких дубликатов является класс сигнатурных методов. Подробный обзор алгоритмов этого класса выполнен в [12]. Общей идеей является представление документа с помощью одного числового значения – «сигнатуры», что сводит проверку схожести документов к сравнению их сигнатур. Совпадение этих значений означает, что документы являются нечеткими дубликатами. Существует множество способов вычисления сигнатур документов:

- использование хэш-функции, вычисленной для всего документа (это позволяет обнаруживать лишь точные дубликаты);
- использование хэш-функции, вычисленной для строки, полученной из сцепленных в алфавитном порядке нескольких слов документа с наибольшими значениями весов, рассчитанных различными методами (например, TF, TF-IDF и OptFreq);
- использование хэш-функции, вычисленной для строки, полученной из сцепленных в алфавитном порядке нескольких наиболее длинных или «тяжелых» (то есть, состоящих из слов с наибольшим суммарным значением весов) предложений документа.

Несколько иной подход предложен в работе [14]: здесь сигнатура представляет собой не хэш-сумму цепочки слов, а саму цепочку. При этом документы признаются дубликатами при совпадении заданного числа элементов их цепочек.

### 3.3 Методы, использующие векторные модели

В задачах интеллектуальной обработки текстов (Text Mining) широко используются векторные модели текстовых документов. При этом каждое сообщение представляется в виде вектора в многомерном признаковом пространстве $D = (D^1, D^2, ..., D^N)$, каждый элемент которого отражает некоторую характеристику документа.

В качестве элементов могут использоваться слова, встречающиеся в текстах коллекции [11]. При этом значениями элементов вектора (1), представляющего некоторый документ, являются веса соответствующих слов, отражающие их значимость для этого документа:

$$d_i = (w_i^1, w_i^2, ..., w_i^n), \qquad (1)$$

где $N$ – общее количество различных слов во всех документах, $w_i^j$ – вес $j$-ого слова в $i$-ом документе. Хотя такой выбор признакового пространства является наиболее распространенным, могут применяться и другие характеристики текстов, например, частота появления различных пар символов или частота появления тех или иных частей речи [17].

В работе [1] векторная модель использована при решении задачи обнаружения дубликатов. При этом вектором представляется не отдельный документ, а пара документов из обучающей выборки. В качестве значения элемента вектора здесь используется произведение весов соответствующего слова в первом и втором документе пары. Полученный вектор подвергается классификации с помощью метода опорных векторов (support vector machine, SVM) [15] для принятия решения о наличии или отсутствии дублирования.

Несколько иной подход предложен в статье [13]. Он также использует векторное представление пары документов, но вектор в целом здесь характеризует схожесть элементов пары, а его отдельные компоненты – близость документов с точки зрения различных критериев. Пары, отмеченные в обучающей выборке как «дубликаты» или «не дубликаты», представлены двумя кластерами точек многомерного пространства. Таким образом, задача обнаружения дублей сводится к классификации новых пар документов, то есть отнесению их к одному из этих кластеров. Классификация основана на выборе кластера, центроид которого находится ближе к точке, представляющей новую пару документов.

### 3.4 Метод, основанный на выявлении близких по смыслу частей текстов

В [16] решается несколько иная, но близкая задача: формирование из группы документов, описывающих некоторое событие, одного сообщения, содержащего только оригинальную информацию о событии. Для этого выполняется поиск и исключение из текстов документов фрагментов, содержащих идентичную информацию. С этой целью выполняется представление документов цепочками значимых слов, сравнение этих цепочек и обнаружение их схожих участков.

### 3.5 Анализ рассмотренных методов обнаружения дубликатов

У сигнатурных методов и алгоритмов, использующих шинглы, есть некоторые схожие черты: эти методы минимизируют вычислительную сложность операции сравнения документов. Поэтому они находят широкое применение в системах, работающих с гигантскими объемами данных (например, в поисковых системах). Обратной стороной медали является их узкая направленность – эти методы и модели представления текстов, которыми они оперируют (шинглы, сигнатуры), пригодны лишь для устранения дублей и не могут быть использованы для других задач. Однако, как было показано выше, очищенные от дубликатов данные впоследствии подвергаются разнообразной обработке и анализу. Поэтому представляется целесообразным применять для обнаружения дублей алгоритмы и модели, которые могут быть использованы для задач интеллектуальной обработки текстов.

Модель и метод, представленные в [16], также предназначены исключительно для решения конкретной задачи, а именно устранения идентичных фрагментов сообщений. Кроме того, для эффективного использования этого метода документы должны быть предварительно распределены по кластерам, соответствующим событиям. В нашей же ситуации устранение дублей, напротив, выполняется на этапе предварительной обработки данных перед использованием интеллектуальных аналитических методов, таких как обнаружение событий.

## 4 Предложенный подход к обнаружению дубликатов

В целях обеспечения возможности применения разрабатываемой модели документов для решения различных задач из области Text Mining, было решено использовать в качестве её основы векторное представление текстов. Однако использование слов документов в качестве признаков (1) позволяет сравнить сообщения лишь с точки зрения состава слов, что недостаточно для принятия правильного решения. Во многих предметных областях существенную роль играют и другие критерии (см. 2.2), и эти критерии должны быть включены в модель.

Таким образом, модель должна предусматривать возможность сравнения документов по различным признакам. Окончательно же решение должно приниматься на основе анализа пары документов с точки зрения всех критериев. Исходя из этого, удобно представить пару документов $(d_i, d_j)$ вектором (2), элементами которого являются результаты сравнения документов по соответствующим признакам:

$$\rho_{i,j} = (\rho_{i,j}^1, \rho_{i,j}^2, ..., \rho_{i,j}^k), \qquad (2)$$

где $\rho_{i,j}^k$ характеризует сходство документов $d_i$ и $d_j$ по $k$-му критерию. На основе этого вектора принимается решение о наличии или отсутствии дублирования. Для этого необходимо задать функцию, выполняющую интерпретацию вектора $\rho_{i,j}$, то есть определяющую вектор в один из двух классов, один из которых означает наличие дублирования (обозначим его $M_+$), другой – отсутствие ($M_-$):

$$D(\rho) = \begin{cases} 1, & \rho_{i,j} \in M_+; \\ 0, & \rho_{i,j} \in M_-. \end{cases} \qquad (3)$$

С учетом выбранного подхода, процесс обнаружения дублей можно разбить на следующие этапы (рис. 2):

1. Построение модели документа, отражающей характеристики новостного сообщения с точки зрения каждого из выбранных критериев.

2. Сравнение моделей двух документов и получение результирующего вектора $\rho_{i,j}$.

3. Интерпретация вектора с помощью решающей функции $D(\rho_{i,j})$.

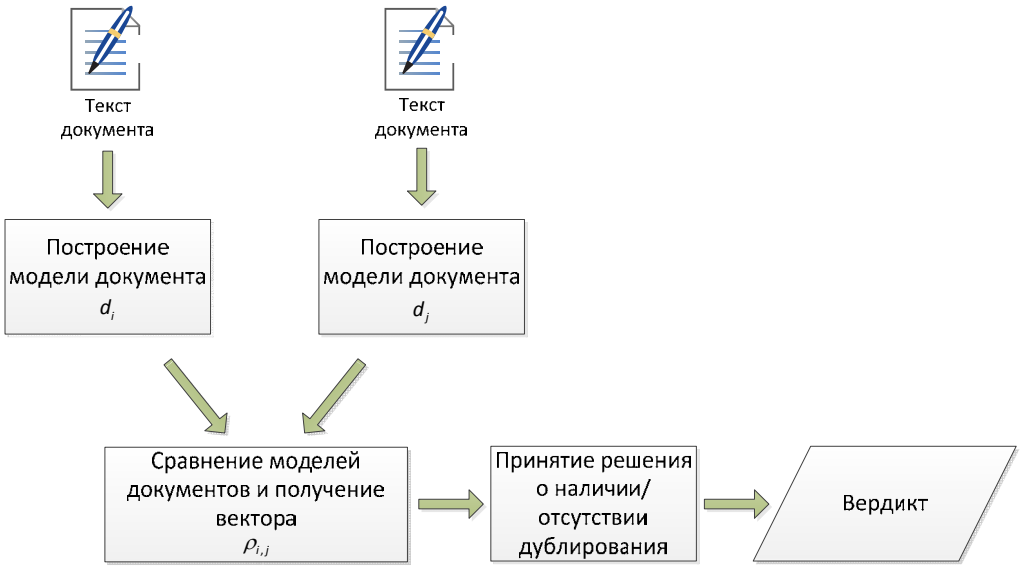



Рис. 2. Этапы обнаружения дубликатов

## 5 Модель документа

Важным этапом решения задачи является выбор критериев для сравнения документов. Набор критериев должен быть достаточно выразительным, чтобы обеспечивать возможность гибкой настройки модели в соответствии со спецификой различных предметных областей. Для определения набора критериев было проведено исследование выборки текстовых документов, автоматически собираемых из различных Интернет-источников.

В качестве источников были выбраны 35 сайтов по различной тематике: основные новостные сайты, публикующие материалы общественно-политической и экономической тематики, официальные сайты органов государственной власти РФ, некоторые сайты органов законодательной и исполнительной власти субъектов РФ. Такой набор сайтов позволил охватить значительное число тем и типов информационных сообщений (ленты новостей, аналитические статьи и обзоры, документы правового характера, документы, содержащие финансово-экономические показатели).

1200 пар документов были проанализированы экспертами вручную. В результате выполненного анализа были выявлены следующие критерии, характеризующие модель распознавания дубликатов.

### 5.1 Содержание текста

В большинстве случаев определяющее значение имеет близость текстов. Прежде всего, это характерно для общественно-политических новостей, вклад которых в суммарную информацию, передаваемую новостным потоком, определяется оригинальностью их текстового содержания.

Для обеспечения возможности сравнения составов слов документов, модель должна включать векторное представление текста сообщения (1). Если слово не встречается в документе, его вес равен нулю. Для остальных слов вес рассчитывается по методу TF-IDF с использованием алгоритма Okapi BM25 [6]. Рассчитанный таким образом вес слова $wt = tf * idf$ пропорционален частоте его употребления в документе $tf$ и обратно пропорционален частоте употребления слова в других документах коллекции $idf$. В результате, модель текста документа $d_i$ представляет собой вектор:

$$d_i^w = (w_i^1, w_i^2, ..., w_i^{N^w}), \quad (4)$$

где $N^w$ – общее количество различных слов во всех документах, $w_i^j$ – вес $j$-го слова в $i$-ом документе.

Векторное представление позволяет использовать для сравнения текстов простые алгебраические методы. В качестве меры сходства часто используют евклидову метрику – расстояние между двумя точками в многомерном пространстве, вычисляемое по теореме Пифагора. Однако она плохо подходит для сравнения документов, похожих по содержанию, но значительно различающихся по размеру. Поэтому при решении задач информационного поиска более распространен другой способ сравнения векторов, называемый косинусной мерой. Близость векторов оценивается на основании значения косинуса угла $\theta$ между ними, что позволяет не учитывать при сравнении длину векторов:

$$sim_{\cos}(d_i^w, d_j^w) = \frac{\sum_{n=1}^{N} w_i^n w_j^n}{\sqrt{\sum_{n=1}^{N} (w_i^n)^2} \sqrt{\sum_{n=1}^{N} (w_j^n)^2}}. \quad (5)$$

На первый взгляд, в целях обнаружения дубликатов лучше использовать евклидово расстояние, поскольку размер должен играть роль при сравнении документов: тексты существенно различающейся длины с очень низкой вероятностью являются нечеткими дубликатами. Однако для обеспечения большей гибкости системы было решено разделить оценку документов с содержательной и структурной точки зрения. Чтобы сделать оценку содержательной близости документов независимой от других характеристик, решено использовать для сравнения косинусную меру близости. При использовании неотрицательных весов слов косинусная мера принимает значения в интервале [0, 1], поэтому в качестве оценки различия векторов используется значение

$$\rho_{i,j}^{w} = 1 - sim_{\cos}(d_i^w, d_j^w). \quad (6)$$

### 5.2 Содержание заголовка

Отдельно при принятии решения учитываются заголовки сообщений, причем их роль значительно варьируется в зависимости от предметной области. При перепечатке новостей общественно-политической тематики с одного сайта на другой нередко изменяется только заголовок, причем иногда – весьма существенно. В таком случае наличие дублирования может быть обнаружено на основе близости остального текста новостей. Однако для сообщений другого типа (например, правового характера), различие заголовков имеет решающее значение, и такие документы не должны признаваться дубликатами, несмотря на близкое содержание.

В модели заголовок текста представляется аналогично основному тексту (4), с той лишь разницей, что в качестве элементов вектора используются слова, встречающиеся в заголовках документов коллекции: $d_i^t = (t_i^1, t_i^2, ..., t_i^{N^t})$, где $N^t$ – общее количество различных слов в заголовках всех документов, $t_i^j$ – вес *j*-го слова в заголовке *i*-го документа.

Сравнение составляющих моделей, отражающих содержание заголовков, также выполняется аналогично сравнению содержания документов (6):

$$\rho_{i,j}^{t} = 1 - sim_{\cos}(d_i^t, d_j^t).$$

### 5.3 Предложения и абзацы

Помимо оценки близости текстового содержания документов в целом, эксперт обращает особое внимание на наличие в сообщениях идентичных структурных элементов текстов – предложений и абзацев. Это связано с тем, что при перепечатке некоторым источником ранее опубликованного документа многие из этих элементов переносятся в текст-дубликат без изменений.

Для сравнения структурных элементов текста документ представляется множествами своих предложений $d_i^c = (c_i^1, c_i^2, ..., c_i^{N_i^c})$ и абзацев $d_i^p = (p_i^1, p_i^2, ..., p_i^{N_i^p})$, где $N_i^c$ и $N_i^p$ – количество предложений и абзацев в *i*-ом документе. В свою очередь, каждое предложение представляет собой последовательность слов, а потому может быть представлено векторной моделью $c_i^j = (c_i^{j,1}, c_i^{j,2}, ..., c_i^{j,N^w})$, где $c_i^{j,k}$ – вес *k*-го слова в *j*-ом предложении *i*-ого документа. Аналогичным образом представлен каждый абзац документа: $p_i^j = (p_i^{j,1}, p_i^{j,2}, ..., p_i^{j,N^w})$.

Поскольку структурные элементы представлены множествами, для определения их сходства можно использовать коэффициент Жаккара. Так, для предложений близость документов будет равна

$$sim_j^c(d_i^c, d_j^c) = \frac{\left| d_i^c \cap d_i^c \right|}{\left| d_i^c \cup d_i^c \right|}. \quad (7)$$

Такая мера близости принимает во внимание лишь количество совпадающих и различающихся предложений, но не учитывает, какие именно предложения совпадают и различаются. Однако совпадение значимых, содержательных предложений должно иметь больший вес, чем одновременное появление в обоих документах одинаковых коротких и незначительных фраз. В связи с этим вместо количества предложений используется их суммарный вес. Вес каждого предложения рассчитывается как сумма весов составляющих его слов.

Кроме того, представленная мера является симметричной, и потому она плохо подходит для сравнения документов, один из которых получен из другого путем удаления нескольких предложений: в этом случае первое сообщение содержит дополнительную информацию относительно второго, но второе не имеет оригинальных данных относительно первого. Чтобы учесть требуемую несимметричность, было решено использовать меру включения вместо меры сходства:

$$sim_{inc}^c(d_i^c, d_j^c) = \frac{\sum_{c \in \left| d_i^c \cap d_i^c \right|} [\sum_{k=1}^{N^w} c^k]}{\sum_{c \in d_i^c} [\sum_{k=1}^{N^w} c^k]}. \quad (8)$$

Для оценки различия документов с точки зрения предложений используются значения $\rho_{i,j}^c = 1 - sim_{inc}^c(d_i^c, d_j^c)$ и $\rho_{j,i}^c = 1 - sim_{inc}^c(d_j^c, d_i^c)$.

Аналогичным образом выполняется сравнение абзацев, результатом которого являются значения меры различия $\rho_{i,j}^p$ и $\rho_{j,i}^p$.

### 5.4 Числовые данные

Кроме документов, содержащих только текстовую информацию, часто встречаются и те, которые включают числовые данные. В ряде случаев даже незначительное изменение этих данных может существенно повлиять на содержание сообщения. Примером таких документов являются новостные сообщения из области экономики (новости о со-

стоянии фондового рынка) или спорта (сообщения о результатах соревнований). Такие документы не должны признаваться дубликатами даже при полном совпадении их текста.

Для сравнения документов с точки зрения числовых значений каждое сообщение представляется набором чисел, извлеченных из его текста: $d_i^n = \{n_i^1, n_i^2, ..., n_i^{N_i^{n_d}}\}$, где $N_i^{n_d}$ – количество различных чисел в *i*-ом документе. Для выполнения оценки сходства таких наборов также используется мера включения, однако элементы сравниваемых множеств не являются взвешенными, а потому учитывается лишь количество одинаковых и различающихся числовых значений:

$$sim_{inc}^n(d_i^n, d_j^n) = \frac{\left|d_i^n \cap d_i^n\right|}{\left|d_i^n\right|}. \qquad (9)$$

Расстояние между документами по данному критерию равно $\rho_{i,j}^n = 1 - sim_{inc}^n(d_i^n, d_j^n)$.

Для некоторых предметных областей важен не только состав набора чисел, но и порядок их следования в тексте документа. Это относится, в частности, к спортивным новостям, где разные последовательности одних и тех же чисел могут соответствовать различным результатам соревнований (например, два сета в теннисном матче, завершившиеся со счетом «6:4» и «4:6»). В таком случае для представления сообщения используется кортеж $d_i^n = \{n_i^1, n_i^2, ..., n_i^{N_i^{n_a}}\}$, где $N_i^{n_d}$ – общее количество чисел в *i*-ом документе.

В этом случае для сравнения документов необходимо выбрать меру различия, учитывающую порядок следования элементов. Такой мерой является расстояние Дамерау–Левенштейна [3], равное количеству операций вставки, удаления, замены и перестановки элементов, необходимых для преобразования одной последовательности символов (в данном случае – чисел) в другую. Эта мера является модификацией расстояния Левенштейна, отличающаяся наличием операции перестановки двух соседних символов (транспозиции). Это важно для нашей задачи, поскольку при перепечатке документа иногда изменяется порядок следования его абзацев и предложений, что приводит к появлению перестановок в последовательности чисел.

Расстояние между сообщениями при использовании этой меры равно $\rho_{i,j}^n = dist_{DL}(d_i^n, d_j^n)$ и является симметричным.

### 5.5 Фотографии и ссылки

Информация в сообщении может быть представлена не только текстом или числовыми данными, но и различными объектами, включенными в текст документа - фотографиями, видеороликами, ссылками на сторонние источники. Присутствие в документе дополнительных фото- и видеоматериалов значительно повышает вероятность его оригинальности

При сравнении фотоматериалов, включенных в сообщение, возникают сложности: определить идентичность фотографий в двух документах проблематично, поскольку одинаковые с точки зрения эксперта фотографии могут иметь различные URL и разный размер. Поэтому было принято решение учитывать не сами фотографии, а их количество в документе: $d_i^{im}$. Также в модель включается компонент, отражающий количество ссылок, присутствующих в тексте сообщения: $d_i^h$. Различие документов с точки зрения этих критериев определяется как разность соответствующих значений: $\rho_{i,j}^{im} = d_i^{im} - d_j^{im}$, $\rho_{i,j}^h = d_i^h - d_j^h$.

### 5.6 Дата и время публикации

Существенным фактором, влияющим на принятие решения о наличии или отсутствии дублирования, является разница во времени публикации сообщений. Так, при дублировании новостных статей перепечатыванию обычно подвергаются свежие новости, недавно опубликованные на сайте первоисточника. С увеличением интервала между моментами появления документов в сети вероятность дублирования быстро убывает

В модели эта характеристика сообщения представлена посредством POSIX-времени момента публикации (которое определяется как количество секунд, прошедших с полуночи 1 января 1970 года до момента, когда документ был опубликован источником): $d_i^{dt}$. Различие между документами определяется как разность между моментами публикации сообщений в секундах: $\rho_{i,j}^{dt} = d_i^{dt} - d_j^{dt}$.

### 5.7 Авторитетность источника

При анализе документов эксперты обращают внимание на источники сообщений, при этом они руководствуются своими представлениями об авторитетности источников. Статья из авторитетного источника (который обычно публикует оригинальные материалы) имеет существенно меньшую вероятность быть признанной дубликатом, чем документ, полученный из источника, регулярно занимающегося перепечаткой чужих сообщений.

Авторитетность источника *s* представляется значением $aut(s) \in [0,1]$, отражающим вероятность публикации им оригинального сообщения. Это значение может быть задано экспертом вручную или получено на основе обучающей выборки как соотношение количества оригинальных документов, поступивших от источника, к общему количеству опубликованных им сообщений.

В модель документа включается компонент, отражающий авторитетность источника, опубликовавшего этот документ: $d_i^a = aut(src(d_i))$, где $(src(d_i)$ – функция, устанавливающая соответствие между документом и его источником.

Таким образом, модель документа представляет собой совокупность компонентов, характеризующих сообщение с точки зрения различных критериев:

$$d_i = (d_i^w, d_i^t, d_i^n, d_i^c, d_i^p, d_i^{im}, d_i^h, d_i^{dt}, d_i^a). \quad (10)$$

# 6 Метод обнаружения дубликатов

## 6.1 Интерпретация результата сравнения

После получения моделей $d_i$ и $d_j$ двух документов необходимо сравнить их и вынести решение о том, являются ли документы нечеткими дубликатами. Для этого выполняется анализ схожести сообщений по каждому из критериев. Результатом сравнения документов с точки зрения некоторого критерия $k$ является значение $\rho_{i,j}^k$. Выполнив попарно сравнение компонентов моделей для каждого из критериев, получим вектор

$$\rho_{i,j} = (\rho_{i,j}^w, \rho_{i,j}^t, \rho_{i,j}^n, \rho_{i,j}^c, \rho_{i,j}^p, \rho_{i,j}^{im}, \rho_{i,j}^h, \rho_{i,j}^{dt}, \rho_{i,j}^a) \quad (11)$$

Ввиду вышеуказанной несимметричности мер сходства, используемых для некоторых критериев, результатом сравнения двух документов являются два различных вектора $\rho_{i,j}$ и $\rho_{j,i}$, характеризующие степень отличия первого и второго сообщения друг от друга. Каждый из этих векторов интерпретируется с помощью функции $D(\rho)$ (3).

Для интерпретации результата сравнения необходимо решить задачу бинарной классификации, то есть отнести вектор к классу $M_+$ или $M_-$. Для настройки параметров классификатора используется обучающая выборка – набор векторов, каждый из которых снабжен меткой $m \in \{M_+, M_-\}$, обозначающей класс, к которому принадлежит этот вектор.

Задача бинарной классификации состоит в том, чтобы для вновь поступившего на исследование вектора $\rho = (\rho^1, \rho^2, ..., \rho^K)$ определить класс, к которому он принадлежит, то есть значение $m$. Для её решения будем использовать метод опорных векторов (SVM). Этот метод основан на построении в $K$-мерном пространстве $(K-1)$-мерной гиперплоскости, разделяющей объекты классов $M_+$ и $M_-$. В зависимости от расположения вектора $\rho$ относительно этой гиперплоскости, выполняется его отнесение к одному из классов.

Возможны следующие результаты интерпретации:

- $D(\rho_{i,j}) = D(\rho_{j,i}) = 0$. В этом случае оба документа признаются оригинальными;
- $D(\rho_{i,j}) \neq D(\rho_{j,i})$. Один из документов является оригиналом, а второй – дублем;
- $D(\rho_{i,j}) = D(\rho_{j,i}) = 1$. Оба документа являются дублями друг относительно друга. То есть, ни один из них не содержит оригинальных данных относительно другого.

На основе полученных результатов принимается решение о дальнейших действиях в отношении документов. Так, в разработанной системе решалась задача проверки документов в момент их поступления от источника, при этом загружаемые сообщения сравнивались на предмет дублирования с документами, уже загруженными в базу. Поэтому интерес представлял лишь один из результатов интерпретации – является ли загружаемый документ дубликатом ранее полученного сообщения. Однако при обработке готовой коллекции документов с целью обнаружения и устранения дубликатов важно выявить все пары сообщений, в которых имеет место дублирование, для чего требуется использовать оба результата.

## 6.2 Метод предварительного отбора кандидатов

Предложенный метод выявления нечетких дубликатов имеет существенный недостаток – высокую вычислительную сложность. Каждое новое сообщение подвергается сравнению со всеми ранее загруженными, и при каждом сравнении выполняется расчет близости документов по множеству критериев. Однако очевидно, что в большинстве случаев в таком тщательном анализе нет необходимости – сильно различающиеся по тексту документы с высокой вероятностью различны и по содержанию. Следовательно, нужно исключать из рассмотрения те из ранее загруженных новостей, которые слишком сильно отличаются от текущей.

С этой целью вышеописанный метод предваряется процедурой отбора документов, дубликатом которых может быть текущая новость (то есть, отбора кандидатов на роль оригинала этой новости). Эта процедура, по сути, также решает задачу обнаружения дубликатов, причем основными требованиями, предъявляемыми к ней, являются минимальная вычислительная сложность и максимальная полнота (поскольку отброшенные из числа кандидатов документы далее рассматриваться не будут).

В качестве такой процедуры рассматривались представленные в работе [12] приближенные методы обнаружения дубликатов, имеющие высокую производительность. Ввиду наличия набора взвешенных слов, было решено использовать для описания сообщения сигнатуру, представляющую собой строку, состоящую из сцепленных в алфавитном порядке нескольких наиболее «тяжелых» слов документа. При этом процедура отбора кандидатов заключается в выборе из ранее загруженных документов тех, которые имеют такую же сигнатуру, как и текущая новость. Такие пары документов с совпадающими сигнатурами должны быть подвергнуты проверке основным методом, представленным в предыдущем подразделе. В работе [12] предложено использовать сигнатуры из 6 слов, но это приводит к низкому значению полноты (0.54). В целях получения высокой полноты, было решено сократить количество слов, составляющих сигнатуру, до двух.

## 7 Экспериментальная проверка метода

В рамках данной работы были проведены эксперименты, направленные на анализ качества работы разработанной подсистемы обнаружения дубликатов, реализующей предложенный метод. Все эксперименты проводились на ПЭВМ со следующими основными параметрами: процессор Intel Core 2 Duo 2,2 ГГц, объем ОЗУ 2 Гб.

Целью первого эксперимента была оценка качества метода предварительного отбора потенциальных дубликатов на примере анализа общественно-политических новостей. Тестирование производилось в течение суток. В качестве входных данных использовались 1502 документа, извлеченных с 20 новостных сайтов. Каждый из документов подвергался сравнению с 10293 загруженными ранее сообщениями. В общей сложности было выполнено 16 586 310 сравнений документов, при этом 259 пар были отобраны для проверки основным методом.

Таблица 1. Оценка качества метода отбора кандидатов

| | $N_p$ | $N_{or}$ | $N_{dup}$ |
|---|---|---|---|
| Всего | 16 586 310 | 16 586 121 | 189 |
| Прошли отбор | 259 | 108 | 151 |
| Отброшено | 16 586 051 | 16 586 013 | 38 |

Где $N_p$ – общее количество пар, $N_{or}$ – число пар, элементы которых не дублируют друг друга, и $N_{dup}$ – количество пар документов-дубликатов.

Из 189 пар дубликатов отбор прошла 151 пара (80%). Таким образом, использование сигнатуры из двух слов позволяет увеличить полноту по сравнению с шестисловными сигнатурами, однако добиться полноты, близкой к 100%, не удалось. Метод часто отбрасывает дубликаты в случаях, когда один из документов является урезанной копией другого – отсутствие нескольких параграфов значительно влияет на веса слов. Анализ результатов экспериментов показывает необходимость доработки метода предварительного отбора.

Отброшенные пары не используются для обучения и тестирования основного метода, однако было обнаружено, что 38 ошибочно отброшенных пар дубликатов по своим характеристикам близки к тем 151, которые прошли отбор. Таким образом, недостаточная полнота метода предварительного отбора ведет к появлению в коллекции большего количества дублирующихся сообщений, но не снижает качество обучения основного метода.

Среди всех 259 пар, прошедших отбор, дубликаты составляют 58%. Столь низкая точность метода доказывает необходимость дополнительного анализа отобранных пар. При этом метод продемонстрировал высокую эффективность (под эффективностью понимается отношение количества пар, отброшенных на этапе предварительного отбора, к общему числу пар): из 16 586 310 пар документов было отброшено 16 586 051 (0,99998%).

В рамках второго эксперимента выполнялась оценка качества основного метода. Для тестирования использовались 26 036 документов, извлеченных с 20 новостных сайтов. С помощью метода фильтрации было отобрано 2650 пар-кандидатов, каждая из которых была проанализирована экспертами на предмет наличия дублирования. Часть пар использовалась для обучения, на остальных выполнялось тестирование метода. Целью эксперимента было определение зависимости показателей качества (точности, полноты и *F*-меры) от мощности обучающей выборки и от учитываемых критериев.

Полученные зависимости представлены на рис. 3 (*а* – при использовании только близости составов слов, *б* – при использовании только схожести параграфов, *в* – при использовании всех критериев, приведенных в разделе 5).

Как видно из рисунка, при использовании 400 обучающих примеров происходит насыщение, и с дальнейшим увеличением обучающей выборки качество работы метода не улучшается. Таким образом, для обучения системы достаточно 400 пар документов, размеченных экспертами.

Проведенный эксперимент доказывает целесообразность многокритериального сравнения документов: при использовании всех критериев достигаются более высокие показатели качества (F-мера в зоне насыщения равна 0,82), чем при анализе документов только с точки зрения слов (0,67) или параграфов (0,64).

При анализе результатов эксперимента было выявлено несколько факторов, негативно сказывающихся на качестве. Одним из них является человеческий фактор: каждый эксперт, принимавший участие в подготовке обучающей и тестовой выборок, имеет свое представление о том, какие документы являются информативно необходимыми, а какие – нет, в результате чего возникают конфликты в суждениях экспертов. Также эксперимент показал, что система не может обнаруживать дублирование в случае переписывания оригинального текста без изменения его содержания (рерайтинга), что говорит о необходимости доработки модели для обнаружения такого рода дублирования. Наконец, при проведении эксперимента была выполнена попытка настройки единой модели распознавания для всех документов, загружаемых с новостных сайтов. Но эти документы принадлежат различным предметным областям – среди общественно-политических новостей попадаются экономические, спортивные, юридические. Для повышения качества работы эти документы должны анализироваться с использованием специализированных моделей распознавания.

Проведенный эксперимент также показал, что среднее время, затрачиваемое на анализ пары документов основным методом, составляет 5 мс. С учетом высокой эффективности метода предварительной фильтрации это означает, что система способна обрабатывать 10 000–50 000 документов в час (в зависимости от количества загруженных ранее сообщений, с которыми требуется сравнивать новые документы).

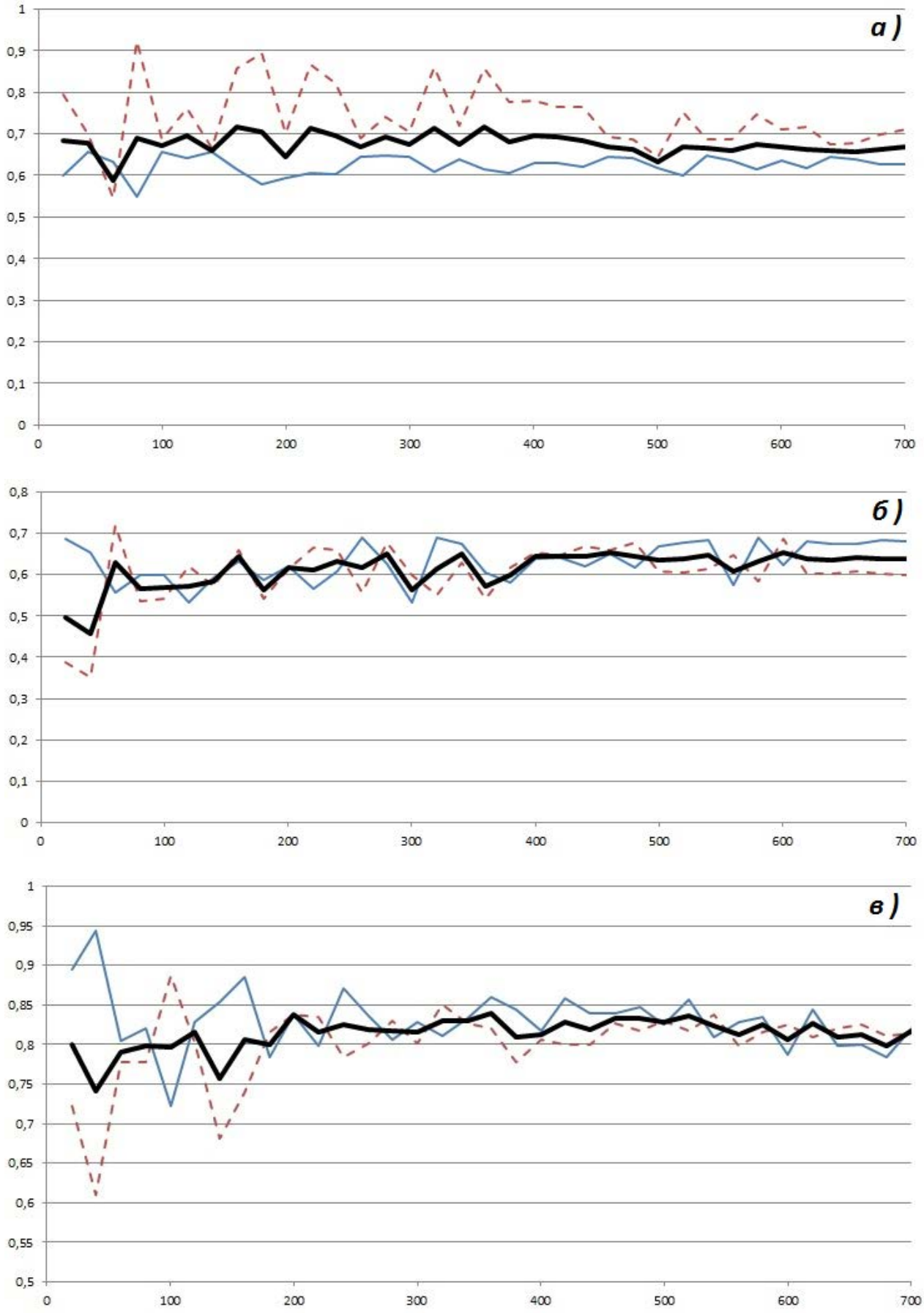



Рис. 3. Зависимость точности (тонкая сплошная линия), полноты (тонкая пунктирная линия) и *F*-меры (жирная линия) от мощности обучающей выборки

## 8 Направления дальнейших исследований

Помимо обнаружения и устранения дубликатов, предлагаемый метод может быть использован для решения других задач интеллектуального анализа текстов. При соответствующей настройке набора учитываемых критериев и порога близости документов, необходимого для вынесения решения о наличии дублирования, метод может быть применен для решения общей задачи обнаружения документов, близких по содержанию к заданному. Это позволит, в частности, выполнять формирование подборок тематически близких документов, а также сообщений, которые с большой долей вероятности связанны с каким-то общим событием. Таким образом, имеется возможность использования разработанного метода для решения задачи динамической кластеризации коллекции документов.

Еще одним перспективным направлением развития метода является снабжение его возможностью не только обнаружения наличия или отсутствия дублирования, но и выделения в тексте близких по содержанию сообщений фрагментов с оригинальной (недублированной) информацией.

## 9 Заключение

В работе предложен метод обнаружения и устранения нечеткого дублирования в потоке текстовых сообщений. В его основе лежит отнесение пар документов к классу «дубликатов» или «недубликатов» с помощью метода опорных векторов.

Предлагаемый метод обладает высокой гибкостью благодаря возможности его настройки для обработки сообщений из различных предметных областей. Это достигается посредством включения в модель документа компонентов, отражающих критерии, которыми руководствуются эксперты при анализе текстовых коллекций вручную.

Для обеспечения низкой вычислительной сложности предложена процедура отбора пар-кандидатов на основе сравнения числовых сигнатур документов. Это позволяет применять основной метод лишь к документам, прошедшим отбор.

Представленный метод был апробирован при решении задачи анализа потока текстовых сообщений, загружаемых из открытых интернет-источников, с целью устранения документов, являющихся дубликатами ранее загруженных материалов.

## Литература


[1] M. Bilenko, R.J. Mooney. Adaptive Duplicate Detection Using Learnable String Similarity Measures. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003), Washington DC, pp. 39–48, August, 2003.

[2] A. Broder. Algorithms for duplicate documents. http://www.cs.princeton.edu/courses/archive/spr05/cos598E/bib/Princeton.pdf

[3] Damerau, F. 1964. A technique for computer detection and correction of spelling errors. Communications of the ACM 7, 3 (1964), 171–176.

[4] D. Fetterly, M. Manasse, M. Najork. A Large-Scale Study of the Evolution of Web Pages, WWW2003, May 20–24, 2003, Budapest, Hungary.

[5] U. Manber. Finding Similar Files in a Large File System. Winter USENIX Technical Conference, 1994.

[6] Robertson S., Walker S., Jones S., Hancock M.-Beaulieu, M. Gatford. Okapi at trec-3. The Third Text REtrieval Conference (TREC-3), 1995.

[7] Андреев А.М., Березкин Д.В., Симаков К.В. Модель извлечения фактов из естественно-языковых текстов и метод ее обучения // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды 8-й Всероссийской научной конференции (RCDL’2006). – Суздаль, 2006.

[8] Андреев А.М., Березкин Д.В., Морозов В.В., Симаков К.В. Метод кластеризации документов текстовых коллекций и синтеза аннотаций кластеров // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды 10-й Всероссийской научной конференции (RCDL’2008). – Дубна, 2008. – С. 220–229.

[9] Андреев А.М., Березкин Д.В., Козлов И.А., Симаков К.В. Метод обнаружения изменений структуры веб-сайтов в системе сбора новостной информации // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды 14-й Всероссийской научной конференции (RCDL-2012). – Переславль-Залесский, 2012. – С. 124–133.

[10] Блох М.Я. Теоретические основы грамматики : учебник. – 2-е изд., исправл. – М. : Высш. шк., 2000. – 160 с.

[11] Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие. – М. : МИЭМ, 2011. – 272 с.

[12] Зеленков Ю.Г, Сегалович И.В. Сравнительный анализ методов определения нечетких дубликатов для Web-документов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды 9-й Всероссийской научной конференции (RCDL’2007). – Переславль-Залесский, 2007. – С. 166–174.

[13] Князева А.А., Турчановский И.Ю., Колобов О.С. Выявление дубликатов в библиографических базах данных // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды 15-й Всероссийской научной конференции (RCDL2013). – Ярославль, 2013. – С. 276–282.

[14] Ландэ Д.В., Дармохвал А.Т., Морозов А.Ю. Подход к выявлению дублирования сообщений в новостных информационных потоках // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды 8-ой Всероссийской научной конференции (RCDL2006). – Суздаль, 2006.

[15] Лифшиц Ю. Метод опорных векторов. Курс лекций «Алгоритмы для Интернета», 2006.

[16] Никконен А.Ю. Устранение избыточности и дублирования сюжетов новостных сообщений // Сборник работ участников конкурса «Интернет-математика 2007».

[17] Шевелёв О.Г. Методы автоматической классификации текстов на естественном языке : учеб. пособие. – Томск : ТМЛ-Пресс, 2007. – 144 с.

## The Method of Detecting Duplicates in a Stream of Text Documents

A. Andreev, D. Berezkin, I. Kozlov, K. Simakov

The problem of duplicate documents elimination from a stream of text messages is considered. A multicriterion model of text document is given. Criteria are chosen to properly represent documents from different domains. An approach for duplicates detection based on binary classification is proposed. A method of candidates preliminary filtration is proposed in order to reduce the computational complexity of the approach.

# Подходы к реализации межведомственного обмена распределенными и разнородными данными на примере ЕСИМО

© Е.Д. Вязилов © Н.Н. Михайлов © А.Е. Кобелев © Д.А. Мельников
ФГБУ «ВНИИГМИ-МЦДЦ» Росгидромета, Обнинск
vjaz@meteo.ru nodc@meteo.ru kobelev@meteo.ru melnikov@meteo.ru

**Аннотация**

Рассмотрены основные подходы, как в области интеграции данных, так и обеспечении межведомственного информационного взаимодействия (обмена данными) в области морской среды и морской деятельности. Реализация рассмотрена на примере Единой государственной системы информации об обстановке в Мировом океане. Для интеграции данных не требуется специальных преобразований данных в их источниках, достаточно только представить метаданные на интегрируемые информационные ресурсы. За счет использования единого словаря параметров и стандартизованных классификаторов пользователь может получить ресурсы в необходимом составе, форматах хранения или в виде таблицы, графика, карты.

## 1 Введение

В последние годы в России началась массовая разработка «единых систем», например, Единая информационная система Федеральной службы по надзору в сфере связи, информационных технологий и массовых коммуникаций, Единая государственная система предупреждения и ликвидации чрезвычайных ситуаций, Единая государственная автоматизированная система контроля радиационной обстановки на территории Российской Федерации, Единая система государственного учета результатов научно-исследовательских, опытно-конструкторских и технологических работ гражданского назначения, Единая государственная система экологического мониторинга, Единая государственная система координат и др. [3, 6].

Похожие работы ведутся и за рубежом. Так уже около 10 лет реализуются панъевропейские проекты по интеграции данных в области океанографии (http://www.seadatanet.org) [12], метеорологии [9]. Создается мировая интегрированная информационная система в области исследования окружающей среды GEOS [13]. Имеется пример интеграции наблюдательных систем Integrated Ocean Observing System (http://www.iooc.us/ocean-observations/integrated-ocean-observing-system/). Безусловно, есть реализации и в других предметных областях. Хорошие обзоры состояния исследований и разработок в области интеграции данных представлены в работах [4, 7, 10].

Во всех этих системах сбор данных ведется из распределенных и неоднородных источников. В основном используются подходы, связанные со стандартизацией обменных форматов и приведением данных к единой централизованной базе данных (БД) [5]. Это традиционный способ интеграции – Extraction Transformation Loading (ETL). При небольшом числе обменных форматов или количестве источников данных этот подход вполне эффективен. Применяется также подход, связанный с организацией единого реестра ссылок [13] на различные приложения или наборы данных. Дальнейшее повышение производительности труда администраторов БД напрямую связано с автоматизацией обмена, актуализацией и обеспечением целостности БД.

Для организации межведомственного взаимодействия во многих существующих системах, претендующих на «единые системы», не решены такие задачи как создание:

– метаданных по имеющимся источникам данных в стандартах ИСО 1915, 1939 и др.;

– БД единых справочников, классификаторов (международных, национальных, ведомственных, локальных);

– средств автоматического маппирования локальных имен параметров в общесистемные и различных систем кодирования данных;

– средств интеграции разнородных и распределенных данных для организации обмена между программными приложениями (чтобы можно было автоматически получить данные из одной системы (БД) и автоматически использовать их в другом приложении).

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

В области исследования морской среды имеется 50-летний опыт сбора океанографических данных от различных ведомств, ее обработки и международного обмена ими и на современном этапе развития информационных технологий существует необходимость взаимодействия на уровне национальных и зарубежных, в т.ч. международных систем. Кроме того, для управления морской деятельностью и комплексного информационного обеспечения федеральных органов исполнительной власти имеется насущная необходимость как в интеграции данных, так доведения созданной информационной продукции до различных уровней управления.

Введенная в эксплуатацию в 2013 г. Единая государственная система информации об обстановке в Мировом океане (ЕСИМО, http://esimo.ru), интегрирующая разнородные, распределенные информационные ресурсы (ИР) 12 министерств и ведомств России (37 поставщиков данных, более 3500 ИР), предназначена для интеграции информации об обстановке в Мировом океане и комплексного информационного обеспечения морской деятельности. В представленном докладе рассмотрены основные подходы, как в области интеграции данных, так и обеспечении межведомственного информационного взаимодействия (обмена данными).

## 2 Основные подходы

Основными подходами по реализации ЕСИМО являются создание единого словаря параметров [2], маппирование классификаторов, расширение состава атрибутов метаданных, имеющихся в стандарте ИСО 19115.

Создание и поддержка единого словаря параметров системы позволяет привести к единой системе наименований все используемые в ИР атрибуты. Особенно это касается атрибутов метаданных, наиболее широко применяемых в различных ИР. При этом появляются возможности объединения и слияния различных ресурсов, например, объединение данных наблюдений и сведений о наблюдательных платформах для совместной визуализации. Таким образом, единый словарь параметров обеспечивает унификацию разнообразных имен атрибутов во всех программных компонентах ЕСИМО.

В организациях – авторах ИР применяются классификаторы различного уровня стандартизации. Для приведения их единым, принятым в ЕСИМО классификаторам, необходимо маппирование локальных классификаторов в международные и национальные. Это позволяет использовать любую нотацию кодирования без какого-либо влияния на информационное взаимодействие. При этом пользователь может всегда вернуться к локальной нотации классификатора.

Общие коды и классификаторы обеспечивают раскодирование значений метаданных и данных при внешнем представлении данных, унификации информационных, программных интерфейсов пользователя. Список необходимых для описания ИР классификаторов, включает следующие (уровень обработанности данных, пространственное представление, частота обновления ИР, частота повторения данных в ИР, форма представления ресурса, тип доставки, организации, страны, океаны и моря, города, субъекты РФ, суда, прибрежные станции, морской деятельности, методы наблюдений, определений, обработки данных). В данных также используются некоторые классификаторы из вышеперечисленных. Кроме того, есть такие классификаторы, как сплоченность льда, формы облаков, балл волнения и другие, всего используется около 400 классификаторов и нотаций кодирования.

Классификация ИР по уровню обработанности данных (наблюденные, диагностические, прогностические, обобщенные – агрегированные), форме представления данных (точка, профиль, сетка, объектные файлы), системе хранения (БД, структурированные файлы, сервисы, каталоги объектных файлов) позволяет не только улучшить поиск ИР, но и более эффективно организовать визуализацию данных.

Включение в описания ИР формализованных сведений о типе платформы измерений, пространственно-временных масштабов представления данных позволяет в дальнейшем построить различные шаблоны визуализации данных.

Поименованная совокупность данных, генерируемых источником от локальной системы, после применения к ним операций (описание метаданных и нормализация данных) становится ИР. Информационный ресурс – это структурированные (БД и файлы) или неструктурированные данные (документ, совокупность документов), предназначенные и оформленные для распространения среди неограниченного круга лиц, либо служащие основой для предоставления информационных услуг. Признаками ИР являются однородность структуры данных, нахождение в одном источнике (одном носителе), один пространственно-временной масштаб разрешения данных (например, срочные данные на фиксированных прибрежных станциях, случайные измерения во времени и пространстве).

Информационная база предметной области и составляющие ее ИР обладают такими свойствами как идентификация, содержание, производство, происхождение, связность, ограничение доступа, жизненный цикл ИР, рис. 1.

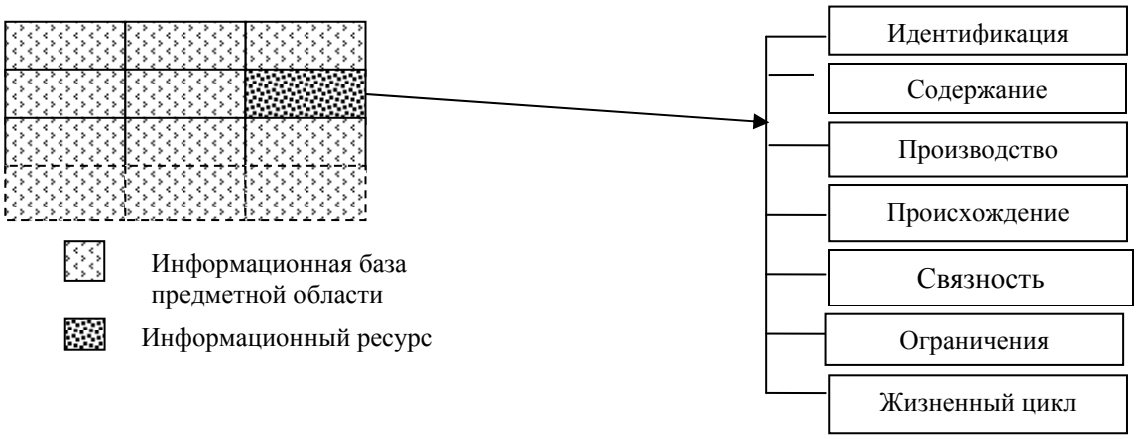



Рис. 1. Категории свойств информационных ресурсов

ИР представляет собой единицу набора данных (часть или вся таблица или несколько таблиц БД, отдельный файл), имеющую уникальность по таким свойствам, как идентификатор ресурса (физический и логический), тематика ресурса (параметры), тип ресурса (данные наблюдений, диагностические, прогностические и обобщенные данные), пространственно-временное разрешение данных, форма представления (буквенно-цифровая, структурированная, текстовая, графическая, пространственная), система хранения данных (БД, структурированные или объектные файлы, приложение, включая картографические сервисы).

Для управления данными введено понятие экземпляр ИР. В зависимости от тематики, объемов данных ИР может представлять один или несколько экземпляров, количество экземпляров зависит от принятой иерархии данных. Например, для океанографических данных экземпляром могут быть: данные, измеренные на одном горизонте; данные, полученные на одном профиле (океанографическая станция, выполненная в одной точке в один момент времени); данные, собранные за весь период рейса научно-исследовательского судна (единица сбора данных).

Метаданные включают описание источников информации, средств управления и обработки данных. Источники информации представлены более детально отдельными описаниями наблюдательных сетей, приборов, экспедиций, проектов, мореведческих организаций, научно-исследовательских судов, прибрежных станций и постов на морях России, космических аппаратов, экспертов, которые также оформлены как ИР.

Схема метаданных за счет использования классификаторов и единого словаря параметров является универсальной – независимой от предметной области; достаточно выразительной для понимания структуры данных; совместимой с международными стандартами и протоколами в области метаданных и информационного поиска.

При этом максимально используются стандарты интероперабельности в области метаданных ИСО 19115, стандарты ИСО серии 19100, стандарты Open Geospatial Consortium.

Протокол обмена определяет форматы и механизмы обмена данными между компонентами технологии и состоит из сообщения - запроса, сообщения – ответа (XML формат) и обменного файла данных в международном формате NetCDF (двоичный файл).

Эти подходы обеспечивают решение задач управления разнородными ИР посредством единообразного доступа ко всем ресурсам и использования поисковых атрибутов для разных форм представления ИР.

## 3 Схема описания информационных ресурсов

Описание ИР строится из блоков, включающих один или несколько элементов и разделов (классов) стандарта ИСО 19115 и представляется в виде набора записей [1].

*Элемент* является неделимой частью описания в составе раздела описания, который может быть использован в различных разделах, обладает уникальным именем и типом представления (строковое, числовое и т.п.). Элемент может быть представлен как ключевым элементом (идентификатором), так и свойством.

*Класс* или раздел описывает набор однородных свойств ИР. Класс - это фиксированный набор элементов, скомпонованный по определенным правилам, задающим последовательность элементов и их повторяемость в пределах этой последовательности.

*Запись* — это композиция классов, отображающих описание ИР и сведения, необходимые для интеграции данных. Аналогично классам, записи строятся по определенным правилам встречаемости классов (необязательные, обязательные, множественные или единичные).

Выделен следующий список классов, отражающий свойства описываемого ИР: идентификация; временные характеристики;

географические характеристики; структурирование данных; сведения о системе кодирования; описание кода; описание элемента; вертикальное обобщение; информация о качестве данных; связь с источником данных/метаданных; сведения о распространении, включая сведения о транспортном файле данных; сведения о проекте; сведения об инструментах; сведения о наблюдательной платформе; дополнительная информация. Модель описания информационных ресурсов отражает данные, помещаемые в архив системы, так и хранимые, и распространяемые данные, как это требуется в соответствии с моделью Open Archival Information System.

## 4 Модель ЕСИМО

Роль и место ЕСИМО в информационном обеспечении морской деятельности представлена на рис. 2.


Рис. 2. Роль и место ЕСИМО в информационном обеспечении морской деятельности

Перечень функциональных задач ЕСИМО включает интеграцию разнородной и распределенной информации об обстановке в Мировом океане; обеспечение федеральных, региональных и местных органов власти Российской Федерации, организаций, осуществляющих морскую деятельность, комплексной информацией об обстановке в Мировом океане; взаимодействие с зарубежными информационными системами морской направленности, обеспечение доступа к их ресурсам. Модель ЕСИМО дана на рис. 3. ЕСИМО состоит из информационно-технологических узлов. Узлы ЕСИМО представляют собой IP-адресуемые сервера, обеспечивающие выполнение функциональных задач ЕСИМО. Сервера узлов созданы согласно типовому техническому решению и на компонентной основе.

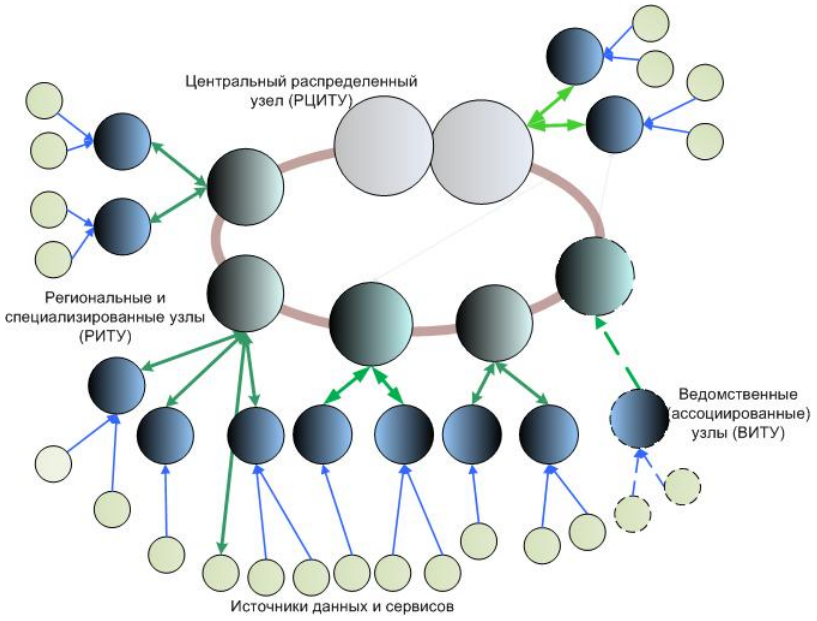



Рис. 3. Модель ЕСИМО

Состав компонентов узла зависит от выполняемых функций и зоны ответственности узла. Выделены следующие категории узлов ЕСИМО: ведомственный информационно-технологический узел (ВИТУ); специализированный информационно-технологический узел (СИТУ); региональный информационно-технологический узел (РИТУ); распределенный центральный информационно-технологический узел (РЦИТУ).

Операторы ВИТУ выполняют подготовку и загрузку информации в ЕСИМО по своей зоне ответственности, обеспечивают информационную безопасность и управление узлами. Они имеют возможность осуществлять обработку данных и анализ информации, информационное обслуживание с использованием средств вышестоящего узла, к которому присоединен ВИТУ.

Операторы РИТУ/СИТУ обеспечивают поддержку всех технологий ЕСИМО в применении к заданному региону или специализации (например, военная подсистема ЕСИМО). Узел взаимодействует с ведомственными узлами, осуществляет синхронизацию метаданных и обмен ресурсами с центральным узлом ЕСИМО, а также обслуживает потребителей информации об обстановке в Мировом океане согласно зоне ответственности на региональном уровне.

Оператор РЦИТУ ЕСИМО обеспечивает взаимодействие с внешними системами и обслуживание пользователей на федеральном уровне, а также осуществляет управление работой единой системы в целом. Узел взаимодействует с предписанными ВИТУ, а также со специализированными и региональными узлами для синхронизации метаданных и обмена ресурсами, осуществляет ведение и распространение кодов и классификаторов, условно-постоянных баз метаданных, единой электронной карты-основы.

## 5 Архитектура системы интеграции данных

Архитектура технологии интеграции данных отображена на рис. 4.

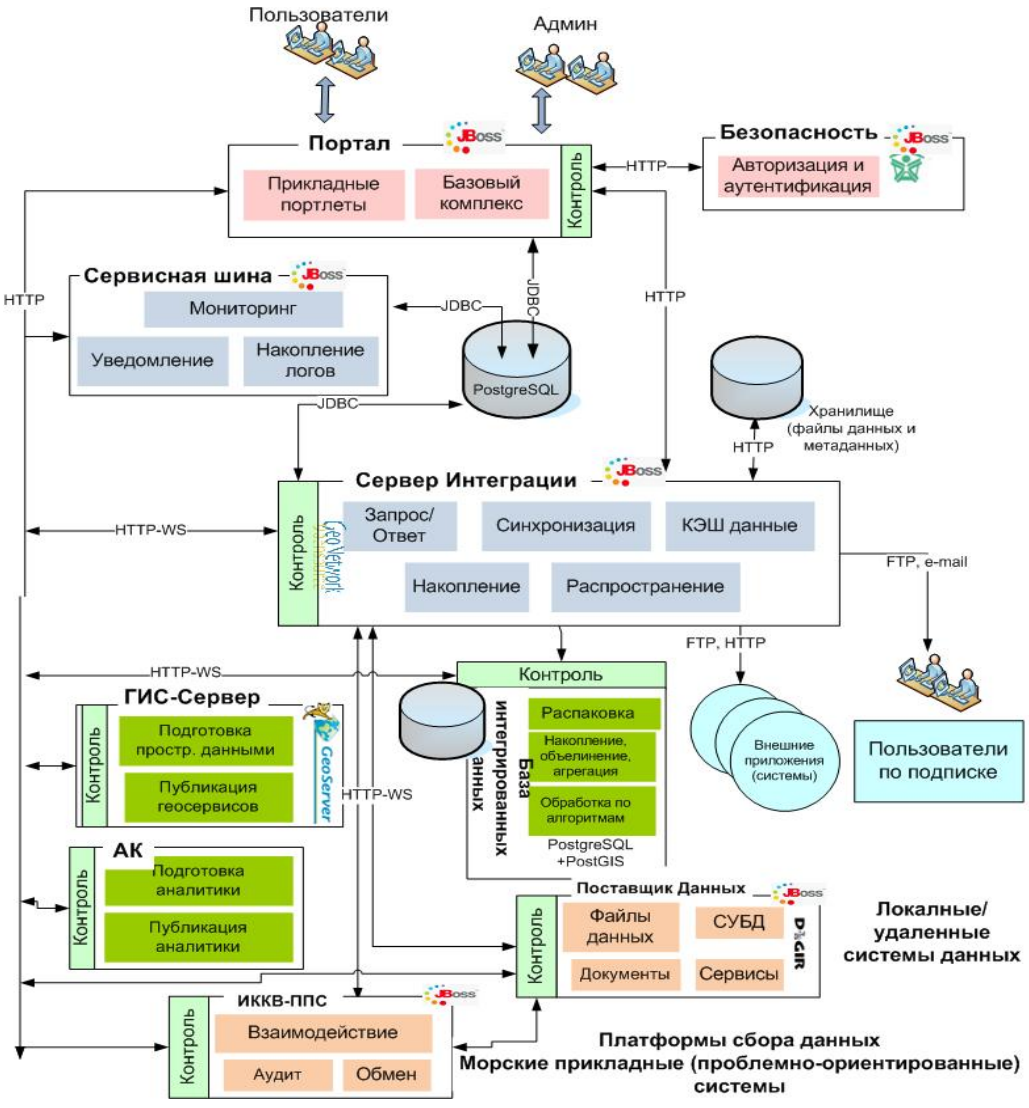



Рисунок 4. Архитектура технологии интеграции ЕСИМО

Программный комплекс «Поставщик данных» предназначен для информационного взаимодействия с локальными системами данных, размещаемые на серверах центров ЕСИМО и организаций-поставщиков информации и выполняет стандартные функции описания метаданных в соответствии со стандартом ИСО 19115, а также операции предварительного контроля данных и их загрузки на Поставщик данных, а также формирования транспортного файла в международном формате NetCdf.

Программный комплекс «Сервер интеграции» предназначен для управления программным средством «Поставщик данных», поддержки средств унификации обмена данными, обмена данными с «Поставщиками данных» и внешними приложениями (ГИС, портал, Аналитический комплекс – АК, информационно-коммуникационный комплекс взаимодействия – ИККВ). Для сбора логов предназначен компонент Сервисная шина. Единая система регистрации, авторизации, аутентификации обеспечивается компонентом Безопасность. Технология интеграции обеспечивает выполнение следующих задач:

– регистрацию и описание распределенных ИР в локальных системах организаций-поставщиков данных без изменения структур и схем хранения данных;

– поиск, доступ к локальным системам данных и обмен данными между программными компонентами технологии и внешними программными приложениями;

– доставку данных во внешние программные приложения и другие системы (рис. 5), в т.ч. на основе подписки данные доставляются на ftp-сервера, или адрес электронной почты или внешнюю БД соответствии с регламентом пополнения ИР.

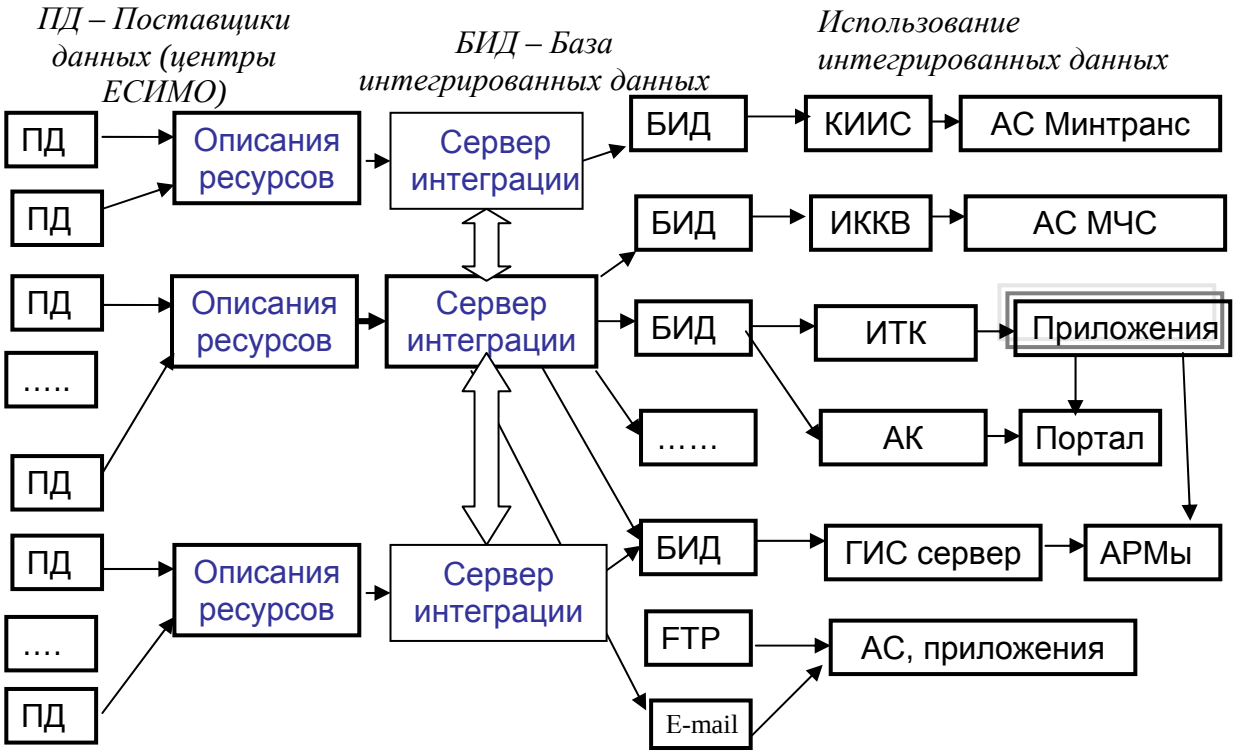



Рисунок 5. Схема использования интегрированных ресурсов

Для повышения оперативности доступа к данным создана база интегрированных данных (БИД), в которой сохраняются все ИР. Таких БИД может быть несколько. Так в ЕСИМО созданы БИД для региональных и специализированных порталов в Санкт-Петербурге, Владивостоке. Прикладная часть программного обеспечения ЕСИМО ориентирована на использование интегрированных данных через БИД, что позволяет развивать приложения с использованием различного инструментария, а также распределенную разработку.

## 6 Использование системы, включая обмен данными

Метаданные используются для поиска, навигации по каталогу ИР, управления контентом, в т.ч. для агрегации данных; мониторинга состояния ресурсов, проведения информационно-аналитических исследований.

Использование ИР осуществляется с помощью портала ЕСИМО и автоматизированных рабочих мест. В зависимости от типа системы хранения ИР

(структурированные, неструктурированные данные, БД, приложения, объектные файлы) организуется соответствующая визуализация.

Установлено более 60 автоматизированных рабочих мест (АРМ) пользователей в МЧС России, Минтранс России, Росгидромете и других ведомствах. При этом производится автоматическая доставка наблюденной, аналитической, прогностической и обобщенной информации на сервера ведомственных автоматизированных систем (АС) Национального центра управления в кризисных ситуациях (НЦУКС) МЧС России с помощью ИККВ и ФГУП «Морсвязьспутник» Минтранс России через Комплексную интегрированную информационную систему (КИИС). В данном случае реализуется межведомственный обмен данными. При этом реализовано несколько вариантов обмена данными.

Наиболее широко используется вариант обмена данными при использовании АРМов. Каждый пользователь АРМа заказывает необходимый ему состав ИР и картографических слоев. Заказанные ИР автоматически в соответствии регламентом или по событию обновления доставляются в БИД. Для ИР, имеющих географические координаты, автоматически строятся картографические слои, оформленные в виде Web map services (WMS) сервисов стандарта Open Geospatial Consortium. Пользователи АРМов при визуализации ИР и картографических слоев автоматически получают необходимые им данные и строят комплексные карты на основе выбранных ими слоев ГИС-визуализатором ЕСИМО (OceanViewer).

Для НЦУКС процессы обмена данными происходят следующим образом. Заказанные по подписке этой организаций 68 цифровых ИР от ФГБУ «ВНИИГМИ-МЦД», ФГБУ «Гидрометцентр России», ФГБУ «ААНИИ», ФГУП «Морсвязьспутник», ФГБУ «НИЦ «Планета», ФГБУ «НПО «Тайфун» ежедневно доставляются Сервером интеграции с Поставщиков данных и перенаправляются в БИД НЦУКС. В НЦУКС на основе этих данных собственной АС строятся карты и визуализируются таблицы.

Обмен данными с ФГУП «Морсвязьспутник» выглядят так. Эта организация заранее определила список картографических тематических и стандартных слоев, оформленных в виде картографических сервисов. Эти сервисы автоматически создаются в соответствии с регламентом обновления ИР (как правило, несколько раз в сутки на ГИС-серверах ФГБУ «ВНИИГМИ-МЦД», ФГБУ «ААНИИ», ФГБУ «ДВНИГМИ»). Сервисы вызываются при необходимости пользователями КИИС при информационном обеспечении пользователей и представляются в виде комплексной карты собственным ГИС-визуализатором, объединяющей как слои пространственного распределения гидрометеорологических данных, так и сведения портах, расположении транспортных судов и их характеристиках.

Последние два варианта реализации обмена данными являются наиболее перспективными, т.к. эти варианты обмена данными позволяют внешним системам автоматически усваивать новую информацию из ЕСИМО.

## 7 Заключение

В результате выполненной работы реализованы следующие инновационные подходы. ЕСИМО не требует специальных преобразований данных в их источниках, достаточно только представить метаданные на интегрируемые ИР. За счет использования единого словаря параметров и стандартизованных для системы классификаторов пользователь может получить ИР в необходимом составе, форматах хранения и представления данных. Такой подход реализован также в системе SeaDataNet [12] Использование таких атрибутов метаданных, как тип наблюдательной платформы (фиксированная точка или движущаяся платформа), временное (регулярные или нерегулярные данные) и пространственное разрешение данных (точка, маршрут, сетка), позволило реализовать возможность автоматического выбора типа представления данных (временной ряд, сеточные данные, профиль), а также объединять данные нескольких информационных ресурсов. Реализации такого подхода в других системах авторам не известна. В состав системы могут включаться существующие модели, например, распространения нефтяных пятен, которые могут работать в различных узлах системы, получать необходимые данные от ЕСИМО и поставлять полученные результаты в систему. Впервые в рамках одной модели интегрируются описания ресурсов не только с фактографическими, но и пространственными данными, а также объектными файлами (приложениями, документами, графическими и видео файлами).

Созданная система интеграции распределенных и неоднородных данных позволила организовать межведомственный обмен данными для 12 ведомств России и начать переход от разрозненных и фрагментарных наборов данных к единому пространству данных [11]. За счет интеграции данных впервые в мире на одной интерактивной карте можно увидеть, как данные наблюдений, анализов, прогнозов, климата, так и сведения о расположении российских транспортных судов, портов, судостроительных организаций.

Опыт, накопленный при создании и эксплуатации ЕСИМО, и масштабируемое программное обеспечение представляет интерес для других ведомств России. Уже имеются результаты внедрения программного обеспечения в Межправительственной океанографической комиссии ЮНЕСКО для создания международного портала Ocean Data Portal (http://www.oceandataportal.org), на котором

представлены зарубежные ИР, доступные пользователям. Это позволяет говорить о реализации автоматизированного взаимодействия между информационными системами национальных центров океанографических данных различных стран и международными системами. Для Всемирной метеорологической организации (ВМО) использовано разработанное программное обеспечение при создании Глобального центра информационной системы ВМО в России для обмена гидрометеорологическими данными между метеорологическими центрами, расположенными в Москве, Вашингтоне, Мельбурне (http://portal.gisc-msk.wis.mecom.ru:8080/portal/). Ведутся переговоры об использовании программного обеспечения в ВМФ, Погранслужбе России, Росгидромете для создания ведомственных систем интеграции данных.

Важными направлениями развития ЕСИМО являются снижение сложности работы с системой за счет повышение уровня автоматизации управления системой, уменьшение затрат на ее эксплуатацию, дальнейшее повышение надежности работы аппаратно-программных средств системы.

## Литература


[1] Белов С.В., Бритков В.Б. Интеграция информационных ресурсов в задачах исследования морской среды // Журнал «Информационные технологии и вычислительные системы». – 2008. – Вып. 1. – с. 73–82.

[2] Вязилов Е.Д., Михайлов Н.Н. Интеграция данных о морской среде и деятельности // Инфраструктура спутниковых геоинформационных ресурсов и их интеграция : сб. науч. статей / под ред. д-ра техн. наук М.А. Попова и д-ра техн. наук Е.Б. Кудашева. – Киев : Карбон-Сервис, 2013. – С. 174–181.

[3] Единая межведомственная информационная система статистики. – 2014. [Электронный ресурс]. – Режим доступа: http://fedstat.ru/indicators/start.do, свободный.

[4] Когаловский М.Р. Методы интеграции данных в информационных системах // Институт проблем рынка РАН. – 2010. [Электронный ресурс]. – Режим доступа: http://www.cemi.rssi.ru/mei/articles/kogalov10-05.pdf, свободный.

[5] Макальский Л.М., Гаврилов А.И., Жебрунов Г.А., Тихонова Е.А. Реализация экспорта/импорта данных между разнородными информационными системами // Информационные технологии моделирования и управления. – 2008. – № 5 (48). – С. 572–576.

[6] Реестр федеральных государственных информационных систем. Последнее изменение: 28.06.2014. – [Электронный ресурс]. – Режим доступа: http://rkn.gov.ru/it/register/#, свободный.

[7] Серебряков В.А. Семантическая интеграция данных. 24.04.2012. – 102 с. [Электронный ресурс]. – Режим доступа: http://sp.cmc.msu.ru/proseminar/2012/serebryakov.2012.04.20.pdf, свободный.

[8] Черняк Л. Интеграция данных: синтаксис и семантика // Открытые системы. – 2009. – № 10. [Электронный ресурс]. – Режим доступа: http://www.osp.ru/os/2009/10/11171290/, свободный.

[9] Baudouin Raoult & Guillaume Aubert & Marta Gutiérrez & Cristina Arciniegas-Lopez & Ricardo Correa. Virtual organization in the SIMDAT meteorological activity: a decentralized access control mechanism for distributed data. – Springer-Verlag, 2009. – Earth Sci. Inform. – 2009. – Vol. 2. – P. 63–74. – DOI 10.1007/s12145-009-0026-7.

[10] Data Integration for Dummies. USA. Published by John Wiley & Sons, Inc. Hoboken, New Jersey, 2014. – 52 p. – ISBN 978-1-118-89658-7.

[11] Franklin M., Halevy A., Maier D. From Databases to Dataspaces: A New Abstraction for Information Management // SIGMOD Record. – Dec. 2005. – Vol. 34, No. 4. – Пер. С. Кузнецова. http://citforum.ru/database/articles/from_db_to_ds

[12] Maillard Catherine, Lowry Roy, Maudire Gilbert, Schaap Dick and the SeaDataNet Consortium. SeaDataNet: Development of a Pan-European Infrastructure for Ocean and Marine Data Management // OCEANS'07, Aberdeen, Scotland, 18–21 June, 2007. – 15 p.

[13] What is GEOSS? The Global Earth Observation System of Systems. [Электронный ресурс]. – Режим доступа: http://www.earthobservations.org/gci_gp.shtml, свободный. – Загл. с экрана.


## The Approaches for Development of Interagency Exchange of Distributed Heterogeneous Data for the ESIMO


Evgeny D. Vyzilov, Nikolay N. Mikhailov, Alexander E. Kobelev, Denis A. Melnikov



The basic approaches for data integration and interagency interaction (data exchange) for marine environment and maritime activities are given. An example of implementation of the Unified state system of information for the World Ocean is considered. Data integration requires No special data transformations in the sources for data integration is required; you just have to submit metadata on the information resources. By a common vocabulary of parameters and standardized classifiers, the user can get the resources in the needed composition, format, or in the form of tables, graphs and map.

# Комбинированная виртуально-материализованная среда интеграции больших неоднородных коллекций данных

© С. А. Ступников ©А. Е. Вовченко

ИПИ РАН,
Москва

ssa@ipi.ac.ru alexey.vovchenko@gmail.com

**Аннотация**

В работе рассматривается архитектура комбинированной виртуально-материализованной среды интеграции неоднородных коллекций данных различного вида (структурированных, слабоструктурированных и неструктурированных). Необходимость поддержки двух различных видов интеграции объясняется тем, что как виртуальный, так и материализованный подходы к интеграции имеют свои достоинства и недостатки. Виртуальная интеграция осуществляется с использованием технологии предметных посредников. Материализованная интеграция реализуется с использованием свободно распространяемой платформы распределенного хранения и обработки данных Hadoop; а также системы организации реляционных хранилищ данных над Hadoop, в качестве которой могут использоваться платформы Big SQL или Hive.

*Работа выполнена при поддержке РФФИ (гранты 13-07-00579, 14-07-00548) и Президиума РАН (Программа фундаментальных исследований Президиума РАН № 16 «Фундаментальные проблемы системного программирования»).*

## 1 Введение

Новая парадигма [24] в науке и информационных технологиях, доминирующая в последнее время, основывается на *исследовании данных* (data exploration). Роль данных особо подчеркивается и становится критической. Объемы данных фактически во всех областях деятельности растут со временем экспоненциально. Поэтому новая парадигма требует создания методов и средств оперирования данными, объемы которых выходят за рамки возможностей технологий баз данных, развивавшихся в последние десятилетия (преимущественно реляционных). Необходимы подходы, позволяющие справляться с разнообразием моделей данных (включая неструктурированные и слабоструктурированные данные), метаданных, семантики данных.

К моделям данных, появившимся в последнее время, относятся разнообразные NoSQL-модели: документные модели (системы SimpleDB, MongoDb, CouchDB), модели с колоночным хранением (системы HBase, HyperTable, Cassandra), модели «ключ-значение» (системы Voldemort, Riak, Redis, Scalaris). Развиваются онтологические и семантические модели, такие, как семейства RDF и OWL; графовые модели (системы Neo4j, Dex, GraphDB, HyperGraphDB, Trinity, Pregel); модели, основанные на многомерных массивах – ММ-модели (SciDB).

Методы создания унифицированного представления различных видов нетрадиционных моделей в канонической информационной модели (общем языке, унифицирующем языки разнообразных моделей данных) в последнее время активно исследовались в ИПИ РАН. Были рассмотрены подходы к унификации моделей данных различных классов: семантических (OWL [10], RDF [3]), графовых [5], ММ-моделей [4], NoSQL-моделей [2]. В качестве канонической модели рассматривался язык СИНТЕЗ [11], поддерживающий комбинированную слабоструктурированную (фреймовую) и объектную модель данных.

Для поддержки новых моделей данных создаются системы управления данными, обладающие масштабируемостью, высокой доступностью, возможностью разбиения коллекций данных произвольным образом на разделы для параллельной обработки.

Растущая популярность использования слабоструктурированных баз данных (в различных видах NoSQL) наряду с реляционными базами, в совокупности с технологиями программирования Hadoop и MapReduce, обеспечивающими параллельную обработку огромных массивов

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

слабоструктурированных данных, объясняется множеством фактических и потенциальных применений. Например, развитие Веб-приложений, социальных сетей, сенсорных сетей, финансовых и торговых приложений, интенсивная работа с данными в науке (в науках о Земле, в биоинформатике), и пр. требуют использования данных, которые с трудом поддаются частичной структуризации. Твиты и посты в блогах являются включают слабоструктурированные текстовые отрывки, изображения и видео-ролики. Преобразование такого контента в структурированный формат для семантического анализа и поиска является трудной задачей, имеющей целью отображение структур и семантики данных в форму, «понимаемую» компьютером для извлечения информации из данных.

Данная работа относится к области конструирования средств поддержки систем с интенсивным использованием данных. Методы унификации моделей данных образуют формальную базу работы.

Целью работы является разработка и реализация комбинированной виртуально-материализованной архитектуры среды интеграции неоднородных коллекций данных различного вида (структурированных, слабоструктурированных и неструктурированных). Такая среда должна поддерживать как виртуальную, так и материализованную интеграцию коллекций данных, представленных как в традиционных, так и нетрадиционных моделях данных.

Необходимость поддержки двух различных видов интеграции объясняется тем, что как виртуальный, так и материализованный подходы интеграции имеют свои достоинства и недостатки.

Виртуальная интеграция осуществляется с использованием технологии предметных посредников [12], образующих промежуточный слой между пользователем (приложением) и неоднородными информационными ресурсами. При этом данные из ресурсов не материализуются в посреднике. К достоинствам виртуальной интеграции относится то, что развертывание и поддержка системы виртуальной интеграции значительно дешевле развертывания и поддержки системы материализованной интеграции (хранилища данных) исходя как из временных, так и из материальных затрат. Обычно системы виртуальной интеграции используются для интеграции ресурсов, данные из которых трудно преобразовывать и (или) владельцами которых являются разные лица. Однако, данные в интегрируемых ресурсах не должны быть слишком большими, либо запросы к ресурсам должны обладать достаточной степенью селективности для того, чтобы объем данных, передаваемых из ресурса, не был слишком велик.

При материализованной интеграции предполагается создание хранилища данных (warehouse), в которое загружаются коллекции данных, подлежащие интеграции. В процессе загрузки происходит преобразование данных из схемы коллекции в общую схему хранилища. При необходимости, хранилища могут масштабироваться на большие объемы данных (хотя это требует соответствующих материальных затрат). Хранилища предоставляют удобную и эффективную платформу преобразования и интеграции данных, а также решения сложных аналитических задач над данными.

Материализованная интеграция реализуется в комбинированной архитектуре с использованием свободно распространяемой платформы распределенного хранения и обработки данных Hadoop [27]; а также системы организации реляционных хранилищ данных над Hadoop (WR-Hadoop для краткости), в качестве которой могут использоваться, например, платформы Big SQL [22] или Hive [9].

В данной статье рассматриваются основные принципы построения комбинированной виртуально-материализованной архитектуры и подходы к ее реализации. Создание программных средств реализации рассматриваемой архитектуры является предметом дальнейшей работы. Ближайшей задачей является выбор конкретной системы WR-Hadoop, одним из вариантов которой является поддержка обеих систем Big SQL и Hive.

Статья организована следующим образом. В разделе 2 рассматриваются основные черты комбинированной архитектуры. В разделе 3 иллюстрируются преобразования коллекций, представленных в нетрадиционных моделях данных, в их интегрированное представление в модели данных сопряжения Hadoop – среда посредников. В разделе 4 рассматривается пример задачи интеграции неоднородных коллекций данных в комбинированной среде.

## 2 Архитектура комбинированной виртуально-материализованной среда интеграции коллекций данных

Архитектура комбинированной среды интеграции неоднородных коллекций данных (рис. 1) основана на существующей архитектуре предметных посредников [12].

Основные черты архитектуры выглядят следующим образом:

- концептуальная схема данных, предлагаемая пользователю системой интеграции (посредником), формируется независимо от интегрируемых ресурсов;
- ресурсы, релевантные концептуальной схеме, регистрируются в посреднике: их схемы связываются с концептуальной схемой при помощи представлений (взглядов). Определение взглядов осуществляется полуавтоматически и требует привлечения экспертов;

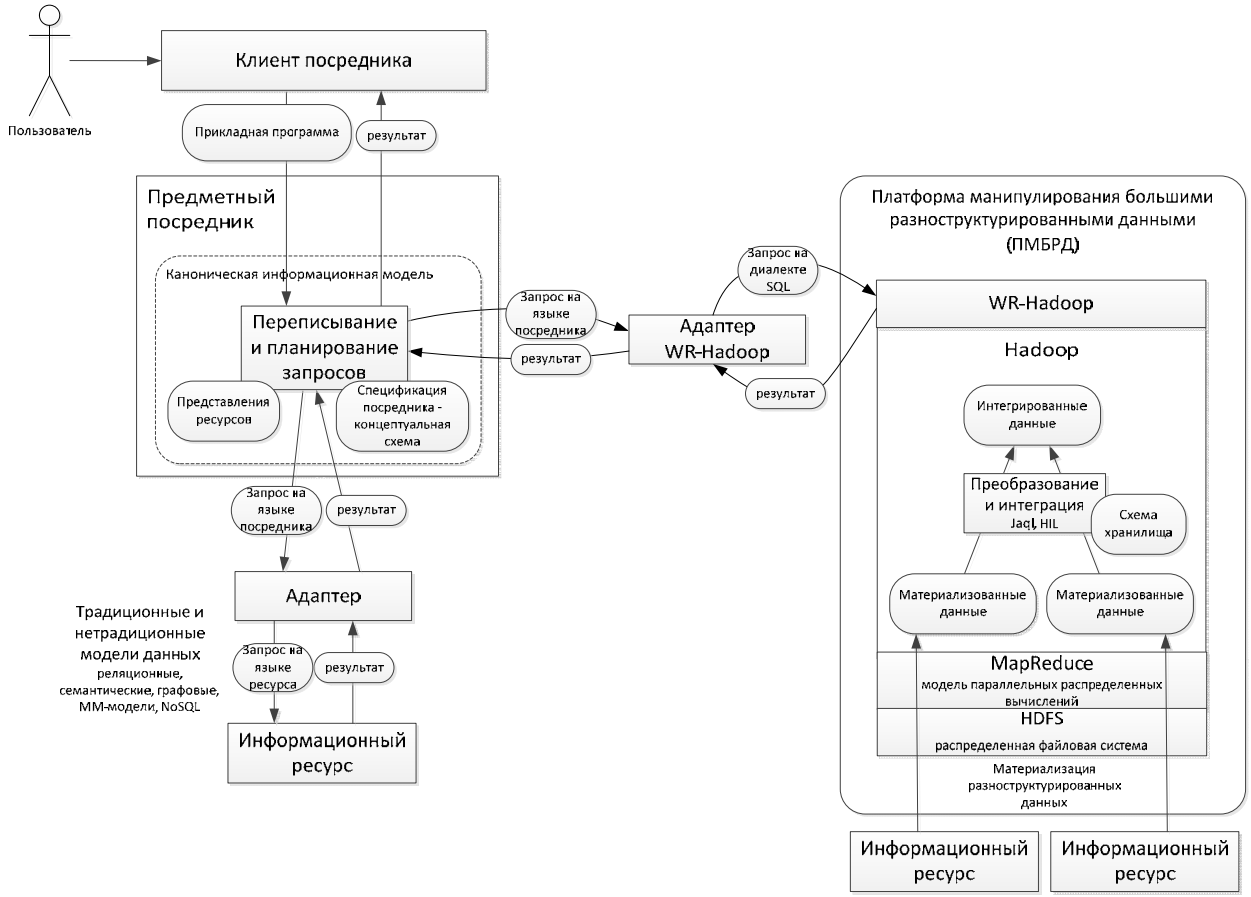



Рисунок 1. Архитектура комбинированной среды интеграции неоднородных коллекций данных

- пользователь задает программы (запросы) над концептуальной схемой. Эти запросы переписываются в посреднике с использованием взглядов в частичные подзапросы в терминах конкретных ресурсов. Переписывание осуществляется в языке правил канонической модели посредников;
- взаимодействие ресурсов и посредника реализуется при помощи адаптеров, располагающихся между посредником и ресурсами. Адаптеры преобразуют запросы из языка правил канонической модели в язык ресурса, а также преобразуют результат исполнения запроса из формата ресурса в формат посредника.

Таким образом, для виртуальной интеграции ресурсов, основанных на традиционных и нетрадиционных моделях, необходимо построение адаптеров соответствующих моделей данных. В области построения адаптеров для среды предметных посредников накоплен большой опыт, разработаны общая архитектура адаптеров и методы конструирования адаптеров [1], реализованы адаптеры реляционной и объектно-реляционной моделей, адаптер веб-сервисов, XML-адаптер и другие. Формальной базой для построения адаптеров служат отображения моделей данных ресурсов в каноническую модель [6], разработанные в результате унификации моделей ресурсов.

### 2.1 Платформа для материализованной интеграции ресурсов в комбинированной архитектуре

Для материализованной интеграции ресурсов в комбинированной архитектуре необходима масштабируемая платформа манипулирования большими разноструктурированными данными (ПМБРД). В качестве такой платформы в данной работе выбрана связка системы *Hadoop* и системы организации реляционных хранилищ данных над Hadoop – WR-Hadoop (в качестве которой могут использоваться, например, системы Big SQL [22] или Hive [9]).

Платформа *Apache Hadoop* [7][27] была впервые представлена в 2005 г. в составе проекта Apache Software Foundation и представляет собой набор программных средств распределенного хранения и обработки больших объемов данных. Платформа Hadoop предназначена для развертывания на вычислительных кластерах и основана на распределенной файловой системе HDFS (Hadoop Distributed File System). В состав кластера входят узлы, хранящие фрагменты файлов. Теоретически могут быть сотни и тысячи таких узлов, основанных на недорогих вычислительных платформах (commodity hardware). Для обеспечения высокой надежности поддерживается избыточность путем

создания копий фрагментов между узлами. С точки зрения пользователя такая структура выглядит как обычная древовидная файловая система.

Масштабируемость платформы на значительные объемы данных достигается за счет параллельной обработки фрагментов на узлах с использованием программной модели параллельных распределенных вычислений MapReduce [21]. Модель MapReduce позволяет разработчикам приложений абстрагироваться от сложностей работы распределенных программ, связанных с распределением данных, планированием нагрузки и обеспечением отказоустойчивости.

Платформы *Hive* [9] и *Big SQL* [22] представляют собой решения для организации реляционных хранилищ данных, разработанные на основе среды Hadoop. В них реализованы известные структуры реляционных баз данных – таблицы, столбцы, разделы. Платформы поддерживают реляционные языки манипулирования данными (HiveQL и Big SQL соответственно) для работы в неструктурированной среде Hadoop: моделью данных Hive является стандарт SQL92 с некоторыми дополнениями, моделью данных Big SQL – практически полный стандарт SQL 2011. Фактически, системы проецируют реляционную структуру на данные, хранящиеся в Hadoop и предоставляют возможность исполнения SQL-подобных запросов на больших наборах данных путем компиляции их в программы MapReduce, исполняемые в среде Hadoop.

Система Hive является свободно распространяемым решением, а Big SQL – проприетарным, распространяемым в составе продукта IBM InfoSphere BigInsights [16].

Следует отметить, что системы Hive и Big SQL не в полной мере реализуют функции хранилищ данных (warehouses). В частности, напрямую не поддерживается процесс извлечения-преобразования-загрузки (ETL). Для реализации недостающих функций в комбинированной архитектуре предлагается использование декларативно-императивных языков высокого уровня над Hadoop (раздел 2.2).

Таким образом, в комбинированной архитектуре обеспечивается возможность распределенного хранения, преобразования и интеграции больших разноструктурированных данных (при помощи Hadoop), а также унифицированный взгляд на материализованные данные через реляционную модель (при помощи Hive или Big SQL).

Потенциально, в качестве ПМБРД может быть выбрана и другая платформа: например, вместо Hadoop могут быть использованы такие системы, как Apache Spark [8, 28], GraphLab [20, 25], Disco [14] и др. Однако, Hadoop в настоящее время является наиболее распространенной и универсальной системой, поддерживающей модель вычислений MapReduce.

В комбинированной архитектуре ПМБРД рассматривается как еще один вид ресурсов, подлежащий виртуальной интеграции. Интеграция становится двухслойной: материализованная интеграция осуществляется внутри ПМБРД, виртуальная – на уровне предметных посредников. Виртуальной интеграции при этом могут подлежать ресурсы

- данные из которых сложно или невозможно материализовать по разным причинам и (или)
- модель данных включает специфические операции и алгоритмы, адаптация которых в реляционной модели сложна и (или) неэффективна. К таким моделям относятся, например, графовые, ММ-модели и т.д.

Для включения ПМБРД Hadoop/WR-Hadoop в среду предметных посредников необходимо разработать адаптер для WR-Hadoop (Hive или Big SQL) в соответствии с подходом, изложенным в работе [1]. При этом модель данных WR-Hadoop выступает в роли *модели сопряжения ПМБРД со средой предметных посредников*. Модель данных WR-Hadoop должна быть унифицирована (отображена в каноническую модель посредников). Концептуально унификация модели данных WR-Hadoop опирается на проведенную при конструировании реляционного адаптера [1] унификацию реляционной модели. В конструкцию реляционного адаптера будут внесены два важных усовершенствования:

- поддержка объектных таблиц;
- поддержка сложных (complex) типов данных, таких, как массивы (ARRAY), структуры (STRUCT) и отображения (MAP).

### 2.2 Языки и инструменты для материализованной интеграции в комбинированной архитектуре

Материализация в ПМБРД осуществляется путем помещения в Hadoop-кластер файлов, экспортированных из информационных ресурсов. Файлы могут быть экспортированы в различных открытых форматах: JSON, XML, CSV, в виде текстовых файлов, а также в бинарном формате JSON (JSON binary). Материализации, как и виртуальной интеграции, могут подлежать данные, представленные в различных традиционных и нетрадиционных моделях. Решение о том, какой способ интеграции следует применить для конкретного ресурса, следует принимать исходя из целей системы интеграции (например, характерных запросов пользователя), эффективности, стоимости и т.д.

Преобразование данных к реляционному виду для последующей интеграции производится при помощи программ на языке *Jaql*.

*Jaql* представляет собой язык запросов и сценариев, разработанный IBM и использующий формат обмена данными JavaScript Object Notation (JSON [19]). Jaql поддерживает произвольную глубину вложенности структур данных, является в

высокой степени функционально-ориентированным, чрезвычайно гибким, и хорошо применимым для работы со слабоструктурированными данными. Язык ориентирован на прозрачное применение модели программирования MapReduce: декларативно-императивные запросы Jaql переписываются в последовательность программ MapReduce, исполняемых в среде Hadoop.

Основными структурами данных, подлежащими манипулированию при помощи Jaql, являются объекты (коллекции пар <имя, значение>) и массивы (упорядоченные списки значений). Поддерживается множество встроенных функций над массивами и объектами. Язык предоставляет широкий набор операций преобразования данных (фильтрация, группировка, сортировка, соединение, объединение и т.д.).

Язык Jaql поставляется в составе InfoSphere BigInsights [16] – программной платформы обработки больших данных, основанной на Hadoop.

Итак, преобразованные к реляционному виду данные сохраняются в формате JSON. Преобразования коллекций, представленных в нетрадиционных моделях данных, в их интегрированное представление в реляционной модели данных иллюстрируются в разделе 5. Сложные потоки обработки данных (очистки, устранения дублирования, слияния) и их интеграции реализуются с использованием комбинации языков Jaql и *HIL*[1].

Декларативный язык *HIL* (High-level Integration Language) разработан IBM для программирования сложных потоков обработки данных (ETL), агрегирующих факты из больших коллекций разноструктурированной информации в целевые коллекции унифицированных сущностей.

Язык позволяет реализовать методы извлечения, сопоставления и группирования, разбора, связывания, устранения дублирования (deduplication) различных разноструктурированных представлений информации об одних и тех же сущностях реального мира (entity resolution). HIL также позволяет реализовать методы и операции слияния (интеграции) данных об одних и тех же сущностях реального мира и их связей, представленных в разных коллекциях, образованных в процессе разрешения сущностей (включая реализации стратегий и устранения конфликтующих данных, операции поглощения и слияния данных). Программы на HIL компилируется в Jaql, что позволяет использовать HIL для преобразования и интеграции данных в Hadoop.

Следует отметить, что методы материализованной интеграции, организации ETL-процессов, очистки данных, предлагаются в составах программных решений ведущих разработчиков баз данных. К таким решениям относится, например, IBM InfoSphere Information Server [17]. Этот программный продукт потенциально может быть использован в комбинированной архитектуре для материализованной интеграции. При этом в качестве хранилища может быть использована платформа IBM InfoSphere Warehouse [18] (являющаяся частью СУБД DB2). Однако, Information Server предлагает лишь ограниченное количество относительно простых методов выделения, отображения и слияния сущностей. Недостаточность таких средств была осознана компанией IBM, в результате чего и появился язык HIL.

Рассматриваемая в данной статье комбинированная среда интеграции нацелена, в частности, на исследование методов интеграции больших разноструктурированных данных. Этим и мотивирован выбор в качестве ПМБРД платформы Hadoop/WR-Hadoop и комбинации языков Jaql и HIL в качестве инструментов материализованной интеграции.

## 3 Преобразование коллекций данных нетрадиционных моделей в интегрированное представление

Рассмотрим пример коллекции данных, представленной в модели данных графовой СУБД Neo4j [26]. Небольшой подграф базы данных о фильмах, включающий основные виды вершин, ребер и атрибутов, изображен на рис. 2.

Вершины графа соответствуют фильмам и людям, ребра графа соответствуют участию людей в создании фильма как актеров (CAST) или режиссера (DIRECTS). Люди характеризуются именем (*name*), фильмы – названием (*title*) и годом создания (*year*), роли актеров – именем персонажа (*character*). Вершины и ребра графа обладают уникальными идентификаторами.

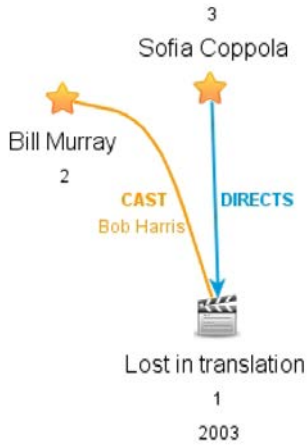



Рисунок 2. Пример графа в модели данных Neo4j

[1] Вопросы выделения сущностей и интеграции разноструктурированной информации при помощи программ на языке HIL выходят за рамки данной статьи.

Графовая БД системы Neo4j может быть сериализована в формате JSON. Представление для вышеприведенного графа в JSON выглядит следующим образом:

```
[{ "id": 1, "type": "node", "labels": ["MOVIE"],
"properties": {"title": "Lost in translation", "year": 2003} },
{ "id": 2, "type": "node", "labels": ["PEOPLE"],
"properties": {"name": "Bill Murray"} },
{ "id": 3, "type": "node", "labels": ["PEOPLE"],
"properties": {"name": "Sofia Coppola"} },
{ "id": 4, "type": "relationship",
"start_node": 1, "end_node": 2,
"relationship_type": "CAST",
"properties": {"character": "Bob Harris"} },
{ "id": 5, "type": "relationship",
"start_node": 3, "end_node": 1,
"relationship_type": "DIRECTS" }]
```

Здесь *start_node* и *end_node* обозначают исходящую и входящую вершины ребра соответственно; *labels* обозначает множество меток, присвоенных вершине.

Реляционное представление данной графовой БД может выглядеть следующим образом. Схема реляционной БД состоит из двух отношений, одно из которых соответствует вершинам графа (*Nodes*, табл. 1), другое – ребрам (*Relationships*, табл. 2).

Таблица 1. Отношение *Nodes*

| id | labels | title | year | name |
|---|---|---|---|---|
| 1 | [“MOVIE”] | “Lost in translation” | 2003 | |
| 2 | [“PEOPLE”] | | | “Bill Murray” |
| 3 | [“PEOPLE”] | | | “Sofia Coppola” |

Таблица 2. Отношение *Relationships*

| id | start_node | end_node | relationship_type | character |
|---|---|---|---|---|
| 4 | 1 | 2 | “CAST” | “Bob Harris” |
| 5 | 3 | 1 | “DIRECTS” | |

Преобразование графовой БД в реляционное представление может быть осуществлено при помощи следующих двух функций, определенных на языке Jaql:

```
createNodesRelation = fn(movie_graph_db)(
      movie_graph_db->filter $.type == "node"->
        transform { id: $.id, labels: $.labels,
          title: $.properties.title, year: $.properties.year,
          name: $.properties.name }  );

createRelationshipRelation = fn(movie_graph_db)(
      movie_graph_db->filter $.type == "relationship"->
        transform { id: $.id,
          start_node: $.start_node, end_node: $.end_node,
          relationship_type: $.relationship_type,
          character: $.properties.character }  );
```

Функции принимают на вход представление графовой БД в формате JSON. Первая из функций возвращает отношение *Nodes* в формате JSON, пригодное для загрузки в соответствующую реляционную таблицу, вторая – отношение *Relationships*. При определении функций используются выражения фильтрации массивов (*filter*) и преобразования элементов массивов (*transform*), связанные оператором организации потоков данных (->).

В качестве еще одного примера коллекции данных рассмотрим базу знаний *DBpedia* (http://dbpedia.org/). Коллекция содержит структурированную информацию, извлеченную из свободной энциклопедии Википедия. Данные в коллекции представлены в модели RDF. В качестве примеров данных рассмотрим RDF-спецификации города *Кембридж* (Великобритания) и его представителя в Парламенте *Эндрю Лэнсли*. Спецификации представлены в формате JSON:

```
{ "http://dbpedia.org/resource/Cambridge": {
"http://dbpedia.org/property/officialName": [ {
  "type": "literal", "lang" : "en",
  "value" : "City of Cambridge" } ],
"http://dbpedia.org/ontology/areaTotal": [ {
  "type": "literal", "value": 115650000,
  "datatype": "http://www.w3.org/XMLSchema#double"}],
"http://dbpedia.org/ontology/isPartOf": [
 { "type": "uri",
 "value": "http://dbpedia.org/resource/East_of_England" },
 { "type": "uri",
 "value": "http://dbpedia.org/resource/Cambridgeshire" } ],
"http://dbpedia.org/ontology/leaderName": [ {
"type": "uri",
 "value": "http://dbpedia.org/resource/Andrew_Lansley"} ]
} }

{ "http://dbpedia.org/resource/Andrew_Lansley": {
"http://dbpedia.org/property/name": [ {
  "type" : "literal", "lang" : "en",
  "value" : "Andrew Lansley" } ],
"http://dbpedia.org/ontology/birthDate": [ {
  "type": "literal", "value": "1956-12-10+02:00",
  "datatype": http://www.w3.org/XMLSchema#date } ],
"http://dbpedia.org/ontology/party": [ {
  "type": "uri",
 "value": "http://dbpedia.org/resource/Conservative_
    Party_(UK)" } ] ,
"http://dbpedia.org/ontology/electionMajority": [ {
  "type": "literal", "value": 7838,
 "datatype" : "http://www.w3.org/XMLSchema#integer"}]
} }
```

Свойства Кембриджа как объекта включают, в частности, название (*officialName*), площадь (*areaTotal*), вышестоящие территориальные образования (*isPartOf*), лидера (*leaderName*). Свойства Эндрю Лэнсли, как объекта базы знаний , включают имя (*name*), дату рождения (*birthDate*), партию (*party*), количество голосов на выборах (*electionMajority*).

Реляционное представление объектов такого рода может выглядеть следующим образом. Схема реляционной БД состоит из двух отношений, одно из которых соответствует городам (*Cities*), другое – персонам (*Persons*). Атрибутам отношений соответствуют свойства объектов. В RDF-хранилищах подобные отношения, группирующие свойства однородных объектов, называются таблицами свойств (*property tables* [29]).

Таблица 3. Отношение *Cities*

| subject | officialName | area Total | isPartOf | leaderName |
|---|---|---|---|---|
| "http://dbpedia.org/ resource/Cambridge" | ["City of Cambridge"] | [115650000] | ["http://dbpedia.org/resource/ East_of_England", "http://dbpedia.org/resource/ Cambridgeshire"] | ["http://dbpedia.org/ resource/Andrew_Lansley"] |

Таблица 4. Отношение *Persons*

| subject | name | birthdate | party | election Majority |
|---|---|---|---|---|
| "http://dbpedia.org/ resource/Andrew_Lansley" | ["Andrew Lansley"] | ["1956-12-10+02:00"] | ["http://dbpedia.org/resource/ Conservative_Party_(UK) "] | [7838] |

Преобразование RDF-объектов персон в кортежи отношения *Persons* может быть осуществлено при помощи функции *createPersonTuple*, определенной на языке Jaql:

```
createPersonTuple = fn(person_rdf) (
    pa = ["http://dbpedia.org/property/name",
          "http://dbpedia.org/ontology/birthDate",
          "http://dbpedia.org/ontology/party",
          "http://dbpedia.org/ontology/electionMajority"],
    properties_record = index(values(person_rdf), 0),
    if( not
       containedIn( "false",
         for($pa in pa)
           if(containedIn($pa,
                names(record(values(person_rdf)))))
             ["true"]
           else ["false"])
    )
    if( arity(person_rdf) == 1 )
    {  subject: index(names(person_rdf), 0),
       name:   for( $name in
         properties_record."http://dbpedia.org/property/
           name")   [$name.value],
     birthday: for( $birthday in
         properties_record."http://dbpedia.org/ontology/
           birthDate")  [$birthday.value],
     party:   for( $party in
        properties_record."http://dbpedia.org/ontology/
           party")   [$party.value],
     electionMajority:    for( $em in
       properties_record."http://dbpedia.org/ontology/
           electionMajority")     [$em.value]
    }
);
```

Функция принимает на вход RDF спецификацию персоны в формате JSON и возвращает кортеж, пригодный для загрузки в соответствующую реляционную таблицу. При определении функции используются условное выражение *if*, выражение цикла *for*, встроенные функции манипулирования массивами и записями *names*, *values*, *record*, *index* [16] и вспомогательная функция *containedIn(elm, arr)*, возвращающая значение *true*, если значение *elm* является элементом массива *arr*:

```
containedIn = fn(elm, arr)(
    exists(for( $iter in arr ) if($iter == elm) [true])  );
```

Функция преобразования RDF-объектов городов определяется аналогичным образом.

## 4 Пример задачи интеграции неоднородных коллекций данных

Рассматриваемая задача состоит в определении отношения населения к экономическим и политическим вопросам в конкретном регионе. В задаче используются следующие информационные ресурсы:

- архивы региональных электронных СМИ;
- социальные сети (например, *ВКонтакте*);
- сервисы микроблогов (например, *Twitter*).

Ресурсы существенным образом отличаются по структуре, а также по характеру и лексике текстов. Например, сообщение из Twitter, представленное в формате JSON, содержит текст, идентификатор сообщения, язык сообщения, дату создания, информацию о пользователе (приведена лишь часть данных, составляющих сообщение):

```
{ “text”: “В первом квартале 2014 ввод жилья в Бобруйской области вырос на 30 процентов http://t.co/9cTJcnkucI”,
“id”: 441892291058622460,
“lang”: "ru",
“created_at”: "Fri Mar 07 11:05:06 +0000 2014",
“user”: { "name":"Петр Иванов",
        "screen_name":"32_minute",
        “id”: 2191013094}
}
```

Сообщение из сети ВКонтакте содержит идентификатор сообщения, идентификатор автора, идентификатор получателя, дату создания, текст, количество «лайков», количество переотправок (приведена лишь часть данных, составляющих сообщение; текст сообщения приведен не полностью):

```
{ "id": 254,
"from_id":2785124,
"to_id":6835,
"date":1387737719,
"text":"Бобруйская область - регион# в котором добывается больше половины всего леса в России. Однако на благосостоянии простых жителей Ухтомского района это никак не сказывается. …",
"likes":{"count": 10},
"reposts":{"count": 5, "user_reposted": 5}
}
```

В рамках задачи анализу подлежат статьи из СМИ и сообщения из социальных сетей за определенный период времени (текущий год, квартал и т.д.) и опубликованные авторами, проживающими в определенном регионе. Такие статьи и сообщения извлекаются из соответствующих ресурсов, загружаются в Hadoop и пропускаются через средства текстовой аналитики[2], развернутые на каждом из узлов кластера. Анализ текста позволяет извлечь из текстов статей и сообщений различные упоминаемые объекты: персоны (и их должности), территориальные образования, организации и т.д. Также анализ текста позволяет оценить тональность сообщения – выявить позитивное, нейтральное или негативное отношение автора текста к теме текста. Таким образом, тональность текста может быть связана с объектами, упоминаемыми в тексте.

Например, информация, извлеченная из приведенного выше сообщения из сети ВКонтакте, может выглядеть следующим образом:

```
{ “user_id”: “6835”,
“message_id”: 254,
“source”: “ВКонтакте”,
“extracted_objects”:{
      “persons”: [{ “name”: “Василий”,
            “surname”: “Петров”,
            “position”: “губернатор”}],
      “territorial_entities”: [
         {“name”: “Бобруйская”,
           “type”: “область”},
         {“name”: “Ухтомский”,
           type: “район”}]
},
“sentiment”: “negative”,
“negative_keywords”:
      [“барак”, “отчаяние”, “прогнивший”]
}
```

В тексте сообщения упоминаются губернатор *Василий Петров*, территориальные образования *Бобруйская область* и *Ухтомский район*. Текст носит отрицательную тональность.

Материализованные статьи и сообщения, обогащенные информацией, извлеченной из текстов, преобразуются к виду, удовлетворяющему единой схеме хранилища[3]. В хранилище помещаются также данные, извлеченные из профилей пользователей социальных сетей.

Определение отношения населения к экономическим и политическим вопросам может быть осуществлено путем различных запросов к схеме хранилища. Могут быть сделаны выводы, например, об отношении населения к конкретным лицам, организациям за определенные промежутки времени и на определенной территории путем подсчета позитивных, нейтральных и негативных сообщений и статей, упоминающих этих лиц или организации.

Отношение населения к экономическим и политическим вопросам может быть также ограничено некоторой темой, например, *«проблемы жилищно-коммунального хозяйства»*. В этом случае анализу подвергаются лишь те сообщения, в которых присутствуют ключевые слова (или их комбинации), относящиеся к теме, например, *{жилье, ЖКХ, отопление, электричество, тариф, барак, ветхий, аварийный}*.

Дополнительные аналитические возможности предоставляет виртуальная интеграция социальных графов (хранимых в графовой базе данных, например, Neo4j [26]) и хранилища сообщений и статей в одном предметном посреднике. Социальные графы (образуемые отношениями «друг» или «подписчик») загружаются в графовую базу данных, предоставляющую операции и алгоритмы анализа графов. Предметный посредник, интегрирующий социальные графы и хранилище статей и сообщений, позволяет формулировать и исполнять программы, определяющие влияние социальных сетей на формирование отношения населения к экономическим и политическим вопросам. Например, может быть отслежено распространение во времени позитивных, нейтральных и негативных сообщений, относящихся к некоторой теме (заданной ключевыми словами) в связных подграфах социальных сетей.

## 5 Родственные работы

В качестве примера подхода, родственного предлагаемой в данной статье комбинированной виртуально-материализованной архитектуре, необходимо упомянуть средства федерализации, поддерживаемые в Big SQL [30]. Эти средства появились в июле 2014 г., в момент создания финальной версии статьи. Федерализация является прямым аналогом виртуальной интеграции в предметных посредниках.

Архитектура системы федерализации BigSQL (рис. 2) средства выполнения программ (*Runtime Engine*), адаптеры (*Wrappers*) и удаленные базы данных. Важным компонентом является *Optimizer*, осуществляющий планирование исполнения запроса с использованием статистической оценочной модели. Для доступа к конкретным ресурсам используются адаптеры. На текущий момент поддерживается федерализация для реляционных СУБД DB2, Oracle, Teradata и Netezza. Поддерживается реализация собственных адаптеров пользователем на языках C++ и Java.

Архитектура и подход к выполнению распределенных запросов в посредниках [1] и Big SQL обладают рядом сходных черт: слои выполнения программ, адаптеров и ресурсов; планирование запросов; настраиваемые адаптеры.

[2] Обсуждение методов и средств текстовой аналитики выходит за рамки данной статьи.

[3] Принципы построения схемы хранилища для задачи выходят за рамки данной статьи, ввиду ограниченности ее объема.

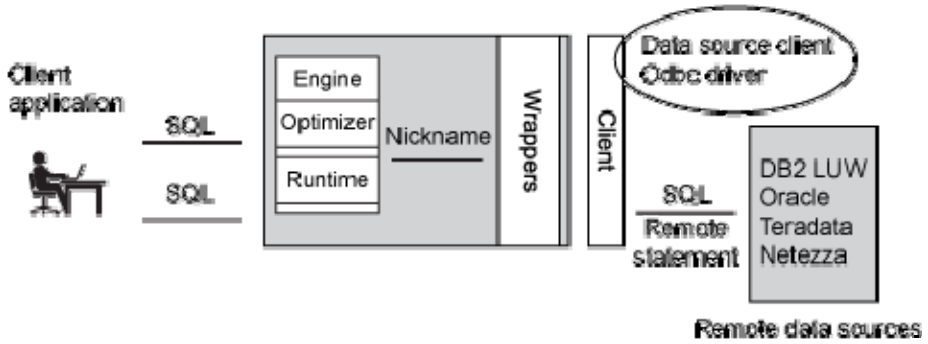



Рисунок 2. Архитектура BigSQL Federalization

Преимуществом подхода предметных посредников является ориентация на разно-модельные ресурсы. Посредники позволяют интегрировать модельно неоднородные ресурсы: реляционные, объектно-реляционные, XML, NoSQL, веб-сервисы и др. Средства BigSQL ориентированы на виртуальную интеграцию реляционных баз данных. Подробный анализ средств федерализации Big SQL и их сравнение с предметными посредниками являются задачами дальнейшей работы.

## Заключение

В статье рассмотрены основные принципы организации архитектуры комбинированной виртуально-материализованной среды интеграции больших неоднородных коллекций данных и подходы к ее реализации. Целью архитектуры является сопряжение возможностей предметных посредников по интеграции разномодельных коллекций данных и возможностей по манипулированию разноструктурированными данными, предоставляемыми платформой Hadoop и ее надстройками. Масштабирование обработки коллекций данных при помощи технологии Hadoop представляется дополнением, существенно расширяющим возможности технологии предметных посредников по решению задач над неоднородными информационными ресурсами. Реализация архитектуры, ее практическое применение и сравнение с родственными подходами являются задачами дальнейшей работы.

## Литература


[1] А.Е. Вовченко Рассредоточенная реализация приложений в среде предметных посредников : дис. … канд. техн. наук : 05.13.11. – М., 2012. – 216 с.

[2] Скворцов Н.А. Отображение моделей данных NoSQL в объектные спецификации // Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL'2012. – Переславль-Залесский: Ун-т города Переславля, 2012. – С. 78–87.

[3] Н.А. Скворцов. Отображение модели данных RDF в каноническую модель предметных посредников // Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL'2013. – Ярославль: Ярославский гос. ун-т им. П.Г. Демидова, 2013. –С. 202–209.

[4] Ступников С. А. Унификация модели данных, основанной на многомерных массивах, при интеграции неоднородных информационных ресурсов // Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL'2012. – Переславль-Залесский: Ун-т города Переславля, 2012. – С. 67–77.

[5] С.А. Ступников. Отображение графовой модели данных в каноническую объектно-фреймовую информационную модель при создании систем интеграции неоднородных информационных ресурсов // Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL'2013. – Ярославль: Ярославский гос. ун-т им. П.Г. Демидова, 2013. – С. 193–202.

[6] С.А. Ступников, Н.А. Скворцов, В.И. Будзко, В.Н. Захаров, Л.А. Калиниченко. Методы унификации нетрадиционных моделей данных. Системы высокой доступности // Радиотехника. – 2014. – Вып. 1. – С. 18–39.

[7] Apache Hadoop Project. 2014. – http://hadoop.apache.org/

[8] Apache Spark project. 2014. – http://spark.apache.org/

[9] Edward Capriolo, Dean Wampler, Jason Rutherglen. Programming Hive Data Warehouse and Query Language for Hadoop. O'Reilly Media, 2012.

[10] Kalinichenko L.A., Stupnikov S.A. OWL as Yet Another Data Model to be Integrated // Advances in Databases and Information Systems: Proc. II of the 15th East-European Conference. – Vienna: Austrian Computer Society, 2011. – P. 178–189.

[11] Kalinichenko L.A., Stupnikov S.A., Martynov D.O. SYNTHESIS: a Language for Canonical

Information Modeling and Mediator Definition for Problem Solving in Heterogeneous Information Resource Environments. Moscow: IPI RAN, 2007. – 171 p.

[12] Kalinichenko L.A., Briukhov D.O., Martynov D.O., Skvortsov N.A., Stupnikov S.A. Mediation Framework for Enterprise Information System Infrastructures. Proc. of the 9th International Conference on Enterprise Information Systems ICEIS 2007. – Funchal, 2007. – Volume Databases and Information Systems Integration. – P. 246–251.

[13] Kevin S. Beyer, Vuk Ercegovac, Rainer Gemulla, Andrey Balmin, Mohamed Eltabakh, Carl-Christian Kanne, Fatma Ozcan, Eugene J. Shekita. Jaql: A Scripting Language for Large Scale Semistructured Data Analysis. VLDB 2011.

[14] Disco framework for distributed computing based on the MapReduce paradigm. 2014. http://discoproject.org/

[15] Mauricio Hernández, Georgia Koutrika, Rajasekar Krishnamurthy, Lucian Popa, Ryan Wisnesky. HIL: a high-level scripting language for entity integration. Proceedings of the 16th International Conference on Extending Database Technology EDBT 2013. – P. 549–560.

[16] IBM InfoSphere BigInsights Information Center. 2014. – http://pic.dhe.ibm.com/infocenter/bigins/v2r1/index.jsp

[17] IBM InfoSphere Information Server Information Center. 2014. – http://pic.dhe.ibm.com/infocenter/iisinfsv/v9r1/index.jsp

[18] IBM DB2 Warehouse Information Center. 2014. http://pic.dhe.ibm.com/infocenter/db2luw/v10r5/index.jsp

[19] Introducing JSON. 2014. – http://www.json.org/

[20] Yucheng Low, Joseph Gonzalez, Aapo Kyrola, Danny Bickson, Carlos Guestrin, and Joseph M. Hellerstein (2010). "GraphLab: A New Parallel Framework for Machine Learning." Conference on Uncertainty in Artificial Intelligence (UAI).

[21] Donald Miner. MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems. O'Reilly Media, 2012.

[22] Cynthia M. Saracco, Uttam Jain. What's the big deal about Big SQL? Introducing relational DBMS users to IBM's SQL technology for Hadoop. IBM DeveloperWorks, 2013. – http://www.ibm.com/developerworks/library/bd-bigsql/bd-bigsql-pdf.pdf

[23] The Apache Hive data warehouse software. 2014. – http://hive.apache.org/

[24] The Forth Paradigm: Data-Intensive Scientific Discovery. Eds. Tony Hey, Stewart Tansley, and Kristin Tolle. Redmond: Microsoft Research, 2009. – http://goo.gl/GqkDX1

[25] The GraphLab Project. http://graphlab.org/projects/index.html

[26] The Neo4j Manual. 2014. – http://goo.gl/cHiOGF

[27] Tom White. Hadoop: The Definitive Guide. O'Reilly Media; Third Edition edition. 2012.

[28] Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, Ion Stoica. Spark: Cluster Computing with Working Sets. HotCloud 2010.

[29] Kevin Wilkinson, Craig Sayers, Harumi Kuno, Dave Reynolds. Efficient RDF Storage and Retrieval in Jena2. Proc. of the First International Workshop on Semantic Web and Databases. – 2003.

[30] Mara Elisa de Paiva Fernandes Matias. Set up and use federation in InfoSphere BigInsights Big SQL V3.0. 22 July 2014 (First published 08 July 2014). – http://www.ibm.com/developerworks/library/ba-federation-biginsights/index.html

## Combined Virtual and Materialized Environment for Integration of Large Heterogeneous Data Collections

Sergey Stupnikov, Alexey Vovchenko

Architecture of a combined virtual and materialized environment for integration of heterogeneous data collections is provided. Collections are assumed to contain structured, semi-structured or unstructured data. Combination of virtual and materialized integration is motivated by advantages and disadvantages of both approaches. Virtual integration is supported by subject mediation technology. Materialized integration is provided by Hadoop (open source software framework for storage and distributed processing of large datasets) accompanied by a system implementing relational warehouse over Hadoop (as examples, Hive and Big SQL are considered).

# Управление технологией наполнения электронной библиотеки «Научное наследие России»

Н.Е. Каленов
Библиотека по естественным наукам РАН
nek@benran.ru

**Аннотация**

В докладе рассматриваются организационная структура управления проектом создания ЭБ «Научное наследие России», основные решения по организации процесса распределенного наполнения метаданными ЭБ, контролю технологической дисциплины, получению динамической статистической информации о ходе реализации проекта.

Начиная с 2007 г., в рамках целевой научной программы РАН разрабатывается интегрированная Электронная библиотека «Научное наследие России» (ЭБ ННР), представленная в открытом доступе в Интернет (http://e-heritage.ru) [2–4]. Основной целью создания Библиотеки является информирование пользователей о выдающихся ученых, внесших вклад в развитие российской науки, и их научных достижениях.

В ЭБ ННР отражаются биографические данные об ученых, наиболее значительные их публикации (библиография и отсканированные полные тексты), архивная и музейная информация, относящаяся к деятельности ученого.

В основу функционирования ЭБ ННР положен принцип распределенной подготовки данных с централизованной редакторской обработкой, загрузкой и поддержкой контента.

Головным исполнителем работ по созданию ЭБ является Межведомственный суперкомпьютерный центр (МСЦ) РАН, разработчиками технологии и программного обеспечения – МСЦ РАН, Вычислительный центр им. А.А. Дородницына (ВЦ) РАН, Библиотека по естественным наукам (БЕН) РАН.

В число участников проекта, обеспечивающих подготовки контента для ЭБ ННР, входит достаточно большое количество академических организаций – центральные библиотеки (БАН, БЕН, ИНИОН, ЦНБ УрО РАН), использующие как собственные фонды, так и фонды своих отделений в академических институтах; Архив РАН со своими филиалами Государственный геологический музей им. В. И. Вернадского (ГГМ) РАН и ряд институтов Москвы и Санкт-петербурга.

Наряду с поставщиками контента организационная структура ЭБ как действующей автоматизированной системы включает Совет Системы, в состав которого входят авторитетные представители организаций – участников; административную группу, редакторскую группу, группу технической поддержки.

Совет Системы определяет принципы формирования контента и предоставления его пользователям, основные направления развития ЭБ, решает вопросы привлечения к созданию Библиотеки новых организаций.

Административная группа, представляющая собой временный научный коллектив в составе МСЦ РАН, осуществляет общее руководство технологическими процессами наполнения ЭБ, рассматривает планы и отчеты участников проекта.

Редакторская группа, состоящая из штатных сотрудников МСЦ РАН, принимает окончательное решение о включении в ЭБ тех или иных изданий, осуществляет «выходной контроль» подготовленной информации, включающий проверку правильности метаданных и качества (постранично) каждого оцифрованного издания, подготавливает прошедшие контроль издания для загрузки в программную оболочку демонстрационной части ЭБ, доступную пользователям по адресу http://e-heritage.ru.

Группа технической поддержки, включающая сотрудников ВЦ РАН и БЕН РАН, обеспечивает сохранность информации, работоспособность программных и технических средств, поддерживающих все элементы ЭБ.

В задачи поставщиков контента входит отбор материалов в соответствии с установленными принципами, формирование метаданных о принятых к включению в ЭБ объектах (персоналии, издания, архивные документы, музейные предметы, фотографии, мультимедийные материалы), оцифровка изданий и обработка информации в соответствии с правилами системы (для ЭБ ННР принято решение, согласно которому отсканированный текст не распознается, за

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

исключением оглавления, по которому обеспечивается навигация внутри издания), передача обработанных материалов в редакторскую группу.

Принцип распределенной подготовки данных требует разработки специальных средств, обеспечивающих исключение возможности дублирования материалов (когда разные организации сканируют одни и те же издания или вводят данные об одном и том же ученом), достаточно жесткий контроль соблюдения «технологической дисциплины» и оперативное исправление ошибок.

Специфика технологии создания ЭБ НИР еще более обостряет эти проблемы. Это обусловлено тем, что, во-первых, в качестве участников, формирующих контент, выступают не только разнородные организации, но и представители различных регионов (в настоящее время - Москва, Санкт-Петербург, Екатеринбург, Борок), а, во-вторых, тем, что подготовка информации для отражения в ЭБ включает ряд растянутых во времени процессов, в которые вовлечены различные службы.

Технологические процессы формирования контента ЭБ НИР на примере подготовки изданий (в предположении, что данные об ученом уже введены в систему) представлены в табл. 1.

Таблица 1

| Номер процесса | Наименование процесса | Организационная структура ЭБ |
|---|---|---|
| 1 | Библиографический поиск изданий конкретного автора | Поставщик контента (библиограф) |
| 2 | Формирование предложений для ввода изданий в ЭБ, передача их в редакторскую группу | Поставщик контента (библиограф) |
| 3 | Рассмотрение предложений, информирование Поставщика об их принятии или отклонении | Редакционная группа |
| 4 | Заказ и получение из фондов библиотеки принятого к сканированию издания | Поставщик контента (библиограф) |
| 5 | Формирование развернутых метаданных об издании | Поставщик контента (библиограф) |
| 6 | Сканирование издания и обработка изображений по принятой в ЭБ НИР технологии | Поставщик контента (технолог) |
| 7 | Обработка оглавления издания. связывание его с отсканированным текстом, передача материалов в редакторскую группу | Поставщик контента (технолог) |
| 8 | Контроль качества метаданных и отсканированных страниц, информирование Поставщика об ошибках или принятии издания | Редакционная группа |
| 9 | [Исправление обнаруженных ошибок] | Поставщик контента (библиограф, технолог) |
| 10 | Формирование издания для отражения в ЭБ по принятой технологии; загрузка метаданных, оглавления и полного текста издания в ЭБ | Группа технической поддержки |

Для реализации перечисленных процессов применительно ко всем объектам, отражаемым в ЭБ НИР, а также для управления процессами децентрализованной подготовки данных специалистами БЕН РАН разработана специальная система, основой которой является настраиваемый программный комплекс SCIRUS [5, 6], обеспечивающий ввод, поиск и экспорт данных, описывающих взаимосвязанные сущности, характеризуемые различным набором полей (детальное описание последней версии комплекса приведено в докладе М.М.Якшина, представленном на данной конференции).

Реализация SCIRUS, настроенная на диспетчеризацию технологии создания ЭБ (http://meta.e-heritage.ru/), поддерживает технологическую базу данных (ТБД), отражающую свойства 5 взаимосвязанных сущностей, среди которых: «публикация / документ», описываемая 24 полями метаданных; «персона», описываемая 11 полями; «организация», описываемая 6 полями; «источник» (метаданные издания на сводном уровне – сборник, журнал и т.п., в который может входить публикация), описываемый 6 полями; «файл», описываемый 2 полями (название и URL).

Функции SCIRUS, связанные с настройкой системы, вводом и корректировкой данных доступны только авторизованным пользователям в зависимости от их прав. Настроенных администратором. Однако для поиска и просмотра данных система открыта для пользователей, входящих без пароля под именем guest.

Поиск в системе возможен по любым полям всех сущностей и их булевым комбинациям.

При формировании запроса пользователь задает сущность, которая должна быть выдана в результате обработки запроса. Результатом поиска является заданная сущность, все связанные с которой сущности удовлетворяют сформулированному запросу. Например, можно получить список

публикаций, авторы которых родились в 18 веке, или список ученых в области математики, которые публиковали свои работы на французском языке (естественно. речь идет о материалах, введенных в ЭБ). Система обеспечивает навигацию по связанным записям. Например, получив в результате обработки запроса список публикаций, можно выбрать автора одной из них и получить список всех его публикаций; аналогично от списка авторов можно перейти к списку их работ.

Результаты поиска могут быть отсортированы по любому полю выдаваемой сущности.

Метаданные персоны содержат «идентифицирующие» сведения об ученом (фамилия, имя, отчество, годы жизни и т.п.), перечень научных направлений, в которых он работал, а также его развернутую биографию.

Метаданные публикации включают, наряду с элементами библиографического описания, поля, характеризующие этапы технологической обработки издания – текстовые поля «Отсутствующие страницы» и «Комментарии к проблеме» и списковые поля «Статус записи» и «Статус приемки». Значения и функции последних двух полей используются для задач диспетчеризации технологических процессов, они достаточно подробно описываются в докладе, представленном М.М.Якшиным.

Система поддержки технологии подготовки метаданных устроена так, что каждый участник может, войдя в нее по своему паролю, осуществлять поиск и просмотр всей информации, введенной в систему; вводить новые данные и редактировать старые, но только введенные под его именем. При поиске может быть задано ограничение — выбирать записи, введенные данным пользователем (в меню предлагается список имен, зарегистрированных в системе). Пользователь, наделенный правами администратора, может редактировать информацию, введенную любым участником системы.

Распределенная технология подготовки данных для ЭБ организована следующим образом (рис. 1).

Руководствуясь согласованными подходами к принципам наполнения ЭБ, каждая организация-участник Программы, определяет издания из своих фондов, которые она считает целесообразным включить в ЭБ. После этого зарегистрированный представитель этой организации входит в систему диспетчеризации и проверяет, не зарегистрирована ли уже в ней данная публикация и ее автор(ы). Если в системе публикация отсутствует, он вводит в предлагаемый шаблон ее метаданные и сведения об авторе (если автор не был введен ранее). При этом поле шаблона “Текущий статус” принимает значение “предложено к оцифровке”.

Представитель редакторской группы периодически входит в систему диспетчеризации и получает список документов, имеющих статус “Предложено к оцифровке”. Редакторская группа принимает решение по каждому из них о целесообразности ввода в ЭБ. Если документ подлежит оцифровке, значение поля “Текущий статус” меняется на “зарегистрировано”, и в запись автоматически вводится номер данного документа, под которым он будет введен в ЭБ. В дальнейшем этот номер (поле «Номер МСЦ») изменению не подлежит. Если по какой-либо причине документ оцифровывать нецелесообразно, значение поля “Текущий статус” меняется на “Оцифровке не подлежит”.

Представитель организации, формирующей контент ЭБ, входит в систему диспетчеризации и выбирает свои записи, имеющие текущий статус “зарегистрировано”. После подбора изданий и отправки на оцифровку их текущий статус меняется — в это поле вводится значение “в работе”. После завершения процесса оцифровки статус записей меняется на “оцифровано”, после передачи в редакторскую группу МСЦ — на “сдано”.

Таким образом, в каждый момент времени административная группа ЭБ может получить сведения, сколько и каких изданий находится в работе, сколько и кем оцифровано и т.п.

Технологическое поле «Статус приемки» заполняется сотрудниками редакторской группы, работающими с данным изданием. При получении электронной копии издания в это поле вводится значение «принято к работе». Если замечаний по материалу нет, он передается в техническую группу, и значение поля меняется на «принято к загрузке на сайт», после загрузки поле принимает значение «опубликовано». Если в материале обнаружены ошибки, редактор присваивает полю «Статус приемки» значение «обнаружены проблемы», заполняет поля «Отсутствующие страницы» и «Комментарии к проблеме» и направляет исполнителю соответствующее сообщение. Исполнитель исправляет ошибки, соответственно меняя значение поля «Статус приемки», после чего редактор передает издание для загрузки в ЭБ.

Таким образом контролируются ход и сроки исправления ошибок.

Загрузка метаданных о персонах и публикациях из технологической системы на сервер ЭБ осуществляется автоматически при загрузке электронного издания. Синхронизация данных осуществляется по значению поля «Номер МСЦ», при этом используется специальный формат, базирующийся на XML и RDFS, принятый для системы ЕНИП РАН [1]. Экспорт данных в этом формате возможен и путем вызова опции “Экспорт в формате ВЦ РАН” непосредственно со страницы результатов поиска данных системы SCIRUS. В этом случае перед вызовом опции необходимо отметить записи, подлежащие выгрузке.

Для наглядности схема выполнения технологических процессов создания и включения публикации в ЭБ представлена на рис. 1.

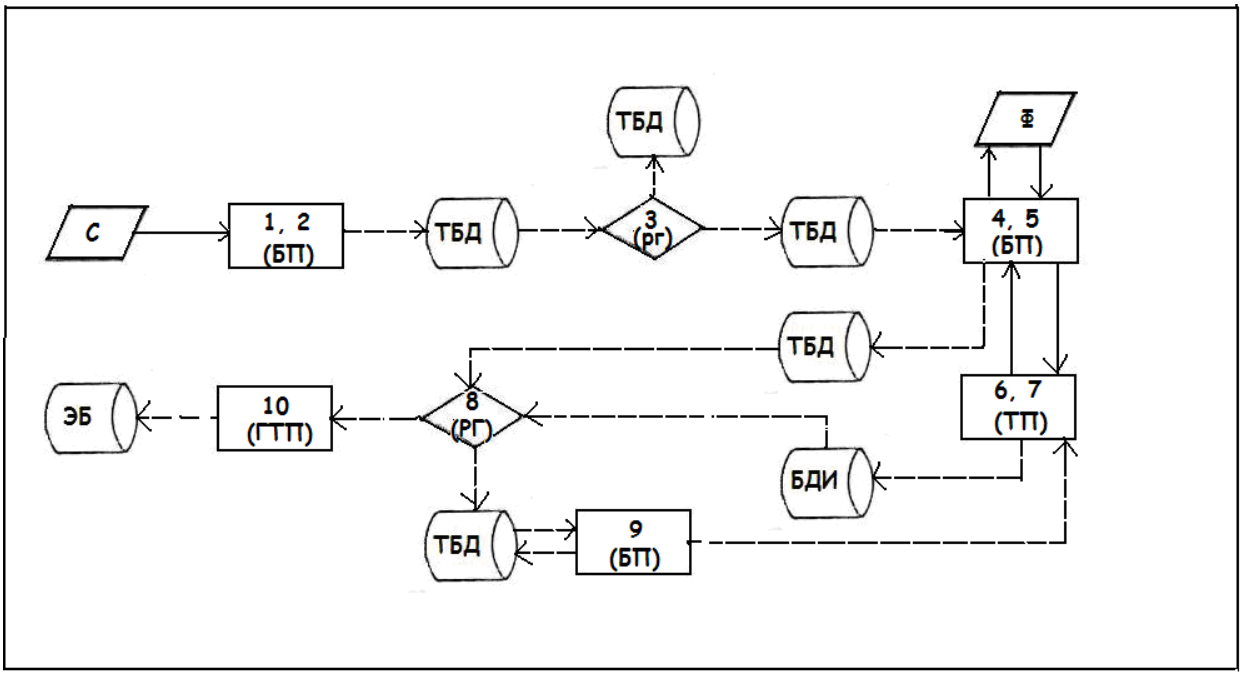



Рис. 1

Здесь приняты следующие обозначения.

Цифры внутри блоков соответствуют номерам процессов, приведенных в таблице 1.

«С» – внешние справочно-информационные ресурсы, которыми пользуются поставщики контента при поиске изданий (каталоги библиотек, библиографические издания и т.п.);

«Ф» – фонды библиотек, где хранятся печатные издания;

«БП» – библиографы поставщика;

«ТБД» – технологическая база данных;

«РГ» – редакторская группа»;

«ТП» – технологи поставщика, обеспечивающие сканирование и обработку изображений;

«БДИ» – база данных изображений (отсканированных страниц);

«ГТП» – группа технической поддержки.

Сплошные линии означают перемещение материальных объектов (издания, включаемые в ЭБ, библиографические материалы, выписки из каталогов), пунктирные – ввод данных в компьютер.

Наиболее узкое место в технологии (и это можно заметить на схеме) связано с возможностью обнаружения на этапе 8 ошибок, связанных со сканированием (в первую очередь, пропущенные страницы). В этом случае приходится снова заказывать книгу (возвращаться к этапу 4) в фондах библиотеки и направлять ее на сканирование. Эта проблема встает достаточно остро в случае, когда сканируются издания, находящиеся в удаленных хранилищах, что характерно для БЕН РАН, фонды которой распределены по 60 отделам в академических организациях московского региона. Поэтому сканирование изданий (процесс 6) требует очень тщательного контроля и должно выполняться высококвалифицированными специалистами.

В настоящее время в технологическую базу данных ЭБ «Научное наследие России» загружены данные о более чем 5000 ученых и более чем 18000 их публикаций. По многим ученым в ЭБ введены оцифрованные портреты, архивные материалы. В ЭБ также введены описания ряда, коллекций ГГМ им. В.И. Вернадского РАН, метаданные отдельных входящих в них экспонатов, ссылки на ученых, связанных с этими экспонатами.

Основное направление развития проекта связано с модификацией пользовательского интерфейса ЭБ, интеграцией с другими отечественными и зарубежными электронными библиотеками, включением в контент мультимедиа.

## Литература

[1] Бездушный А.А., Бездушный А.Н., Серебряков В.А., Филиппов В.И. Интеграция метаданных Единого Научного Информационного Пространства РАН // Материалы Международной : Монография / ВЦ РАН. – М., 2005. – С. 238.

[2] Каленов Н.Е., Савин Г.И., Сотников А.Н. Электронная библиотека «Научное наследие России» // Информационные ресурсы России. – 2009. – № 2. – С. 19–20.

[3] Каленов Н.Е., Савин Г.И., Серебряков В.А., Сотников А.Н. Принципы построения и формирования электронной библиотеки «Научное наследие России» // Программные продукты и системы. – 2012. – № 4. – С. 30–40.

[4] Каленов Н.Е., Сотников А.Н., Ильина И.Н. Архивная информация в электронной библиотеке «Научное наследие России» // Фундаментальная наука: проблемы изучения, сохранения и реставрации документального наследия научной

конференции / отв. ред. В.Ю. Афиани. – М.: Архив РАН, 2013. – С. 25–35.

[5] Сенько А.М. Информационная система SciRus: принципы построения и перспективы развития // Научный серис в сети ИНТЕРНЕТ: технологии параллельного программирования. Всероссийская науч. конф., Новороссийск, 18–23 сент. 2006. – М, 2006. – С. 58–59.

[6] Якшин М.М. WEB-интерфейс системы «Наука России» // Современные технологии в информационном обеспечении науки: Сб. науч. тр. / под ред. Н.Е.Каленова. – М., 2003. – С. 47–52.

## Management of the Technology for Filling the Digital Library Entitled “The Scientific Heritage of Russia”

Nikolay E. Kalenov

The main decisions on the management of the Digital Library project “Scientific Heritage of Russia”, on the processes of filling of distributed metadata for the DL, technological discipline of control of the DL filling are considered.

# Автоматизированное понимание таблиц на основе системы исполнения правил

© А.О. Шигаров
Институт динамики систем и теории управления СО РАН
Иркутск
shigarov@icc.ru

## Аннотация

В работе обсуждаются вопросы автоматизации процесса понимания таблиц, т.е. восстановления изначально отсутствующей в них информации о семантических отношениях (пары вида, ячейка-роль, метка-значение, метка-метка, метка-измерение). Предлагается подход, при котором понимание таблицы реализуется как исполнение правил анализа табличной структуры. На основе этого подхода разработана система для массового преобразования неструктурированной табличной информации, представленной в формате табличного процессора Excel, к структурированному виду. Результатом понимания таблиц являются структурированные данные — таблицы в канонической форме, которые структурно соответствуют таблицам реляционной базы данных. Полученные экспериментальные результаты показывают эффективность применения предлагаемого подхода для широкого класса сводных таблиц из статистических отчетов.

## 1 Введение

По оценки исследователей Merrill Lynch [16] примерно 80 процентов всей бизнес информации представлено в неструктурированном виде. Такая информация не имеет предопределенной формальной модели данных (например, научная статья, финансовый отчет, сообщение электронной почты) [1] и является противоположностью структурированной информации (например, реляционным базам данных).

Многие исследователи, в том числе, W. Inmon [11-12], отмечают важность вопросов интеграции неструктурированной информации. Одним из наиболее интересных вопросов является интеграция неструктурированных текстов, включая таблицы. Многие слабоструктурированные (ASCII-текст, файлы печати PDF и др.) и полуструктурированные (документы Word, книги Excel, HTML страницы и др.) документы [7] содержат таблицы. Такие таблицы главным образом адресованы для восприятия человеком. Они не предназначены напрямую для высокоуровневой машинной обработки, например, выполнения запросов к данным по аналогии с SQL (Structured Query Language). Поэтому они также являются примером неструктурированной информации.

На практике решения многих задач связаны с необходимостью извлекать информацию из таких таблиц и загружать её в базы данных. Поскольку, таблицы, представленные в неструктурированном виде, часто оказываются единственным доступным источником информации. Только после преобразование такой табличной информации к структурированной форме она становится доступной для использования в бизнес-аналитике, включая, аналитическую обработку в реальном времени (OLAP), интеллектуальный анализ данных, и извлечение знаний.

В литературе рассматриваются следующие задачи, которые являются преобразованием неструктурированной табличной информации к структурированному виду.

1) Каноникализация таблицы [2, 19] — приведение её к канонической форме, которая структурно соответствует таблице реляционной базы данных.

2) Извлечение информации из таблицы [5] является аналогом задачи извлечения информации из текста и состоит в выборочном извлечении фактов, формирующих целевую базу данных.

3) Понимание таблицы [5, 9] состоит в восстановлении отношений между метками (заголовками) и значениями данных, а также между метками и измерениями (доменами).

Как определяется в работе [9] понимание таблиц в общем случае включает следующие этапы: (1) обнаружение таблицы (поиск позиций ограничивающего прямоугольника таблицы внутри источника); (2) распознавание таблицы (разделение её на отдельные ячейки); (3) функциональный

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

анализ (определение того, какую роль играет ячейка в таблице); (4) структурный анализ (определение связей между ячейками); и (5) интерпретацию таблицы (извлечение фактов из таблицы). В настоящей работе обсуждается автоматизация следующих из перечисленных этапов понимании таблиц: (3) функционального и (4) структурного анализ, и (5) интерпретации таблицы.

## 2 Родственные работы

Существует огромное разнообразие способов изображения таблиц. Это приводит к высокой сложности анализа и обработки неструктурированной табличной информации. Как показано в обзорах [4, 5, 13, 22], посвященных проблемам анализа и обработки таблиц, сейчас наиболее изучены, хотя и не решены полностью, проблемы обнаружения и распознавания таблиц. При этом проблемы высокоуровневого анализа и интерпретации таблиц остаются менее изученными.

Вопросы понимания таблиц, связанные с задачами их (3) функционального и (4) структурного анализ, а также (5) интерпретации, рассматриваются в ряде работ [2, 4, 6, 8, 10, 19, 23–24]. Ниже приводится анализ некоторых из них.

В работах Douglas S. и др. [2] и Tijerino Y. и др. [19] рассматривается преобразование (структурирование) табличной информации, называемое каноникализацией таблицы. В работе Douglas S. и др. предлагается метод интерпретации и каноникализации таблиц, которые содержатся в спецификациях, используемых в строительной промышленности. Для этого они предлагают использовать обработку естественного языка на основе онтологии предметной области (подъязыка спецификаций строительной промышленности).

Предлагаемый Tijerino Y. и др. [19] способ каноникализации основан на использовании библиотеки фреймов, содержащей знания о лексическом содержании таблиц. Каждый фрейм данных описывает один тип данных и используется для отнесения выражений на естественном языке (табличных заголовков и значений) к этому типу. Для описания типов данных ими предлагается использовать регулярные выражения, словари и некоторые открытые ресурсы, например, WordNet [21].

В перечисленных работах [2–19] предлагаются методы каноникализации таблиц, основанные на анализе и интерпретации представленной в таблицах естественно-языковой информации. На практике этого не всегда достаточно, для более точного и полного извлечения информации из таблицы часто также требуется анализ пространственной и графической информации.

W. Gatterbauer и др. в работе [8] напротив предлагают предметно-независимый метод извлечения информации из HTML таблиц, основанный на анализе исключительно пространственной и стилевой информации в формате CSS2 (Cascading Style Sheets Level 2). В частности, ими предлагается выполнять интерпретацию таблиц (восстановление семантических отношений) на основе эвристик о стилевой информации подготовленного для набора наиболее общих типов изображения web-таблиц.

В работе D.W. Embley и др. [6] предлагаются методы обнаружения таблиц внутри HTML страниц, и извлечения из них информации. При этом предполагается, что таблица может включать вложенные таблицы на связанных страницах. В частности, для поиска атрибутов (меток) и значений (данных) среди содержания ячеек таблицы предлагается использовать онтологии, специально разрабатываемые для извлечения данных. Такие онтологии извлечения помимо понятий (объектов), отношений и ограничений содержат привязанные к объектам фреймы, которые с помощью регулярных выражений позволяют связать содержание таблицы с объектами онтологии. Для связывания атрибутов со значениями, дополнительно к онтологиям извлечения используется набор эвристик о пространственной структуре и содержании таблиц.

В отличие от приведенных исследований нами предлагается автоматизировать понимание таблиц за счет анализа и интерпретации, как их естественно-языковой, так и пространственной и графической (стилевой) информации.

## 3 Представление фактов о таблицах

Для понимания таблиц нами предлагается подход, основанный на исполнении правил анализа структуры таблиц. Идея, лежащая в основе предлагаемого подхода, состоит в следующем. Обычно внутри тематической коллекции документов от одного поставщика таблицы компонуются и форматируются однообразно. Для такой коллекции документов можно определить набор формализованных правил анализа табличной структуры, который удовлетворяет всем или почти всем ее таблицам. Эти правила можно представить в виде базы знаний, а процесс восстановления семантических отношений в таблице реализовать как логический вывод. При этом база фактов, используемая в процессе логического вывода, может включать информацию о пространственном, графическом и естественно-языковом содержании таблицы.

### 3.1 Базовые предположения о таблицах

На основе ограничений табличной структуры, характерных для представлений табличной информации в широко распространенных форматах данных, таких как Excel, Word, HTML и LaTeX, предлагается достаточно общая модель таблицы CELLS, которая ориентирована на представление фактов о табличной информации в процессе логического вывода. В модели сделано несколько общих для этих представлений предположений.

1) Ячейка может располагаться в одной или нескольких соседних строках и в одном или нескольких соседних столбцах (например, атрибуты COLSPAN и ROWSPAN в HTML) и имеет прямоугольную форму в пространстве строк и столбцов, как показано на рис. 1, *а*.

2) Внутри ячейки не могут располагаться другие ячейки или таблицы (это не допускается в Excel).

3) Содержимое ячейки может являться либо меткой (заголовком), либо вхождением (данными). Используемые здесь термины «вхождение» и «метка» соответствуют смыслу терминов «entry» и «label» соответственно из работы Wang X. [20].

4) Метки могут адресовать вхождения либо в строках — метки строк, либо в столбцах — метки столбцов.

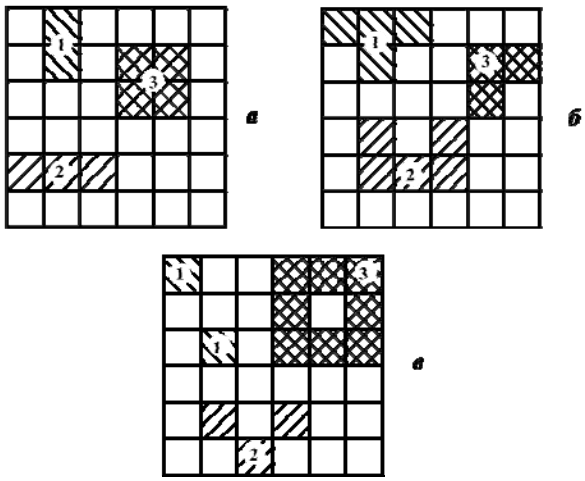



Рис. 1. Примеры объединения плиток сетки в ячейки таблицы, обозначенные как 1, 2 и 3: так ячейка может объединять несколько плиток в Excel, Word, HTML и LaTeX (*а*); так ячейка может визуально (для восприятия человеком) включать несколько плиток с помощью разграфки (*б*); скорее всего, так ячейки никто не представляет (*в*)

Очевидно, что сделанные предположения описывают широкий класс обрабатываемых таблиц. Пример сводной таблицы, полностью укладывающейся в данную модель, приводится на рис. 2.

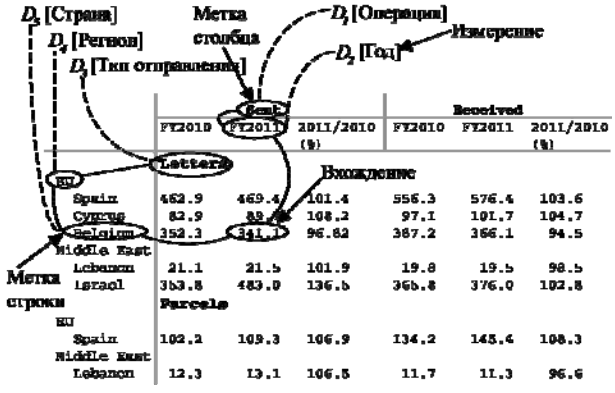



Рис. 2. Пример сводной таблицы

### 3.2 Модель таблицы

Модель включает два уровня: физической и логической структуры, которые в упрощенном виде можно описать следующим образом.

1) Уровень физической структуры $Tp=(Sr, Sc, C)$ состоит из: (1) пространства строк — $Sr$ и столбцов — $Sc$; (2) набора ячеек — $C$, в котором каждая ячейка — $c=(p, c', S)$ включает: координаты в пространстве строк $Sr$ и столбцов $Sc$ — $p=(cl, rt, cr, rb)$, содержание — $c'$, стилевая информация (цветовые схемы, шрифтовые метрики, выравнивание, стили оформлении границ и др.) — $S$.

2) Уровень логической структуры $Tl=(D, Lr, Lc, E)$ состоит из: (1) набора представленных в обрабатываемой таблице измерений — $D=\{Di\}$, каждое из которых содержит значения $Di=\{dj\}$; (2) дерева меток строк — $Lr$ и (3) столбцов — $Lc$, отражающих связи между метками, не являющимися значениями измерений $Di$ из набора $D$ — $l=(l')$, где $l'$ — содержание метки; (4) набора вхождений — $E$, в котором каждое вхождение — $e=(e', D', L')$ включает: содержание — $e'$, набор связанных с ним значений измерений $Di$ из набора $D$ — $D'$, набор связанных с ним меток из деревьев $Lr$ и $Lc$ — $L'$.

### 3.3 Структуры данных

Предлагаемая в работе модель таблицы реализована в виде ряда структур данных, основные из которых перечислены далее: CELL, ENTRY, LABEL, LABELNODE. Структура CELL предназначена для представления ячейки и прежде всего информации о её физической структуре, однако она также включает уровень логической структуры ячейки (т.е. она позволяет накапливать информацию о ее связях с другими ячейками, ее роли и типе данных). На практике это позволяет разрабатывать правила анализа табличной структуры в более лаконичной манере по сравнению со случаем, при котором используются дополнительные структуры данных для представления информации уровня логической структуры. Структуры ENTRY, LABEL, LABELNODE используются исключительно на уровне логической структуры. ENTRY служит для представления вхождения, а LABEL — метки. Структура LABELNODE является оболочкой для структуры LABEL и обеспечивает представление деревьев меток.

Все предложенные структуры данных и алгоритмы реализованы на платформе Java. Это обеспечивает возможность использовать их напрямую для представления фактов о таблицах в процессе логического вывода, выполняемого в системе исполнения правил с поддержкой спецификации JSR-94 (Java Rule Engine API).

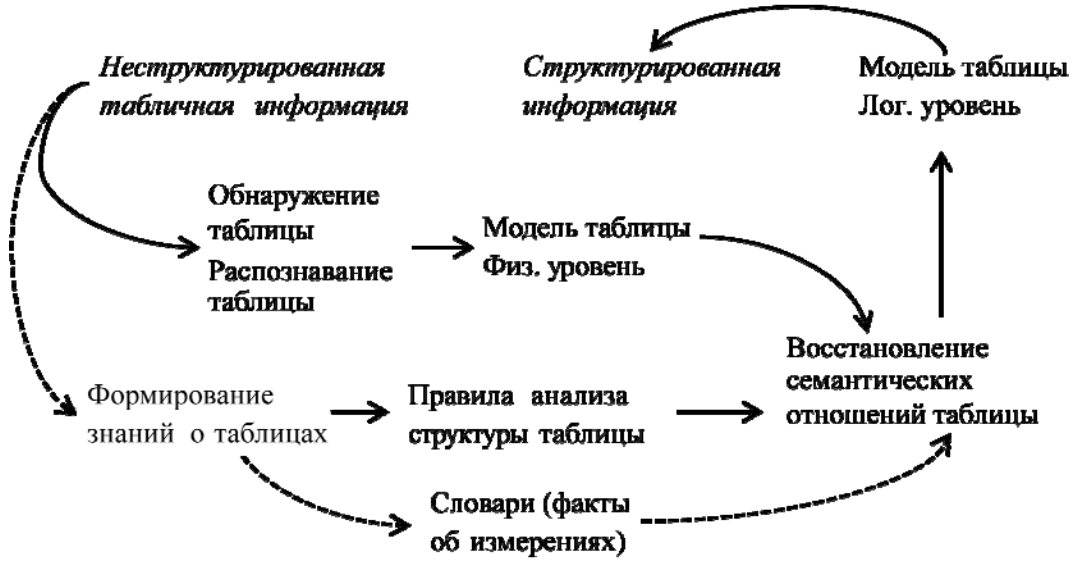



Рис. 3. Схема структурирования табличной информации

*а*

```
...
when
    $c : CCell( cl == 1, style.getFont().getColor() == "#ff0000")
then
    modify ( $c ) { setRole( Role.ROWLABEL ) }
```

*б*

```
...
when
    $c1 : CCell()
    $c2 : CCell( rt == $c1.rb + 1, ( $c1.cl <= cl && cr < $c1.cr ) ||
                ( $c1.cl < cl && cr <= $c1.cr ) )
then
    $c1.addConnectedCell( $c2 )
```

*в*

```
...
when
    $c : CCell( text matches "(?i).*(total)" )
then
    modify ( $c ) { setIgnored( true ) }
...
```

Рис. 4. Примеры правил анализа табличной структуры

## 4. Представление и исполнение правил анализа табличной структуры

Схема преобразования табличной информации от неструктурированной к структурированной форме показана на рис. 3. Предполагается, что этапы обнаружения и распознавания таблицы выполняются в сторонних системах. Например, для извлечения таблиц из PDF документов могут использоваться системы Tabula [18] или PDFGenie [15], для документов, напечатанных в файлы формата EMF, может использоваться технология, предложенная в работах [24]. Выходом таких систем являются таблицы в форматах Excel, HTML или XML, которые могут быть приведены к физическому уровню модели CELLS.

В процессе загрузки таблиц из полученных файлов Excel, HTML или XML в структуры данных, реализующих модель CELLS, табличная информация подвергается предобработки. Это включает опционально: удаление лишних пробельных и служебных символов из текстового содержания, исключение из таблицы пустых строк и столбцов и восстановление отсутствующих настроек стилей границ ячеек. Последнее необходимо, поскольку видимые и физические границы ячейки не всегда совпадают. Визуально они могут быть образованы границами соседних ячеек. Приведение стилей физических границ ячеек в соответствии с её

видимыми границами позволяет упростить правила анализа структуры таблицы.

Полученные в результате данные о таблице, которые формируют базу фактов для логического вывода. Кроме того, факты могут быть дополнены внешней информацией об измерениях.

Для обработки набора таблиц формируется база знаний, которая состоит из продукционных правил анализа табличной структуры. Они отображают доступную информацию: позиции (координаты), графическое форматирование и естественно-языковое содержание ячеек, в отсутствующие изначально отношения между метками, вхождениями и измерениями. Полученные в процессе вывода новые факты о семантических отношениях должны быть достаточными для каноникализации таблицы.

В качестве система исполнения таких правил может использоваться свободная системы Drools Expert [3], реализующая спецификацию JSR-94. При этом сами правила могут быть представлены на языке выражений MVEL [14].

На Рис. 4 приводится ряд простых примеров возможных правил анализа структуры на языке MVEL. Если ячейка $c находится в 1-ом столбце, а её текст выделен красным цветом, то она выполняет роль метки строки (рис. 4, *а*). Если ячейка $c1 расположена непосредственно над ячейкой $c2 и при этом полностью охватывает её по столбцам, то они связаны (рис. 2, *б*). Если ячейка $c содержит текст, удовлетворяющий регулярному выражению "(?i).*(total)", то её необходимо игнорировать при формировании выходных данных (рис. 2, *в*). Примеры правил, которые применялись при тестировании системы CELLS, можно найти по адресу http://cells.icc.ru/test.

В процессе логического вывода накапливается информация о логической структуре таблицы. Для этой информации выполняется постобработка, которая включает: приведение текстового содержания ячеек к эталонным написаниям, сопоставление меток с измерениями и формирование канонической формы таблицы.

Из восстановленной информации модели таблицы CELLS формируется таблица в канонической форме, которая включает следующие поля: DATA — данные (вхождения); ROW_LABEL — пути меток от листьев до корней из невырожденного дерева *Lr*; COL_LABEL — пути меток от листьев до корней из невырожденного дерева *Lc*; *D1,..., Dn* — поля значений измерений *Di* из набора *D*. Каждый кортеж в такой канонической форме представляет связь между вхождением, путями в деревьях меток и значениями восстановленных измерений. Дополнительно поле ROW_LABEL/COL_LABEL может быть разделено на несколько отдельных полей, каждое из которых будет соответствовать одному уровню вложенности в дереве меток строк/столбцов.

Пример канонической формы обработанной таблицы приводится на Рис. 5. Сформированная каноническая таблица может экспортироваться в реляционную базу данных с помощью стандартных средств интеграции данных известных систем управления базами данных (СУБД). Например, службы "SQL Server Integration Services" [17], позволяют импортировать данные из таблиц с простой "решеточной" структурой в форматах Excel, CSV в базы данных под управлением СУБД "SQL Server".

| *Данные* | *Операция* | *Год* | *Тип отправления* | *Регион* | *Страна* |
|---|---|---|---|---|---|
| 462.9 | Sent | 2010 | Letters | EU | Spain |
| 82.9 | Sent | 2010 | Letters | EU | Cyprus |
| ... | ... | ... | ... | ... | ... |
| 12.3 | Sent | 2010 | Parcels | Middle East | Lebanon |
| 469.4 | Sent | 2011 | Letters | EU | Spain |
| 89.7 | Sent | 2011 | Letters | EU | Cyprus |
| 341.1 | Sent | 2011 | Letters | EU | Belgium |
| 21.5 | Sent | 2011 | Letters | Middle East | Lebanon |
| 483.0 | Sent | 2011 | Letters | Middle East | Israel |
| 109.3 | Sent | 2011 | Parcels | EU | Spain |
| 13.1 | Sent | 2011 | Parcels | Middle East | Lebanon |
| 556.3 | Received | 2010 | Letters | EU | Spain |
| ... | ... | ... | ... | ... | ... |
| 11.3 | Received | 2011 | Parcels | Middle East | Lebanon |

Рис. 5. Каноническая форма таблицы из рис. 1: все метки сопоставлены измерениям, поэтому поля COL_LABEL и ROW_LABEL отсутствуют

## 3 Экспериментальные результаты

Экспериментальная оценка представленного подхода выполнена с помощью системы CELLS, в которой реализованы структуры данных, представляющие модель таблицы CELLS, и алгоритмы: 1) загрузки исходной табличной информации в формате Excel (тестовых данных со специальной разметкой); 2) структурирования табличной информации, восстановленной в процессе логического вывода; 3) экспорта результатов в формате Excel.

Для экспериментальной оценки сформирована коллекция тестовых данных, которая включает 97 таблиц в формате Excel, собранных из 7 различных источников. Коллекция доступна по адресу http://cells.icc.ru/test. Её краткое описание приводится в табл. 1. Для формирования коллекции исходная табличная информация была преобразована из формата PDF в Excel.

Источниками тестовых данных послужили слабоструктурированные документы в низкоуровневом формате файлов печати PDF — государственные и финансовые статистические отчеты с богатым табличным содержанием. Для формирования коллекции исходная табличная информация была преобразована из формата PDF в Excel. При этом, насколько это было возможно, в полученных тестовых таблицах было сохранено графическое форматирование, представленное в соответствующих им PDF источниках.

Таблица 1. Экспериментальные результаты

| Код источника | Кол-во таблиц | Кол-во ячеек | Кол-во вхождений | Кол-во меток | Кол-во связей между метками* | Кол-во правил | Время исполнения правил (мс) |
|---|---|---|---|---|---|---|---|
| JAPAN_STAT [1] | 15 | 1088 | 734 | 257 | 102 | 10 | 417 |
| AEROFLOT [2] | 13 | 2047 | 727 | 321 | 167 | 16 | 526 |
| BOEING [3] | 21 | 2156 | 964 | 470 | 196 | 14 | 663 |
| CHINA_STAT [4] | 18 | 7216 | 4180 | 862 | 551 | 12 | 964 |
| CHEVRON [5] | 7 | 812 | 268 | 141 | 89 | 12 | 283 |
| USDA_NASS [6] | 7 | 1553 | 1175 | 313 | 174 | 16 | 638 |
| TOBACCO [7] | 16 | 2844 | 2195 | 508 | 335 | 10 | 730 |

[1] Statistical Handbook of Japan 2007. Statistics Bureau of Japan. Chapter 5, 8.
[2] OJSC «Aeroflot – Russian Airlines» Consolidated Financial Statements For the Year Ended December 31, 2006. P. 4–10, 25–26.
[3] Boeing Co, Annual Report 2010. P. 50–55, 83–85.
[4] China statistical yearbook 2003. National Bureau of Statistics of China. P. 23–48, 555, 559, 571, 584, 590, 664, 708, 774, 765.
[5] Chevron Corp. News Release November 2, 2012. Chevron Corp. P. 1, 5–9.
[6] USDA NASS. 2003 Agricultural Statistics Annual. USDA (U.S. Department of Agriculture). National Agricultural Statistics Service. Chapter VI. P. 5–7, 12.
[7] Tobacco: World Markets and Trade 2005. USDA (U.S. Department of Agriculture). Foreign Agricultural Service.
* Исключая связи корней деревьев меток.

Тестовые данные имеют дополнительную разметку для определения местоположения таблицы внутри листа Excel (рис. 6), а также аккуратную декомпозицию на ячейки. Там, где это возможно, их физическая структура и разграфка совпадают. Это позволяет избежать этапов обнаружения и сегментации таблицы.

| $START | | | | | | |
|---|---|---|---|---|---|---|
| | Company name | Place of incorporation and operation | Activity | Percentage held as of December 31, 2006 | Percentage held as of December 31, 2005 | |
| | LLC "Airport Moscow" | Moscow region | Cargo handling | 50,00% | 50,00% | |
| | CJSC "Aerofirst" | Moscow region | Trading | 33,30% | 33,30% | |
| | CJSC "TZK Sheremetyevo" | Moscow region | Fuel trading company | 31,00% | 31,00% | |
| | CJSC "AeroMASH – AB" | Moscow region | Aviation security | 45,00% | 45,00% | |
| | | | | | | $END |

Рис. 6. Дополнительная разметка тестовой таблицы: маркеры «$START» и «$END» указывают соответственно верхний левый и нижний правый угол таблицы в пространстве строк и столбцов

На рис. 7 показаны некоторые таблицы из тестовой коллекции данных. Их структуры включают типичные для этой коллекции особенности. Так, таблица, рис. 7, *а*, содержит иерархии меток строк и столбцов. Тело таблицы, рис. 7, *б*, пересекают перерезы: «Price per 100 pounds» и «Price per bushel». В таблице, рис. X, *в*, столбцы с метками строк чередуются со столбцами с данными.

Полученные экспериментальные результаты приводятся в табл. 1. Логический вывод выполнялся в системе Drools Expert (5.4.0.Final). При этом использовался процессор Intel Core 2 Quad, 2,66 ГГц. Экспериментальные результаты показывают эффективность применения предлагаемого подхода для широкого класса таблиц.

## 4. Заключение

Предлагаемый подход базируется на предположении о том, что для одного или нескольких схожих источников можно разработать непротиворечивый набор правил анализа структуры содержащихся в них таблиц. Однако разработка достаточно универсальных баз знаний для многих разнородных источников имеет слишком высокую цену и не всегда возможна из-за противоречий, содержащихся в самих источниках. Поэтому данный подход предназначен в основном для задач управления данными, прежде всего для массовой интеграции табличной информации из наборов похожих источников.

| Item | Total | National forest | Non-national forest | | |
|---|---|---|---|---|---|
| | | | Municipal | Private | Others |
| Forest land area (1,000 ha) | 25 121 | 7 838 | 2 796 | 14 440 | 46 |
| Forest growing stock (1 mil. m3) | 4 040 | 1 011 | 433 | 2 590 | 5 |
| Planted forests | | | | | |
| Land area (1,000 ha) | 10 361 | 2 411 | 1 232 | 6 705 | 12 |
| Growing stock (1 mil. m3) | 2 338 | 368 | 255 | 1 712 | 3 |
| Natural forests | | | | | |
| Land area (1,000 ha) | 13 349 | 4 770 | 1 426 | 7 126 | 27 |
| Growing stock (1 mil. m3) | 1 701 | 642 | 178 | 878 | 3 |

а

| Kind of seed | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 |
|---|---|---|---|---|---|---|---|---|---|---|
| | Price per 100 pounds | | | | | | | | | |
| | *Dollars* | *Dollars* | *Dollars* | *Dollars* | *Dollars* | *Dollars* | *Dollars* | *Dollars* | *Dollars* | *Dollars* |
| Alfalfa, uncertified varieties | 152.00 | 161.00 | 168.00 | 185.00 | 185.00 | 205.00 | 184.00 | 165.00 | 158.00 | 280.00 |
| Alfalfa, certified varieties | 269 | 266 | 274 | 277 | 282 | 288 | 287 | 277 | 278 | 157 |
| Clover, ladino | 324 | 321 | 320 | 318 | 307 | 308 | 298 | 285 | 285 | 280 |
| Clover, red | 148 | 148 | 134 | 172 | 184 | 194 | 178 | 143 | 132 | 130 |
| Lespedeza, Korean | 132 | 84,5 | 66 | 99 | 90 | 89 | 76,15 | 77,5 | 160 | 98 |
| Sunflower | 300 | 297 | 297 | 313 | 355 | 380 | 400 | 395 | 407 | 407 |
| Cottonseed, all | 62,7 | 63,5 | 68,2 | 73 | 74,9 | 79,3 | 82,4 | 128 | 154 | 213 |
| Biotech [1] | | | | | | | | | 217 | 271 |
| Non-biotech | | | | | | | | | 87 | 94 |
| Grain sorghum, hybrid | 74,5 | 82,1 | 78,7 | 84 | 92 | 96 | 97,6 | 93 | 93 | 96 |
| | Price per bushel | | | | | | | | | |
| | *Dollars* | *Dollars* | *Dollars* | *Dollars* | *Dollars* | *Dollars* | *Dollars* | *Dollars* | *Dollars* | *Dollars* |
| Corn, hybrid, all [2] | 72,7 | 73,4 | 77,1 | 77,7 | 83,5 | 86,9 | 88,1 | 87,5 | 92,2 | 92 |
| Biotech [1] | | | | | | | | | 110 | 113 |
| Non-biotec | | | | | | | | | 85,3 | 85,8 |
| Wheat (spring) | 5,98 | 7,37 | 7,12 | 8,1 | 7,3 | 6,85 | 6,1 | 6,1 | 6,2 | 6,5 |
| Wheat (winter) | 7,73 | 7,9 | 7,8 | 8,5 | 10 | 8,25 | 7,35 | 7,05 | 7,2 | 7,7 |
| Rice | 15,4 | 22 | 15,1 | 17,5 | 19 | 19,5 | 19,1 | 17,25 | 15,7 | 14,9 |
| Barley (spring) | 5 | 5,18 | 5,37 | 6,49 | 6,13 | 6,04 | 5,8 | 5,8 | 5,8 | 5,8 |
| Soybeans for seed, all | 12,4 | 13,6 | 13,4 | 14,8 | 16,1 | 17,15 | 17 | 17,1 | 20,7 | 22,5 |
| Biotech [1] | | | | | | | | | 23,9 | 27 |
| Non-biotec | | | | | | | | | 17,9 | 15 |
| Flaxseed | 7,37 | 7,74 | 8 | 8,14 | 9,31 | 10 | 8,5 | 7,9 | 7,6 | 7,6 |

б

| 线路名称 | Name | 客运量（万人）Passenger Traffic (10 000 persons) | 旅客周转量（百万人公里）Passenger-kilometers (million passenger-km) | 线路名称 | Name | 货运量（万吨）Freight Traffic (10 000 tons) | 货物周转量（百万吨公里）Freight Ton-kilometers (million ton-km) |
|---|---|---|---|---|---|---|---|
| 京沪线 | Beijing-Shanghai | 5496 | 32975 | 京沈线 | Beijing-Shenyang | 3438 | 82790 |
| 新石线 | Xinjiang-Rizhao | | | 哈大线 | Harbin-Dalian | 3233 | 60717 |
| 沪杭 | Shanghai-Hangzhou | 654 | 6188 | 津沪线 | Tianjin-Shanghai | 5304 | 100909 |
| 浙赣线 | Hangzhou-Ganzhou | 3785 | 33028 | 沪杭线 | Shanghai-Hangzhou | 202 | 4939 |
| 鹰厦线 | Yingtan-Xiamen | 10869 | 88717 | 京广线 | Beijing-Guangzhou | 7187 | 131196 |
| 京九线 | Beijing-Kowloon | 814 | 1906 | 南北同蒲线 | Datong-Taiyuan-Fenglingdu | 11168 | 30412 |
| 京广线 | Beijing-Guangzhou | 491 | 1708 | 太焦柳线 | Taiyuan-Jiaozuo-Liuzhou | 8206 | 56729 |
| 石太线 | Shijiazhuang-Taiyuan | 1800 | 9364 | 京九线 | Beijing-Kowloon | 2644 | 61919 |
| 石德线 | Shijiazhuang-Dezhou | 1575 | 6452 | 兰新线 | Lanzhou-Urumqi | 3366 | 63348 |
| 焦柳线 | Jiaozuo-Liuzhou | 655 | 2512 | 滨洲线 | Harbin-Manzhouli | 3137 | 21181 |
| 京包线 | Bingjing-Baotou | 541 | 1288 | 滨绥线 | Harbin-Suifenhe | 1178 | 16384 |
| 包兰线 | Baotou-Lanzhou | 1245 | 3615 | 京包线 | Bingjing-Baotou | 5881 | 57077 |
| 北同蒲线 | Taiyuan-Datong | 4759 | 33838 | 石太线 | Shijiazhuang-Taiyuan | 3760 | 21301 |
| 南同蒲线 | Fenglingdu-Taiyuan | 74 | 1459 | 石德线 | Shijiazhuang-Dezhou | 379 | 11664 |
| 陇海线 | Lianyungang-Lanzhou | 1055 | 16149 | 浙赣线 | Hangzhou-Ganzhou | 2464 | 45035 |
| 宝中线 | Baoji-Zhongwei | 413 | 1865 | 陇海线 | Lianyungang-Lanzhou | 6357 | 100027 |

в

Рис. 7. Примеры тестовых таблиц

Подход положен в основу развиваемой авторами системы понимания таблиц в формате Excel. Полученные экспериментальные результаты показывают эффективность её применения для широкого класса таблиц, представленных в формате Excel. В то же время необходимо дальнейшее исследование возможностей для упрощения правил анализа структуры таблицы за счет развития структур данных представления табличной информации и дополнительных алгоритмов её предобработки и постобработки.

Работа выполнена при финансовой поддержке РФФИ грант № 14-07-00166 и Совета по грантам Президента РФ СП-3387.2013.5.

## Литература


[1] Blumberg R., Atre S. The problem with unstructured data // DM Review, 2003. http://soquelgroup.com/Articles/dmreview_0203_problem.pdf

[2] Douglas S., Hurst M., Quinn D. Using Natural Language Processing for Identifying and Interpreting Tables in Plain Text // Proc. of the 4th Annual Symposium on Document Analysis and Information Retrieval. Las Vegas. 1995. P. 535–546.

[3] Drools Expert (JBoss Community). http://www.jboss.org/drools/drools-expert.html

[4] e Silva A.C., Jorge A.M., Torgo L. Design of an end-to-end method to extract information from tables // Int. J. on Document Analysis and Recognition. 2006. Vol. 8, No. 2. P. 144–171.

[5] Embley D.W., Hurst M., Lopresti D., Nagy G. Table-processing paradigms: a research survey // Int. J. on Document Analysis and Recognition. 2006. Vol. 8, No. 2. P. 66–86.

[6] Embley D.W., Tao C., Liddle S.W. Automating the Extraction of Data from HTML Tables with Unknown Structure // Data & Knowledge Engineering. Elsevier. 2005. Vol. 54, No. 1. P. 3–28.

[7] Feldman R., Sanger J. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data // Cambridge University Press. 2006. 422 p.

[8] Gatterbauer W., Bohunsky P., Herzog M., Krüpl B., Pollak B. Towards Domain-Independent Information Extraction from Web Tables // Proc. of the 16th Int. Conf. on World Wide Web. ACM New York, NY, US, 2007. P. 71–80.

[9] Hurst M. Layout and Language: Challenges for Table Understanding on the Web // In Proc. of the 1st Int. Workshop on Web Document Analysis. 2001. P. 27–30.

[10] Hurst M. The Interpretation of Tables in Texts. PhD thesis. School of Cognitive Science, Informatics, the University of Edinburgh. UK, 2000.

[11] Inmon W.H. Matching unstructured data and structured data // The data administration newsletter. 2006. http://www.tdan.com/view-articles/5009.

[12] Inmon W.H., Nesavich A. "Tapping into Unstructured Data: Integrating Unstructured Data and Textual Analytics into Business Intelligence", 1st edition, Prentice Hall PTR, 2007.

[13] Lopresti D., Nagy G. A tabular survey of automated table processing // Lecture Notes in Computer Science. 2000. Vol. 1941. P. 93–120.

[14] MVEL. http://mvel.codehaus.org

[15] PDFGenie, http://www.pdftron.com/pdfgenie

[16] Shilakes C.C., Tylman J. Enterprise Information Portals // Merrill Lynch. 1998.

[17] SQL Server Integration Services, http://msdn.microsoft.com/ru-ru/library/ms141026.aspx

[18] Tabula, http://tabula.nerdpower.org

[19] Tijerino Y., Embley D., Lonsdale D., Nagy G. Towards ontology generation from tables // World Wide Web: Internet and Web Information Systems. 2005. Vol. 8, No. 3. P. 261–285.

[20] Wang X. Tabular Abstraction, Editing, and Formatting. PhD thesis. Waterloo, Ontario, Canada. 1996.

[21] WordNet, http://wordnet.princeton.edu

[22] Zanibbi R., Blostein D., Cordy J.R. A survey of table recognition: Models, observations, transformations, and inferences // Int. J. on Document Analysis and Recognition. 2004. Vol. 7, No. 1. P. 1–16.

[23] Кудинов П.Ю. Адаптивные методы извлечения информации из статистических таблиц, представленных в текстовом виде : дис. … канд. техн. наук. М., 2011. С. 105.

[24] Шигаров А.О. Технология извлечения табличной информации из электронных документов разных форматов : дис. … канд. техн. наук. Иркутск, 2009. С. 143.


## Automated Table Understanding Using a Rule Engine

Alexey O. Shigarov

The paper discusses issues on automation of the table understanding (i.e. recovering relationships of table elements). We propose an approach to table understanding based on the use of a rule engine. A table model oriented on the logical inference and algorithms for processing tabular information are also considered in the paper. The CELLS system for structuring tabular information presented in Excel spreadsheet format has been developed using the proposed approach, model and algorithms. The performance evaluation of the system shows that the approach can be applied to a wide range of tables.

# Модель мышления и понимания в автоматической обработке запросов пользователя

© А.С. Тощев
Казанский (Приволжский) федеральный университет,
Казань
atoschev@kpfu.tu

**Аннотация**

Описан механизм машинного понимания для обработки и решения проблем на естественном языке, поставленных и сформулированных пользователями. Обоснован теоретический подход, базирующейся на теории мышления Мински. Предложены архитектура и программная реализация системы, использующей выработанный алгоритм.

## 1 Введение

В настоящее время в области IT набрало большую популярность системы удаленной поддержки информационной инфраструктуры, так называемой «Аутсорсинг». Ввиду развития рынка компаниям становится невыгодно держать свой штат службы поддержки, и они отдают свою инфраструктуру сторонней компании.

После анализа статистической информации обработки инцидентов было выяснено[1], что большинство проблем, которые решает удаленная служба поддержки, носят весьма тривиальный характер:

- Установить приложение
- Переустановить приложение
- Решить проблему с доступом к тому или иному ресурсу

Данные проблемы решают специалисты технической поддержки. Обычно техническая поддержка делится на несколько линий:

1. Первая линия. Решение уже известных, задокументированных проблем, работа напрямую с пользователем

2. Вторая линия. Решение ранее неизвестных проблем

3. Третья линия. Решение сложных и нетривиальных проблем

4. Четвертая линия. Решение архитектурных проблем инфраструктуры

Каждая линия поддержки представлена специалистами. В среднем команда, обслуживающая одного заказчика насчитывает 60 человек.

### 1.1 Проблема

Основной тенденцией в развитии области удаленной поддержки инфраструктуры является попытки удешевить и улучшить стоимость предоставления услуг.

Компании, работающие на этом рынке, вкладывают большие деньги в автоматизацию. Кроме того современное развитие науки и техники, а точнее вычислительных мощностей позволяет автоматизацию даже самых ресурсоемких процессов.

Дальнейшим развитием области является замена человеческих специалистов на автоматические системы. Многие ведущие компании ведут разработки в этом направлении. Например, компания HP. Данная компания имеет свои системы по регистрации подобных инцидентов и сейчас ведется работа над автоматизацией системы.

Кроме компании HP подобную систему разрабатывает Wolfram Alpha [2], данная система может понимать и отвечать на вопросы пользователя. Например, если спросить ее «2 + 2», то она ответит «4». Это лишь один тривиальный пример.

## 2 Постановка задачи

Задачами данного исследования являются разработать архитектуру системы, практически реализующую модель мышления для обработки и решения запросов на естественном языке, созданных пользователями в системах типа Служба технической поддержки;

- разработать модели и методы обучения системы;
- протестировать эффективность работы системы в сравнении со специалистами-людьми;
- разработать адаптивную архитектуру, демонстрирующую способность системы адекватно

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13-16 октября 2014 г.**

реагировать на свое состояние, например, определять степень нагрузки и распределять ресурсы.

• Подсчет статистических результатов работы комплекса

## 3 Модель мышления

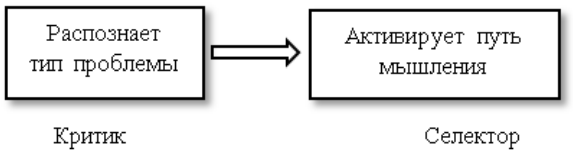
Для решения проблемы автоматической обработки инцидентов (проблем), возникающих в области поддержки информационной инфраструктуры было решено отталкиваться от человеческого понимания.

Человеческое понимание тесно связано с мыслительной деятельностью и является одной из его функций [3]. Существует множество моделей мышления, например, модель Рассела и Норвига [4], модель Мински шести уровней мышления [5] (с. 381–432). Нами была выбрана последняя, так как она лучше подходит для реализации как компьютерной системы. Модель Мински состоит из шести уровней мышления и триплета Критик – Селектор – Образ мышления. Каждый последующий уровень инкапсулирует предыдущий.


Примерами человеческого поведения в рамках модели Мински являются следующие.

Уровень инстинктивных реакций: человек услышал звук и повернул голову. Уровень обученных реакций: человек увидел быстро приближающийся автомобиль, он запомнил эту ситуацию и теперь знает, что нужно отойти в сторону. Уровень рассуждений: чтобы понять, что нужно предложить покупателю на встрече, продавец рассмотрела несколько альтернатив и выбрала лучшую. Уровень рефлексии: человек размышляет над тем, что он недавно сделал для того, чтобы стать более высококвалифицированным профессионалом. Уровень саморефлексии: нежелание опаздывать заставляет человека заранее продумывать его планы. Уровень самосознательной рефлексии: человек продумывает, что он сделает, опираясь на сравнение со своими идеалами. Каждый последующий уровень воспринимает сигналы предыдущего и контролирует его.

Другой важной составляющей модели Мински является триплет Критик – Селектор – Образ мышления.




Каждый из Критиков распознает разные типы проблем. Когда Критик фиксирует достаточное количество внешних воздействий, он активирует Образ мышления, который будет полезен и наиболее адекватен в данной ситуации. Селекторы отвечают за выделение ресурсов памяти. С точки зрения программного комплекса Селекторы отвечают за выбор данных.

Модель Мински описывает человеческое мышление, нами эта модель была дополнена и адаптирована для задачи обработки и решения запросов на естественном языке, созданных пользователями в системах типа Служба технической поддержки.

### 3.1 Реализация Модели мышления

На базе доработанной модели были создана архитектура приложения с расширением исходной модели и реализована система, работающая по данной архитектуре.

Шесть уровней мышления были реализованы отдельным компонентом «Цикл мышления», который запускает и контролирует все действия системы (Критики, Образы мышления), а также общий контекст системы и контекст текущих задач, инкапсулируя необходимую информацию. В функции «Цикла мышления» входит определение целей работы системы.

Критики были реализованы нами как программные функции (вероятностные предикаты), которые в качестве одного из параметров возвращают вероятность, с которой данная функция может обработать входящие данные, тем самым среди всех Критиков выбирается и используется наиболее вероятный. После выбора, активации и работы Критика он в качестве результата формирует объект Селектор. Селекторы возвращают данные из текущего контекста запроса. Образ мышления реализован как компонент, который может модифицировать текущий контекст, изменяя данные в нем.

При реализации уровней мышления нами была дана новая интерпретация значения уровней, предложенных Мински с точки зрения поставленной задачи обработки и решения запросов на естественном языке, созданных пользователями в системах типа Служба технической поддержки.

На уровне инстинктивных реакций система совершает базовую обработку «инстинктивно», используя встроенные шаблоны, но не логические рассуждения.

На уровень обученных реакций система переходит, если решение на первом уровне найти не удалось. На этом уровне активируется Критик

классификации проблем, который обрабатывает входящий запрос, строя семантическую сеть.

Третий уровень включает все логические (вероятностные) рассуждения системы.

Четвертый уровень – уровень рассуждений – производит постановку целей для системы и контролирует два предыдущих уровня. Механизм целей имеет иерархическую структуру, во главе которой стоит базовая цель «Помочь пользователю». Подцелями базовой цели могут быть «Понять запрос», «Понять проблемы», «Найти решение». Также четвертый уровень контролирует время выполнения входящего запроса и, если оно время превышает определённый предел, производит перераспределение ресурсов.

На пятом уровне происходят инициализация контекста запросов и коммуникации с пользователем.

Шестой уровень контролирует общее состояние системы, ресурсов, проблемы функционирования аппаратного комплекса и выставляет общий статус системы. Если все запросы укладываются в отведенное время, то выставляется положительный статус, иначе выставляется отрицательный статус. По общему статусу можно определить, необходимо ли внешнее вмешательство в работу системы: замена компонентов, увеличение ресурсов.

Обмен информацией между уровнями идет посредством разработанной нами концепции контекстов. В системе предусматривается два класса контекстов: краткосрочный и долгосрочный. Краткосрочные контексты существуют во время выполнения запросов и не пересекаются друг с другом. Долгосрочный контекст существует на более высоких уровнях и объединяет знания системы.

Как сказано выше, на втором уровне запрос преобразуется в семантическую сеть из концепций. Важно отметить, что в системе только две предустановленных концепции – это объект и действие. Всем остальным концепциям система обучается посредством взаимодействия с тренером. Обучение также проходит через все 6 уровней модели, после чего новая концепция записывается в базу знаний.

С точки зрения технических особенностей нужно отметить, что для хранения данных выбрана нереляционная база данных, так как она оптимизирована для представления семантических сетей и объектов.

## 4 Сравнение с подобными системами

Для сравнения использовалась система Wolfram Alpha [2]. Она имеет более общий характер и в отличие от созданной нами системы сможет ответить только на общие, но не специфичные вопросы. Кроме того, наша система при построении семантической сети запроса использует Wolfram Alpha для поиска синонимов, чтобы найти концепции из базы знаний. Например, программное обеспечение, софт, программа ссылаются на одну концепцию.

Для сравнения со специфическими системами использовалась HP Open View [6]. Названная система включает комплекс программ для обработки входящих запросов на естественном языке, но не умеет понимать запрос, а направлена на регистрацию запросов при помощи человека-специалиста. Кроме того, решение проблемы, сгенерированной пользователем, также выполняется человеком. В качестве автоматизации система предлагает блок «Самообслуживание», когда пользователь может выбрать из списка необходимое ему действие (возможный запрос), а система автоматически выполнит его на компьютере пользователя (в нашей системе это отнесено к первому уровню – уровню инстинктивных реакций).

## 5 Результаты

Для тестирования системы была составлена выборка типичных запросов из системы обработки заявок. По результатам тестирования удалось добиться 61% успешности обработок заявок.

## Литература


[1] Результаты анализа инцидентов ОАО «ICL КПО-ВС» http://tu-project.com/for-business/

[2] Вольфрам Альфа. https://www.wolframalpha.com/

[3] Соотношение мышления и понимания. – URL: http://litpsy.ru/obshhaya-psixologiya/psixologiya-poznaniya/sootnoshenie-myshleniya-i-ponimaniya/

[4] Рассел С., Норвиг П. Искусственный интеллект. Современный подход. – Вильямс, 2007. – 1408 с.

[5] Мински М. Машина эмоций. – Саймон & Шустер Пейпербэкс, 2007. – 400 с.

[6] Пекар М. Фогнет: руководство по HP OpenView – Фогбукс, 2008. – 251 с.


## Thinking Model and Machine Understanding in Automated User Request Processing


Alexander S. Toschev



A mechanism of machine understanding in processing and resolving of problems generated and formulated by users in natural language is considered. The theory described is based on the Minsky thinking model. An architecture and software implementation of the computer system based on the described algorithm are presented.

# Метод построения схем реляционных баз данных, использующий семантическую информацию

© И.П. Убалехт
ОмГТУ,
Омск
ivan@ubaleht.com

**Аннотация**

В данной работе предлагается метод построения неизбыточных и непротиворечивых схем реляционных баз данных, использующий семантическую информацию. Предлагаемый метод сочетает в себе элементы проектирования схем баз данных методом синтеза (проектирование от атрибутов и функциональных зависимостей к отношениям) и элементы проектирования от сущностей предметной области. В статье предлагается на инфологическом уровне использовать семантически расширенные аналоги функциональных зависимостей, предлагается способ формализации процесса получения функциональных зависимостей, предлагается алгоритм автоматизированного построения схем реляционных баз данных и прототип графической нотации.

## 1 Введение

В настоящее время существует достаточно много работ, посвящённых исследованию принципов и методов формирования схем реляционных баз данных (БД). Несмотря на это процесс формирования схем реляционных БД остаётся недостаточно формализованным. Формализация процесса формирования схем реляционных баз данных остаётся важной задачей теории реляционных баз данных. Помимо этого актуальной является задача разработки методов формирования схем реляционных БД, имеющих такие характеристики как: высокая автоматизация процесса построения схем; развитые средства, обеспечивающие пользователю (проектировщику схем БД) наглядность и удобство управления процессом построения схем БД.

В данной работе предлагается метод построения неизбыточных и непротиворечивых схем БД. Цель метода – получение схемы реляционной БД, находящейся, как минимум в третьей нормальной форме (3НФ).

Ключевые свойства метода, предлагаемого в данной работе:

– высокая формализация процесса получения схем БД;

– наличие выразительных средств, обеспечивающих наглядность и удобство построения схем БД для конечного пользователя;

– высокий уровень автоматизации построения схем БД.

## 2 Обзор работ и публикаций

Кратко рассмотрим существующие в настоящее время подходы к проектированию схем реляционных БД. Рассмотрим новые публикации в этой области, модели, близкие к модели, на которой основан предложенный в статье метод.

Можно выделить четыре основных подхода к проектированию схем реляционных БД:

1. Построение схем через декомпозицию отношений. Данный подход известен со времён создания реляционной модели данных и описан во многих источниках. Сущность подхода заключается в последовательной декомпозиции (нормализации) первоначально заданных отношений через применение к ним ряда правил [13, 16, 17, 23]. Метод применяется на даталогическом уровне.

2. Построение схем реляционных баз данных посредством синтеза схемы из множества функциональных зависимостей (ФЗ). В дальнейшем будем называть этот подход методом синтеза. Данный подход также давно известен [2, 16, 17, 23] и также используется на даталогическом уровне – на уровне реляционной базы данных (рис. 1, б).

3. Подходы, ведущие проектирование от семантики предметной области (ПрО). Данная группа подходов включает в себя большое количество различных моделей [7,8,13].

Общее для всех подходов и моделей данной группы то, что они используются на инфологическом уровне и с их помощью получается концептуальная схема ПрО. На инфологическом

**Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.**

уровне может быть формализована только часть элементов и взаимосвязей ПрО, например, при использовании классической ER-модели (рис. 1, а), но может быть формализовано и большинство элементов и взаимосвязей, например, при использовании UML модели с языком OCL, ORM модели [7, 8] или расширенной ER модели, предложенной в статьях [3, 10] (рис. 1, в). При использовании семантического метода, предложенного в настоящей работе, также большинство элементов и взаимосвязей ПрО формируются на инфологическом уровне (рис. 1, в).

4. Уточнение и нормализация схемы на основе информации, получаемой из уже готовой и заполненной (либо заполненной частично) базы данных. Будем называть этот подход построением схемы базы данных обратным методом. Суть подхода в том, что в данных уже заполненной базы данных либо в промежуточном прототипе базы данных ищутся закономерности, в том числе и те, которые можно расценивать как ФЗ. На основе этих найденных закономерностей оптимизируется схема базы данных. Этот процесс может быть итерационным и совмещённым с подходом № 2 (методом синтеза), как описано в статье [21].

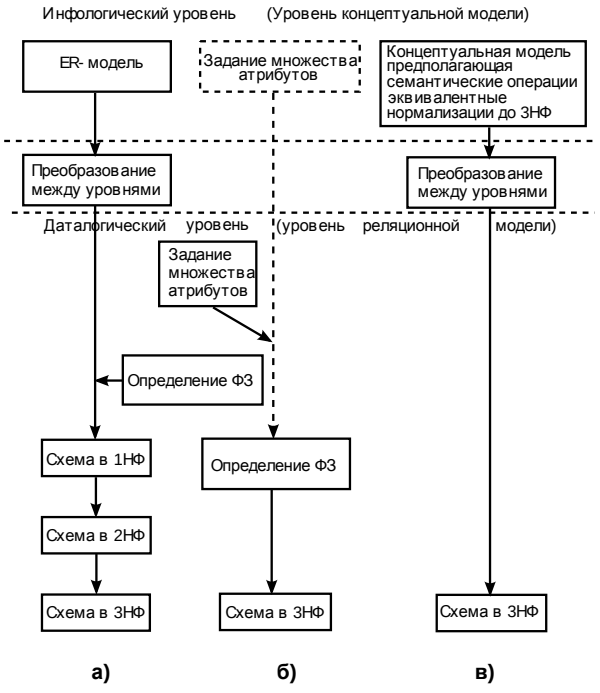



**Рис. 1.** Процесс получения схем реляционных баз данных: а) при использовании ER-модели, б) при использовании метода синтеза, в) при использовании моделей, позволяющих определять большинство элементов и взаимосвязей на инфологическом уровне

Метод, предлагаемый в данной работе, можно в основном отнести к подходу № 3, но он содержит элементы и других описанных выше подходов.

В настоящее время активно создаются новые методы, модели и подходы позволяющие получать неизбыточные и непротиворечивые схемы реляционных БД. Большинство из них можно отнести к подходам, ведущим проектирование от семантики ПрО в представленной выше классификации. Кратко рассмотрим некоторые современные публикации в этой области, имеющие параллели с методом, предлагаемым в данной работе.

Модель «Объект-Роль» (Object-Role Model) предоставляет очень развитые выразительные средства для проектирования БД на инфологическом уровне и для вербализации данных. Данная модель существует довольно давно, однако постоянно выходят новые публикации, посвящённые исследованию этой модели [7, 8].

Семантическая модель «Сущность – Связь – Отображение» является развитием идей ER-модели. Так же, как предыдущая модель, имеет хорошие выразительные средства для детального описания на инфологическом уровне закономерностей исследуемых ПрО [12]. В предлагаемом в данной работе методе так же, как и в двух вышеупомянутых моделях, большое значение играет инфологический уровень, семантическая информация и высокий уровень формализации взаимосвязей между элементами концептуальной схемы.

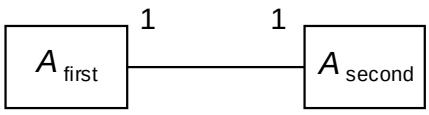
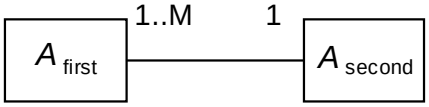
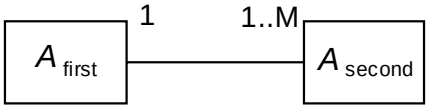
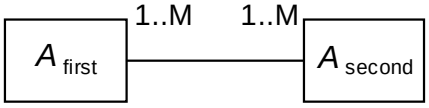
В статьях [3, 10] предлагается современный вариант расширения ER-диаграмм – Articulated Entity Relationship (AER) diagram, который помогает в автоматизированной нормализации схем БД. В методе AER-диаграмм и в методе, предложенном в данной работе есть параллели, это формализация ФЗ на инфологическом уровне и расширение ER-нотации.

В ряде статей [18–20] Панченко Б.Е. вводится понятие реляционного каркаса, с помощью которого можно автоматизировано синтезировать высоко-нормализованные и безаномальные схемы БД. Важную роль в этой модели играют многозначные зависимости.

В работе [22] рассматривается собственная специальная реляционная модель, использующая «отношения-сущности» и «отношения-связывания» для получения схем БД в доменно-ключевой нормальной форме.

Статьи [14, 15] посвящены формализации процесса нормализации схем реляционных БД. Для решения задачи получения оптимальных схем БД в этих статьях предлагается адаптировать известные алгоритмы, например, алгоритм Балаша. В настоящей работе для получения оптимальных схем БД также предлагается алгоритм, который в будущем предлагается описать, например, как алгоритм информированного поиска путей на имплицитном графе.

**Таблица.** Множество состояний бинарной связи *RS*

| Тип связи | Количественное отношение | Описание |
|---|---|---|
| Тип 1 |  | Один-к-одному. Для каждого значения из $A_{first}$ имеется строго одно значение из $A_{second}$, для каждого значения из $A_{second}$ имеется строго одно значение из $A_{first}$. |
| Тип 2 |  | Многие-к-одному. Для каждого значения из $A_{first}$ имеется строго одно значение из $A_{second}$, для каждого значения из $A_{second}$ имеется не менее одного значения из $A_{first}$. |
| Тип 3 |  | Однин-ко-многим. Для каждого значения из $A_{first}$ имеется не менее одного значения из $A_{second}$, для каждого $A_{second}$ имеется строго одно значение из $A_{first}$. |
| Тип 4 |  | Многие-ко-многим. Для каждого значения из $A_{first}$ имеется не менее одного значения из $A_{second}$, для каждого значения из $A_{second}$ имеется не менее одного значения из $A_{first}$. |

## 3 Модель построения схем баз данных с учётом семантической информации

Основные характеристики предлагаемого метода:

– сочетание элементов проектирования схем БД через синтез отношений (проектирование от атрибутов и ФЗ к отношениям) и элементов проектирования через декомпозицию (проектирование от сущностей предметной области);

– использование двухстрочных отношений для определения семантики связей между атрибутами (выявление элементов связанности). При переходе на даталогический уровень элементы связанности преобразуются в ФЗ. Таким образом, на инфологическом уровне предоставляется механизм для формализации ФЗ;

– алгоритм получения безаномальных схем БД, используемый в данном методе можно охарактеризовать как алгоритм локальной декомпозиции с глобальным синтезом схем БД;

– графическая нотация с помощью которой удобно выявлять семантические аналоги ФЗ на инфологическом уровне. Нотация ориентированна на встраивание элементов данного подхода в наиболее распространённые на практике ER-диаграммы в нотации Баркера.

Формализуем основные понятия модели.

**Определение 1.** *Элементом области связанности* или просто *элементом связанности* будем называть тройку $EAI< A_{first}, A_{second}, RS>$, где $A_{first}$, $A_{second} \in \boldsymbol{A}$, $\boldsymbol{A} \subseteq \boldsymbol{U}$; $RS \in \boldsymbol{RS}$. $\boldsymbol{A}$ – множество атрибутов входящих в область связанности; $\boldsymbol{U}$ – множество всех атрибутов заданной ПрО; $\boldsymbol{RS}$ – множество состояний бинарной связи между атрибутами $A_i, A_j \in \boldsymbol{A}$, где $i, j$ любые целые числа от 1 до $n$, $n$ – мощность множества $\boldsymbol{A}$.

В таблице представлено $\boldsymbol{RS}$ – множество всех вариантов состояний бинарной связи *RS*. Связь *RS* присутствует между каждыми двумя атрибутами в области связанности. Графическое отображение связей *RS* между атрибутами представлено рис. 2 б. В данной работе не рассматриваются варианты связи *RS* учитывающие возможность неопределённых значений (null) атрибутов.

**Определение 2.** Пусть $\boldsymbol{A}$ – *произвольное* множество атрибутов $\in \boldsymbol{U}$; $\boldsymbol{U}$ – множество всех атрибутов заданной ПрО; $\boldsymbol{A} \subseteq \boldsymbol{U}$; $A_{first}, A_{second} \in \boldsymbol{A}$. *Областью связанности AI* будем называть такое

множество троек $EAI< A_{first}, A_{second}, RS >$ (множество *элементов связанности)*, что: $A_{first}$, $A_{second}$ каждой тройки *EAI* является элементом декартова произведения $\boldsymbol{A} \times \boldsymbol{A}$. Количество троек $EAI \in AI$ равно количеству элементов множества $\boldsymbol{A} \times \boldsymbol{A}$.

Рассмотрим по шагам схему предлагаемого метода получения неизбыточных и непротиворечивых схем БД. На шаге 1 производится определение первоначальной схемы областей связанности $\boldsymbol{AI}(S_{start})$. На шаге 2 производится установка связей между атрибутами внутри элементов связанности (выявление семантических аналогов ФЗ). На шаге 3 – запуск алгоритма распределения элементов связанности по областям связанности (см. алгоритм). Шаг 2 и 3 выполняются итеративно до тех пор, пока не сработают условия, определяющие, что исходная схема эквивалентна состоянию как минимум ЗНФ. На шаге 4 осуществляется перевод конечной нормализованной схемы областей связанности $\boldsymbol{AI}(S_{finish})$ с инфологического уровня в схему реляционных отношений на даталогическом уровне. На даталогическом уровне схема отношений будет как минимум в ЗНФ.

Шаг 1. На данном шаге имеется множество областей связанности, построенных исходя из неформального понимания проектировщиком предметной области. Основой некоторой формализации на данном этапе может служить проектная документация. Проектировщик определяет области связанности – это наборы атрибутов, между которыми предполагается существование некоторых связей.

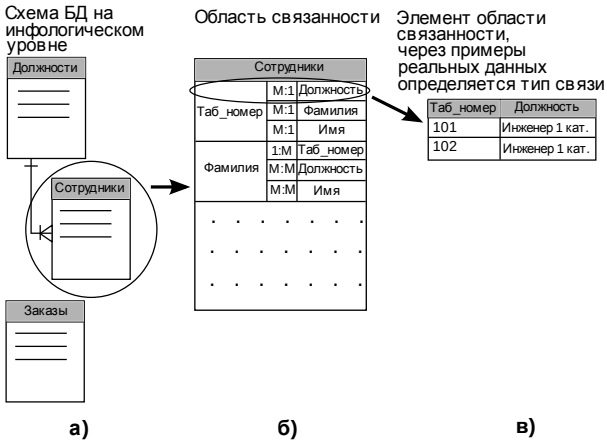



**Рис. 2.** Определение связи между атрибутами *Таб_Номер* и *Должность* через использование двухстрочного отношения: а) схема БД на инфологическом уровне; б) детализирована область связанности *Сотрудники*; в) детализирован элемент связанности *Таб_Номер – Должность* области связанности *Сотрудники*, связь между атрибутами которого определяется через использование двухстрочного отношения, заполненного реальными данными

Шаг 2. В определённых на шаге 1 областях связанности в соответствии с таблицей устанавливаем типы связей между атрибутами внутри всех элементов связанности, то есть определяем все элементы связанности.

Для определения типов связей между атрибутами внутри элементов связанности используется формальная модель. Эта модель включает: интерактивный графический язык, близкий по возможностям языку Query-By-Example (QBE), логику двухстрочных отношений [17] (рис. 2) и реляционное исчисление.

Рассмотрим данную модель подробнее, модель должна осуществлять следующее: на инфологическом уровне используя семантическую информацию позволять пользователю задавать элементы связанности внутри областей связанности; поддерживать удобную графическую нотацию; строго формализовать переход от элементов связанности на инфологическом уровне к ФЗ на даталогическом уровне.

Покажем корректность перехода от элементов связанности на инфологическом уровне к ФЗ на даталогическом уровне. Каждый элемент связанности можно представить как двухстрочное отношение. Опираясь на [17] определим понятие двухстрочного логического отношения.

**Определение 3.** Пусть *r* – отношение со схемой *R*, *A* – атрибут в *R*, отношение *r* будем называть двухстрочным логическим отношением, если оно содержит в точности два кортежа $t_1$ и $t_2$ и если с отношением *r* ассоциировано $\Psi r$ – присваивание истинностных значений атрибутам из *r*. $\Psi r$ – является функцией из *R* в {истина, ложь}, такой что

$$\Psi r(A) = \begin{cases} \text{истина, если } t_1(A) = t_2(A), \\ \text{ложь, если } t_1(A) \neq t_2(A). \end{cases}$$

Таким образом, элемент связанности можно представить как двухстрочное логическое отношение. В качестве значений в строках этого отношения могут быть данные, соответствующие семантике ПрО. Далее нужно формализовать переход от двухстрочного логического отношения к ФЗ в произвольных конечных отношениях.

В соответствии с [17] приведём теорему об эквивалентности ФЗ в отношениях с произвольным (конечным) множеством кортежей, ФЗ в двухстрочных логических отношениях и импликацией, доказательство см. в [17].

**Теорема.** Пусть *F* – множество ФЗ над схемой *R* и $A \rightarrow B$ есть зависимость над *R*. Тогда следующие утверждения эквивалентны: из *F* следует $A \rightarrow B$ для произвольных (конечных) отношений; из *F* следует $A \rightarrow B$ для двухстрочных логических отношений; из *F* следует $A \rightarrow B$ как логическая формула (импликация).

Рассмотрим на примере описанные выше преобразования этого шага. На рис. 2 представлен пример предметной области. Заданы области

связанности, которые можно соотнести с сущностями: *Сотрудники*, *Должности*, *Заказы* (рис. 2, а). Пользователь просматривает область связанности – *Сотрудники* (на рис. 2, б выделена область связанности *Сотрудники*, показан момент анализа пользователем этой области связанности). Пользователю достаточно задать только часть связей между атрибутами в области связанности, остальная часть связей может быть автоматически достроена через применение правил вывода. Далее на рис. 2 в показан момент анализа пользователем отношения (связи) между атрибутами *Таб_Номер* и *Должность*. При анализе пользователь заполняет две строки примерами реальных данных, соответствующих семантике и ограничениям атрибутов *Таб_Номер* и *Должность* в данной ПрО. Из рис. 2 видно, что в соответствии с введёнными реальными данными между атрибутами *Таб_Номер* и *Должность* у пользователя получилась связь типа 2 – многие-к-одному (см. таблицу), что может быть представлено следующим логическим выражением (термом):

$\{t^{(2)} | (\exists u)(Сотрудники(t) \wedge Сотрудники(u) \wedge (t[Таб\_номер] \neq u[Таб\_номер] \vee t[Должность] = u[Должность]))\}$

Получившаяся связь многие-к-одному и представленное логическое выражение эквивалентны тому, что в соответствии с приведенной выше теоремой между атрибутами *Таб_Номер* и *Должность* в произвольном конечном отношении на даталогическом уровне будет ФЗ.

Как видно из примера выше, связи, семантически выраженные в двухсторочных отношениях, можно рассматривать как логические выражения, которые в соответствии с приведённой теоремой эквивалентны ФЗ в произвольных конечных отношениях. Таким образом, области связанности с установленными связями моделируют будущие отношения с ФЗ, которые появятся при переходе на даталогический уровень.

Шаг 3. На предыдущих шагах задана начальная конфигурация областей связанности. Внутри всех областей связанности и для всех элементов связанности определены типы связей. На данном шаге элементы связанности должны стать условиями формирования (преобразования) областей связанности. Под влиянием элементов связанности, заданных на предыдущем шаге, с областями связанности могут автоматически производиться следующие действия: декомпозиция областей связанности; отсоединение не связанных с данной областью связанности атрибутов; поиск недостающих атрибутов и присоединение их к областям связанности.

Эти действия можно алгоритмизировать, алгоритм по преобразованию областей связанности назовём SemanticNormalization.

Применение алгоритма SemanticNormalization можно отразить следующим выражением:

$(\boldsymbol{AI}(S_{start}), \text{Rules}, \text{Heuristic}()) \Rightarrow \boldsymbol{AI}(S_{finish})$,

где $\boldsymbol{AI}(S_{start})$ – множество областей связанности заданных на шаге 1, образующих схему $S_{start}$;

Rules – множество правил, которые определяют какие из операций (рис. 3) нужно применить к текущей области связанности *AI*, состав правил в Rules в данной статье не уточняется, правила в Rules могут образовывать формальную логическую систему;

Heuristic( ) – эвристическая функция;

$\boldsymbol{AI}(S_{finish})$ – множество областей связанности, образующее схему $S_{finish}$ после работы алгоритма SemanticNormalization.

Цель алгоритма – за конечное число операций по преобразованию областей связности перейти от начальной ненормализованной схемы $\boldsymbol{AI}(S_{start})$ к конечной схеме областей связанности $\boldsymbol{AI}(S_{finish})$ такой, что при переводе её на даталогический уровень она образует схему БД как минимум в ЗНФ.

Ниже приведен алгоритм SemanticNormalization. Вход: AI(S ) и Rules SemanticNormalization($\boldsymbol{AI}(S_{start})$, Rules)

```
begin
  while (BufferOfAttributes is Change) or
(BufferOfAttributes != Empty)  or (∃ AI is Change)
  begin
    //перебор всех областей связанности
    foreach(AI in AI)
    begin
      //применение операций и правил к областям
      //связанности и получение  множества
      //областей связанности с новой схемой
      //AI(S_new)
      AI(S_new) =  AppliedOperationsToCurrentAI(
      Operations ={ DeleteAttributeOperation(),
      DecompositionOperation() }, Rules);
    end foreach
    //проверка связанности атрибутов из буфера не
   //присоединённых атрибутов BufferOfAttributes
   //с областями связанности из AI(S_new)
    if(Heuristic is possible)
      AI(S_newest) = Heuristic(AI(S_new), Buffer);
    else
      //если эвристика не возможна, то полный
       //перебор
      foreach(Element  in BufferOfAttributes)
      begin
        foreach(AI in AI)
          AddAttributeOperation(AI, Element);
        end foreach
      end foreach
    end if
  end while
end
```

Алгоритм SemanticNormalization работает, используя множество строгих правил вывода Rules и эвристику. Множество правил Rules определяет, какую из перечисленных ниже операций нужно применить к текущей области связанности *AI*. Далее опишем операции, которые можно производить над областями связанности (рис. 3): операция отсоединения лишнего атрибута от области связанности – DeleteAttributeOperation (рис. 3, а); операция добавления атрибута из буфера несвязанных атрибутов BufferOfAttributes к текущей области связанности – AddAttributeOperation (рис. 3, б); операция разделения области связанности – DecompositionOperation (рис. 3, в).

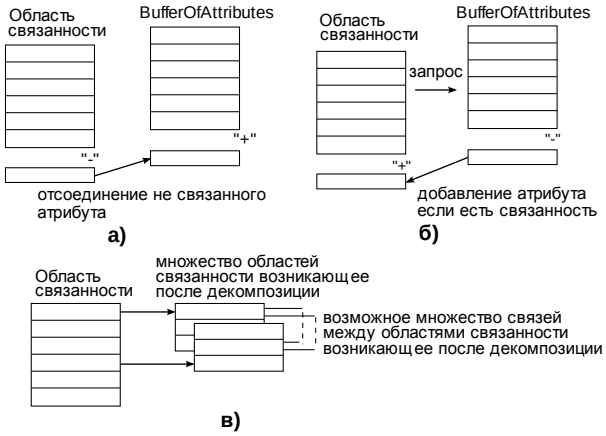



**Рис. 3** Операции, производимые над областями связанности: а) DeleteAttributeOperation, б) AddAttributeOperation, в) DecompositionOperation

Алгоритм прекращает свою работу тогда когда:

– в буфере BufferOfAttributes нет элементов;

– после очередной итерации к буферу атрибутов BufferOfAttributes не добавлено ни одного элемента;

– применяя к каждой области связанности множество правил Rules невозможно произвести ни одной операции из операций, представленных на рис. 3.

После завершения работы алгоритма SemanticNormalization формируется множество областей связанности со схемой S . При переводе множества областей связанности образующих конечную схему S , во множество реляционных отношений на даталогическом уровне, это множество отношений образует схему БД находящуюся как минимум в 3НФ.

Шаг 4. Множество областей связанности со схемой $S_{finish}$ с инфологического уровня переводится во множество отношений на даталогическом уровне. Этот процесс хорошо исследован см., например, источники [13, 16] и подробно в этой работе не описывается.

## 4 Заключение

В статье рассмотрен метод построения неизбыточных и непротиворечивых схем реляционных БД, использующий семантическую информацию. Новизна представленного метода заключается в следующем:

– метод сочетает элементы проектирования схем БД через синтез отношений (проектирование от атрибутов и ФЗ к отношениям), это придаёт методу строгость и элементы проектирования через декомпозицию (проектирование от сущностей предметной области), это предаёт методу наглядность;

– представленный метод можно отчасти рассматривать как семантическое расширение метода синтеза. В проанализированных современных источниках [1, 4–6, 9, 11] при практическом использовании метода синтеза совершенно не формализовывался процесс получения исходного множества ФЗ. Представленный метод строго формализует процесс получения исходного множества ФЗ через применение двухстрочных отношений с примерами реальных данных и логических формул на инфологическом уровне. Такой способ формализации определения ФЗ представляется вполне хорошо реализуемым в программном продукте;

– представлены элементы графической нотации, которые могут использоваться в диаграммах на инфологическом уровне для работы пользователя с семантической информацией. Представленные элементы графической нотации хорошо сочетаются с ER-диаграммами в нотации Баркера (данная нотация ER-диаграмм является в настоящее время наиболее используемой) и могут хорошо дополнять ER-диаграммы при практической реализации данного метода в программном продукте. Предлагаемая нотация может успешно конкурировать с другими современными вариантами расширения ER-диаграмм, например, с нотацией представленной в [3, 10], которая является более громоздкой;

– разработан алгоритм с локальной декомпозицией и глобальным синтезом областей связанности, использующий семантическую информацию и работающий на инфологическом уровне, нормализующий схему будущей БД до состояния эквивалентного 3НФ.

## Литература


[1] Bahmani A., Naghibzadeh M., Bahmani B. Automatic database normalization and primary key generation // 21th Canadian Conference on Electrical and Computer Engineering. – Niagara Falls, Canada, 2008.

[2] Bernstein P.A. Synthesizing Third Normal Form Relations from Functional Dependencies // ACM Transactions on Database Systems (TODS), 1976. Vol. 1, Iss. 4. P. 277–298.

[3] Dhabe P.S., Patwardhan M.S., Deshpande A.A., Dhore M.L., Barbadekar B.V., Abhyankar H.K. Articulated entity relationship (AER) diagram for

complete automation of relational database normalization // International Journal of Database Management Systems (IJDMS), 2010. Vol. 2, No. 2. P. 84–100.

[4] Dongare Y.V., Dhabe P.S., Deshmukh S.V. RDBNorma: – A semi-automated tool for relational database schema normalization up to third normal form // International Journal of Database Management Systems ( IJDMS ), 2011. Vol. 3, No. 1. P. 133–154.

[5] Du H., Wery L. Micro: A normalization tool for relational database designers // Journal of Network and Computer Application, 1999. Vol. 22, No. 4. P. 215–232.

[6] Georgiev N. A web based environment for learning normalization of relational database schemata. Masters thesis. – Umea, Umea university, 2008.

[7] Halpin T. Conceptual Schema and Relation Database Design. – 2th ed. – Sydney: Prentice-Hall of Australia Pty., Ltd, 1995.

[8] Halpin T., Morgan T. Information Modeling and Relational Databases. – 2th ed. Kaufmann Publishers, 2008. 943 p.

[9] Kung H., Tung H. A web-based tool to enhance teaching/learning database normalization. 9th Annual Conference of the Southern Association for Information Systems (SAIS). – Jacksonville, USA, 2006.

[10] Patwardhan M.S., Dhabe P.S., Deshpande A.A., Londhe S.G., Dhore M.L., Abhyankar H.K. Diagrammatic approach for complete automation of relational database normalization at conceptual level // International Journal of Database Management Systems (IJDMS), 2010. Vol. 2, No. 4. P. 132–151.

[11] Yazici A., Ziya K. (2007), JMathNorm: A database normalization tool using mathematica // International Conference on Computational Science 2007 (ICCS 2007). – Beijing, China, 2007.

[12] Бабанов А.М. Семантическая модель «Сущность – Связь – Отображение» // Вестник Томского государственного университета. Управление, вычислительная техника и информатика, 2007. №1. С. 77–91.

[13] Дейт К. Дж. Введение в системы баз данных : пер. с англ. – 7-е изд. – М.: Издат. дом «Вильямс», 2001. 1072 с.

[14] Клименко И.В. Метод построения семейства максимальных транзитивно независимых множеств атрибутов // Вопросы современной науки и практики. Университет им. В.И. Вернадского, 2011. Т. 35, вып. 4. С. 70–78.

[15] Клименко И.В. Модификация алгоритма Балаша для решения задачи нормализации реляционных баз данных // Вопросы современной науки и практики. Университет им. В.И.Вернадского, 2012. Т. 37, вып. 1. С. 43–49.

[16] Кузнецов С.Д. Основы баз данных: учебное пособие. – 2-е изд. – М.: Интернет-Университет Информационных Технологий; БИНОМ. Лаборатория знаний, 2007. 484с.

[17] Мейер Д. Теория реляционных баз данных: пер. с англ. – М.: Мир, 1987. 608 с.

[18] Панченко Б.Е. Об алгоритме синтеза реляционного каркаса. Постановка задачи и формализация // Компьютерная математика. 2012. № 1. С. 84–93.

[19] Панченко Б.Е. Каркасное проектирование доменно-ключевой схемы реляционной базы данных // Кибернетика и системный анализ. 2012. № 3. С. 174–187.

[20] Панченко Б.Е. Алгоритм синтеза реляционного каркаса - неформальное описание // Проблемы управления и информатики. 2013. № 1. С. 83–103.

[21] Радченко В.А., Мальков Ю.А., Балюк С.А., Горпиненко Ю.С. Синтез логической схемы реляционной базы данных на основе выявления множества функциональных зависимостей // Системы обработки информации. 2011. Т. 95, вып. 5. С. 218–224.

[22] Тукеев У.А., Алтайбек А.А. Концептуальная, логическая модели и алгоритм проектирования баз данных в доменно-ключевой нормальной форме // Труды 13-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2011). – Воронеж, 2011. С. 119–125.

[23] Ульман Дж. Основы систем баз данных: пер. с англ. – М.: Финансы и статистика, 1983. 334 с.

## Method of Creation of Schemes of Relational Databases Using Semantic Information

Ivan P. Ubaleht

This paper proposes a method for creating non-redundant and consistent schemes of relational databases using semantic information. The proposed method combines elements of the database schema design method of its synthesis (design from attributes and functional dependencies to relations) and design proceeding from the domain entities. This paper proposes to use semantically extended analogs of functional dependencies on the Infological level; a method for definition of functional dependencies; algorithm of automatic creation of schemes of relational databases and a prototype of graphical notation.

# On Several Social Network Analysis Problems

© George Chernishev © Vsevolod Sevostyanov © Kirill Smirnov
Saint-Petersburg University, Russia
chernishev@gmail.com vsevost@gmail.com kirill.k.smirnov@math.spbu.ru

© Ilya Shkuratov
Saint-Petersburg University, Russia
shkuratov.ilya@gmail.com

## Abstract

In this paper we describe our approach to several problems offered at the ACM SIGMOD Programming Contest 2014. These problems belong to the area of a social network analysis and involve several types of queries to a social graph. The considered graph is modeled by the standard SNB benchmark. We briefly introduce this benchmark, the contest and the problems. Next, we describe our contribution, which is the following: the algorithms for evaluation of these queries and their efficient implementation. Furthermore, we present parallelization techniques for these algorithms and describe overall architecture of our solution.

## 1 Introduction and Related Work

In this paper we study several problems offered at the ACM SIGMOD Programming Contest 2014 [1], a yearly programming contest focused on a data management topics.

This contest has a number of features, which distinguish it from a well-known ICPC series:

- Participants are offered some science-intensive task, which is usually an unsolved problem of current importance.
- The contest runs for several months and no on-site participation is required.
- Topic specificity — the clear data management focus is present. For example, contests of previous years involved construction of distributed query processing engine (2010), multidimensional index (2012) or document stream filtering system (2013).
- The participation is allowed to both graduate and undergraduate students, without any restriction on a number of attempts.

While this contest is not so well known as the ICPC, it is nevertheless popular. For example, last year there were more than 100 registered teams. The contest is relatively young — it runs for $6^{th}$ time this year.

In this paper we also describe the contest: the rules, the task, its timeline and required qualifications. Moreover, we present our experiences and provide a solution of the team "GenericPeople" (Ilya Shkuratov and Vsevolod Sevostyanov), which was ranked[1] 17 out of 33 teams on the preliminary (public) tests. While our approach is not the best, it still has merit:

- our solution can serve as an example demonstrating the required qualifications and which may help to assess the required effort and work intensity. These factors may be of interest for a person who is thinking about the participation;
- the solution successfully passed through all available tests (datasets of three different sizes) within the time limits specified by the contest organizers (5 and 10 minutes);
- the proposed algorithms passed all correctness tests;
- parallelization techniques of these algorithms may be of interest;
- the number reported in the leaderboard is the sum over all query types, at the present time we can say nothing regarding their individual performance;
- at last, the number was reported for three datasets; the proposed algorithms may behave differently (better or worse) on another dataset.

Thus, we deem current study as worthy to be presented and of some interest for the reader. Another motivation for this paper is the concise presentation of the

**Proceedings of the 16th All-Russian Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" – RCDL-2014, Dubna, Russia, October 13-16, 2014.**

[1] `http://www.cs.albany.edu/~sigmod14contest/leaders.html`, last accessed 02/05/2014.

solution for the contest problem, which is usually lacking. After the contest all what is left are the posters of the top five performing teams without detailed explanation (it is given orally at the conference). Also, these posters are (or at least were in the past years) not going into the conference proceedings and are kept on a website, which may disappear. Moreover, we present our experiences and describe (at least partially) the way we went through in order to produce a working solution. It is impossible to pass on all these aspects via poster.

This year contest was dedicated to a social network analysis topic. Social network is essentially a graph, whose vertices represent users and edges denote relations between them. An example of such relation may be "know each other", "follow" and so on. Additionally extra information like a place of work or study, geographical information, various tags, images, likes etc. is known.

In the past years massive amounts of such information were made available for analysis, forming a strong incentive for both academy and industry to come with means for its efficient storage and processing. Social data play a significant role in the whole "Big Data" movement.

A lot of analysis tools employ the MapReduce [6] programming model. Industrial examples of such systems are PIG (Yahoo) [13], SCOPE (Microsoft) [4], Hive (Facebook) [19], Dremel (Google) [14]. Academic examples are Starfish [9], HadoopDB [3] and many others[2]. An alternative (which can be considered a poor man's solution) sometimes employed in production environment, is to use scripts written in scripting language like Python to commence the analysis. A data scientist has to analyze the problem and implement all necessary algorithms manually. While it may not favor the rapid development, it may allow to achieve a more efficient processing. Naturally, this approach is more flexible than using a standard tool and allows a fine-tuning of algorithms. However, it requires extensive technical expertise: knowledge of algorithms and data structures, the understanding of the data processing and so on. The tasks of the contest are representative examples of this "manual" approach and can be considered as a training for a data scientist.

Another aspect of the contest task is the graph analysis component. Graph analysis is a mature area of research which studies the efficient storage and processing of graph data. There are several graph database management systems (a special type of DBMS) and graph programming frameworks. These DBMS feature special query languages, query processing algorithms and data storage.

Some examples of the graph DBMS are Neo4j [12], InfiniteGraph [10] and the framework examples are Apache Giraph [2], Signal/Collect [17]. It is necessary to mention that two latter systems also follow the MapReduce model.

The contestants were given the task which consists of the datasets and four types of queries. The social graph was generated using the SNB [16] tool.

[2]A list can be found in `http://dl.acm.org/citation.cfm?id=1454166`, last accessed 22/07/2014.

The goal was to develop a program which computes the results as fast as possible. The contestants had not only to devise the algorithms for efficient query processing on a large graph, but also to parallelize them. This is a must, given the fact that the evaluation of the resulting implementation was performed on a server-class equipment (8 cores).

Another important aspect was the order of computation for each sub-query. The contestants had to bear in mind the size of intermediate results and the memory bound. In other words, the contestants had to perform the work of a query optimizer: gather needed statistics, assess selectivities and develop an optimal processing strategy for each query type. Also, given the hardware multi-core capability, efficient inter-query type orders are also of interest.

The contribution of this paper is the following:

- The description of the ACM SIGMOD Programming Contest 2014 and its task;
- The contest from the participant's point of view: our experiences;
- The algorithms to handle the problems offered at the contest;
- A parallelization techniques for each of these algorithms;
- A general system architecture: subquery computation orders, inter-query type orders and chunk-based data loading.

Now, we are going to describe our experience. The SNB description and its data schema is presented in the appendix section. Detailed description of our approach and data statistics can be found in the report [5].

## 2 Contest description and experiences

Let's describe this year contest from the participants' point of view. We have already briefly described the contest and its specifics in the introduction section. You can find detailed information regarding the ACM SIGMOD Programming Contest series in the reference [18].

Our research group is a frequent participant of this contest; we had achieved good results twice in the past: in the 2010[3] (team "spbu") and 2013[4] (team "Rota Fortunae") year. Both times our teams achieved $3^{rd}$ place in the final ranking.

### 2.1 General information

This year contest followed the general scheme described in the reference [18]. However, there were several notable divergences:

[3]`http://dbweb.enst.fr/events/sigmod10contest/results/#winner`, last accessed 22/07/2014.

[4]`http://sigmod.kaust.edu.sa/finalists.html`, last accessed 22/07/2014.

1. The contest started noticeable later compared to previous years;
2. There were no $2^{nd}$ round, unlike early years. This change happened in 2013;
3. The absence of the dedicated correctness testing phase during the evaluation (it was performed concurrently with the performance evaluation);
4. There was a series of datasets which were progressively disclosed by the organizers, as the performance of the submissions improved;
5. The task did not explicitly required parallelization or concurrency support, but instead, implied it. It was possible to submit purely sequential implementation;
6. It was possible to submit only the executable, without source code during the preliminary evaluation. The final evaluation required source code and this led to some compatibility difficulties;
7. Contestants were allowed to choose programming languages other than C++.

The provided task was a science-oriented problem related to social network analysis. The problem was to execute a number of queries to a graph representing some social network. The goal was to produce a correct answer and minimize the overall processing time. The graph and queries are fully described in the next section.

Below you can see the timeline of the contest.

- January 25, 2014 — Contest announced.
- February 1, 2014 — Detailed specification of the requirements and test data available.
- February 16, 2014 — A medium data set (10k people) with query workload and answers are available on the Task page. New query workload and answers for the small data set (1k people) are available on the Task page.
- March 1, 2014 — Team registration begins. Leaderboard available.
- March 11, 2014 — Workloads on a medium data set (10k people) have been added to the evaluation system.
- March 17, 2014 — Workloads on a large data set (100k people) have been added to the evaluation system.
- April 15, 2014 — Final submission deadline.
- May 15, 2014 — Finalists announcement.
- June 22-27, 2014 — Conference: announcement of the winner and the poster presentations.

In the overall the contest run for two and a half months. Also you can see that several datasets were progressively added to the evaluation pool. These datasets were progressively disclosed by the organizers as the performance of submissions improved. This is a rather new model of evaluation (appeared in 2013 contest) and it was employed in the following way. As soon as the several submissions were achieving some performance level, where it was hard to discern their quality due to inaccurate measurements (thread scheduling effects, for example), a new, larger dataset was added.

### 2.2 Communication with the contest organizers

Information about the order and rules of the contest were provided on a special web page [1], which was the main mean of communication between the organizers and the contestants. It also describes test data sets, the task and an evaluation environment. Later opportunities to register a team and submit solutions were added.

The organizers also created a Google Group in order to discuss any technical issues (e.g. code page problems) and to provide additional information that might be of interest to all of the contestants: test data-sets publication dates, disk space availability, size of data set for the final evaluation and so on.

### 2.3 Required skills and our experiences

Since the organizers of the contest considers Linux as its target platform, we decided to use C++ programming language as it looks to us an highly-optimizable one. Those who want to take part in the contest are advised to learn Linux development utilities such as gcc, make, valgrind (especially callgrind might be useful), gdb, etc. Also two bash scripts were required: one should build the solution and the other — run it with certain parameters.

You also may encounter restriction on size of submitted solution. It was 8 MB this year, thereby it was helpful for us to learn a couple of gcc flags. The first one is *-s*. It removes unneeded symbols from an executable, thus reducing its size without the loss of performance. The second flag may be useful, if you use external libraries: *-MM* instructs the compiler to generate source files dependencies. This helped us to familiarize with boost headers dependencies, strip boost from unneeded header files and further reduce submitted archive size.

Understanding compiler optimization methods may be of use as well. It allowed us to cope with the gcc optimizer bug, namely incorrect copy propagation after global common subexpression elimination pass. It leads to usage of the original pointer to the buffer instead of its copy, which cause segmentation fault on an attempt to free this buffer. The workaround is to add a dummy use of the original pointer after working with the buffer.

Another important skill is an ability to find necessary information on the subjects of the competition, i.e. the ability to work with digital libraries. Usually the task of

the competition (or one of the tasks) is an unsolved scientific problem. Thus one may find useful information about methods have been tried or perspective approaches. These gave us several hints for the given task.

### 2.4 Tools

Aside from the usual requirements this year contest posed an additional one: knowledge of some scripting language or a tool for data analysis. This language can be used for data mining: to detect hidden dependencies in the source data and to collect necessary statistics. We used Python programming language; other examples include R and Octave tools.

### 2.5 Data

The schema for the data used in the task formulation is presented on Figure 3. Data were stored as a set of CSV files. It is worthy to mention that not all of the files were needed for the query processing. Also, organizers had provided data only for two datasets — the one containing thousand and the one containing ten thousand of persons. These datasets are sufficient for the debug purposes, but they are not enough to tune algorithms for the final evaluation, which involved a graph of million of persons. The benchmark generation parameters were kept in secret and it was impossible to generate that graph by ourselves.

## 3 Problems

The contest offered [1] the following problems (we fully provide them here for the better understanding of the reader and in case of the original website outage):

1. **Query Type 1 (Shortest Distance Over Frequent Communication Paths).**

   Given two integer person ids $p1$ and $p2$, and another integer $x$, find the minimum number of hops between $p1$ and $p2$ in the graph induced by persons who:

   (a) have made more than x comments in reply to each other's comments (see comment_hasCreator_person and comment_replyOf_comment);

   (b) know each other (see person_knows_person, which presents undirected friendships between persons; a friendship relationship between persons $x$ and $y$ is represented by pairs $x|y$ and $y|x$).

2. **Query Type 2 (Interests with Large Communities).**

   Given an integer $k$ and a birthday $d$, find the $k$ interest tags with the largest range, where the range of an interest tag is defined as the size of the largest connected component in the graph induced by persons who:

   (a) have that interest (see tag, person_hasInterest_tag);

   (b) were born on $d$ or later;

   (c) know each other (see person_knows_person, which presents undirected friendships between persons; a friendship relationship between persons $x$ and $y$ is represented by pairs $x|y$ and $y|x$).

3. **Query Type 3 (Socialization Suggestion).** Given an integer $k$, an integer maximum hop count $h$, and a string place name $p$, find the top-$k$ similar pairs of persons based on the number of common interest tags (see person_hasInterest_tag). For each of the $k$ pairs mentioned above, the two persons must be located in $p$ (see person_isLocatedIn_place, place, and place_isPartOf_place) or study or work at organizations in $p$ (see person_studyAt_organization, person_workAt_organization, organisation_isLocatedIn_place, place, and place_isPartOf_place). Furthermore, these two persons must be no more than $h$ hops away from each other in the graph induced by persons and person_knows_person.

4. **Query Type 4 (Most Central People).** Given an integer $k$ and a string tag name $t$, find the $k$ persons who have the highest closeness centrality values in the graph induced by persons who:

   (a) are members of forums that have tag name $t$ (see tag, forum_hasTag_tag, and forum_hasMember_person);

   (b) know each other (see person_knows_person, which presents undirected friendships between persons; a friendship relationship between persons $x$ and $y$ is represented by pairs $x|y$ and $y|x$).

   Here, the closeness centrality of a person p is:

   $$\frac{(r(p)-1)\cdot(r(p)-1)}{(n-1)\cdot s(p)},$$

   where $r(p)$ is the number of vertices reachable from $p$ (inclusive), $s(p)$ is the sum of geodesic distances to all other reachable persons from $p$, and $n$ is the number of vertices in the induced graph. When either multiplicand of the divisor is 0, the centrality is 0.

## 4 Algorithms

In this section we describe algorithms for the tasks of the contest. Due to the space constraints they are presented in a brief, a detailed version featuring algorithm listings can be found in the report [5].

In the rest of this paper we refer to the graph induced by "know each other" relation as $graph$, and to the breadth-first search of that graph as BFS. This graph is used in every query type and BFS (as we show further) plays the key role in all of them. Thus, a shorthand notation would be useful.

### 4.1 Query Type 1 (Shortest Distance Over Frequent Communication Paths)

#### 4.1.1 Algorithm description

An obvious strategy for evaluation of such query would be the following:

1. Run BFS from person $p1$ to person $p2$ and return hops count;
2. During the BFS traversal one needs to check the replies condition. For each edge, considered on a given BFS step, one has to calculate the number of mutual replies for the corresponding persons. If it is less than $k$, then the transition is not possible — the edge does not exist.

This "naive" approach needs no preparation and can be ran just after the $graph$ construction. For each pair of adjacent persons it is necessary to calculate the number of replies and this may take some time. Thus, the described BFS has the complexity $O(m \cdot n \cdot (|V| + |E|))$ where $n$ denotes a cardinality of "comment is reply of comment" relation and $m$ — cardinality of "comment has creator person".

Therefore, we propose a pretreatment phase that will compute number of replies once, which effectively eliminates the repeated calculations. Our goal is to find persons that made *not less than* $k$ comments replying to each other. For each pair of persons connected by an edge $e$ in the $graph$ we will determine the number of mutual replies $k_e$ and attribute it to $e$. In this way, BFS on each step compares two numbers: given $k$ and pre-calculated $k_e$.

### 4.2 Query Type 2 (Interests with Large Communities)

In order to reduce the overhead related to connected component size estimation one needs to take into account restrictions which are specified by the query. To tackle the first restriction (the common tag requirement) we built a "tag-person" index. It allows to search persons which are interested in a given tag. We employ the resulting list during the node traversal. It allows us to avoid visiting nodes (persons) which are not interested in a given tag. Also we avoid expenses related to probing person interest list for a given tag.

The second restriction which we have to take into account — the birthdate restriction. This restriction can be tackled by projecting our graph to a given time interval. By doing so, we avoid excessive comparisons related to birthdate which take place during the query processing. In this case the comparisons are moved to the preprocessing phase, thus providing us no benefit. However, this approach may be beneficial, if used differently. The idea is to produce a decomposition of the whole time interval into disjoint several time slices. During the query processing we can use the projection corresponding to an interval $d$, specified by the query. These projections are constructed during the preprocessing phase. Thus, we can avoid some excessive comparisons during the query processing phase.

Thereby, the estimation of the connected component size for a single tag is essentially a BFS, performed on a graph whose time slice conforms to the date specified by the query. This algorithm can be easily parallelized. For example, one can divide tag set between threads equally and then construct a final result by joining results for the individual tags.

### 4.3 Query Type 3 (Socialization Suggestion)

#### 4.3.1 Algorithm description.

The common sense may provide the following idea of the straightforward evaluation:

1. for each vertex $v$ in the $graph$ perform BFS while keeping in mind the given hops count $h$;
2. upon completion BFS returns the list of reached people $rp$;
3. for $v$ and each person $v_r$ from $rp$ check information about their work places, study places and location for correlation with $p$;
4. if one of the places where both $v$ and $v_r$ are involved is $p$ or its subplace, then calculate the number of common interests $ci$;
5. store (sorted by $ci$) the resulting pairs ($v$, $v_r$);
6. return the top-$k$ pairs as a result.

This algorithm requires examination of all the persons returned by BFS. Since $graph$ is a social its edge count follows power law, therefore there are some hubs and connectors with large degree and many vertices with only a few incident edges [11]. Hubs and connectors shorten the paths between persons and thus, the size of $rp$ may be significant. The time complexity of this algorithm is

$$O(|V| \cdot (|V| + |E| + |rp| \cdot |person.places| + |person.interests|)).$$

It is desirable to reduce the number of persons to examine without the loss of result correctness. In order to do that we suggest to group persons by some of place types. SNB provides three place types: $city$, $country$ and $continent$. The type $country$ seems to be a good choice (see [5] for the explanation).

Using the proposed partitioning we suggest a following improvement: use the type of $p$ to determine which country $c$ to process and then perform BFS for each person $v$ from $c$ bearing in mind the given hops count $h$. That

way only persons from $c$ are stored in $rp$, which reduces its size and allows us to reach our goal.

Described approach time complexity is

$$O(|persons\ in\ p| \cdot (|V| + |E| + |rp| \cdot |person.interests|)).$$

### 4.4 Query Type 4 (Most Central People)

#### 4.4.1 The calculation of closeness centrality metric

First of all, we should note, that our graph is an undirected graph, therefore **r(p)** can be calculated once for each connected component. Thus, the problem is how to compute **s(p)**.

**An algorithm selection.** Given the fact that our graphs is an undirected one and the edges are of unit weights, a simple BFS modification would suffice for the evaluation of **s(p)**. For this purpose we can label each visited vertex with the distance to the initial one. In this approach we do not increase asymptotic complexity of BFS and do not use additional memory. We would require $O(|V| + |E|)$ time and $O(|V| + |E|)$ memory. This estimation is better than estimation for many classical algorithms oriented for general cases of problem "minimal distance from one vertex to all other". For example, Dijkstra algorithm [7] for graphs with non-negative weights, based on Fibonacci heap [8] uses $O(|V| + |E|)$ memory and $O(|V| \cdot \log |V| + |E|)$ time. Moreover, our approach is easily parallelizable: we can compute **s(p)** in parallel for different vertices.

**The cut-off heuristic.** One can note that *closeness centrality* is inversely proportional to **s(p)** within a connected component. Thus, we can propose a criterion for a vertex to enter the *top-k* of a given connected component which uses it's **s(p)**. Let's define a threshold:

$$\Theta = \max_{p \in current\_top\_k} s(p) \quad .$$

Now, we can interrupt the computation of **s(p)**, if the current value had exceeded the threshold $\Theta$.

Despite the simplicity of this cut-off heuristics it drastically decreased the evaluation time for the fourth query type. Unfortunately, we do not know the number and parameters of queries of this type during the final evaluation. But the implementation of this heuristic allowed to decrease the evaluation time for more than 380 seconds on a graph containing 100 thousand persons. The resulting time was 220 seconds.

We also construct a special index structure for this type of query. More details can be found in the report [5].

**Other approaches.** In the last few days of the contest we found the solution that fits almost perfectly into the described problem [15]. It is developed for directed graphs with non-negative weights and reuses the CCV of a single vertex in order to estimate CCV for other vertices and reduce the further computations. Authors also use estimates in order to produce the cut-off of vertices which not to get into *top-k*. That method could be modified to take into account the memory restrictions. The experiments described by authors show that this approach may be particularly efficient for unweighted, undirected graph of a large size. It can reduce the amount of computations for a majority of vertices or even avoid their processing at all.

## 5 System architecture

**Graph structure.** Considering the graph structure we bear in mind the following: **(i)** the cardinality of vertices may run up to a million, **(ii)** BFS is crucial for the evaluation of every query type. Therefore, our approach must have low memory footprint and provide efficient BFS evaluation. In order to satisfy these requirements we use representation similar to adjacency lists, but with arrays instead, that is, each vertex contains a pointer to an array of adjacent vertices. It allows us to meet the memory constrains and avoid unnecessary comparisons in the BFS implementation.

**Layers.**Three layers may be distinguished in our implementation: **(i)** file loading, **(ii)** structure initialization and preparation, **(iii)** query evaluation.

This layered structure is rather natural to the task and allows some flexibility in the setting up the order of query evaluation. That is a rather important feature for the performance improvement. The use of the first layer is to provide the interface to chunk-based file loading. It copes with the problem of big files which can be up several gigabytes in size. The use of the second layer is to parse loaded files and to build indexes and other structures required for the query evaluation. The last layer is responsible for the final results formation.

## 6 Experiments

In this paper we present some experiments illustrating the performance of our approach. Unfortunately, we could not provide detailed experimental data from the contest due to several reasons: (i) we do not have access to the final benchmarks (they are not yet released to public); (ii) we no more have access to the hardware used for the evaluation by the organizers (it was a server-class one); (iii) the two largest benchmarking query sets are unavailable too (we used the largest available dataset — the medium dataset, containing 10k persons).

Thus, we had to perform experiments on our own. The hardware and software setup was the following: i7-4930K CPU (6 cores), P9 X79WS motherboard, 4GB RAM; Ubuntu 14.04, kernel 3.13.0-24, x86_64.

The first series of experiments is presented on Figure 1. They illustrate the basic approach when we sequentially evaluate queries of the same type. The results show the contribution of each query type to the overall processing time. In this series we vary the number of threads. Eventually we get a *U-shaped* graph, which shows that it's not useful to employ more than four threads for the processing in this scenario. It is the result of the algorithm parallelization imperfection (not all algorithms use

all cores all the time) and of the synchronization overheads. This leads us to the idea of pre-treatment phase which will allow us to balance the load. The load balancing will be done by grouping tasks together into stages and reordering of query types.

To examine our idea, we had split the query evaluation into the following stages (the stages are described in the [5]): **(i)** Q3 evaluation and Q1 preparation part 1, **(ii)** Q1 preparation part 2, Q2 preparation and Q4 preparation, **(iii)** Q1 evaluation, **(iv)** Q2 evaluation, **(v)** Q4 evaluation. Tasks belonging to one stage are executed in parallel. Figure 2 shows the results for this kind of processing. Despite that in fact we used our idea in the first two stages only, the performance boost of the evaluation with six threads is about 28% (compared to the best performance from Figure 1) and 56% comparing the performance with the six threads. This may be considered a good result for the medium dataset, which we use for testing. Efficiency of such task grouping is determined by the "closeness" of the tasks executed in parallel in terms of time. The closer times of execution, the more efficiently we use the processor. We can perform the load balancing in two ways: by varying the number of threads for one task and by varying the number of tasks. Hence we can use this approach to tune performance further. However, effects of the load balancing may vary with the dataset. Taking such variation into account is rather difficult and requires a more detailed study of the data structures and the algorithms involved.

## 7 Conclusions

In this paper we described the ACM SIGMOD Contest 2014, its tasks, timeline and our experiences. Also we presented our approach to the offered problems and described the advantages over the naive processing. We discussed algorithms as well as parallelization techniques and presented the general system architecture. Its key points are the following: query type intermixing, query type reordering, continuous query processing and block file loading techniques.

## References


[1] ACM SIGMOD 2014 Programming Contest website. http://www.cs.albany.edu/~sigmod14 contest. Accessed 23/05/14.

[2] Apache Giraph website. https://giraph.apache.org/. Accessed 23/05/2014.

[3] Azza Abouzeid, Kamil Bajda-Pawlikowski, Daniel Abadi, Avi Silberschatz, and Alexander Rasin. 2009. HadoopDB: an architectural hybrid of MapReduce and DBMS technologies for analytical workloads. Proc. VLDB Endow. 2, 1 (August 2009), 922–933.

[4] Ronnie Chaiken, Bob Jenkins, Per-Åke Larson, Bill Ramsey, Darren Shakib, Simon Weaver, and Jingren Zhou. 2008. SCOPE: easy and efficient parallel processing of massive data sets. Proc. VLDB Endow. 1, 2 (August 2008), 1265–1276.

[5] On Several Social Network Analysis Problems: a Report. George Chernishev, Vsevolod Sevostyanov, Kirill Smirnov, Ilya Shkuratov. Technical report. http://www.math.spbu.ru/user/chernishev/papers/sigmod2014contest-report.pdf

[6] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters. Commun. ACM 51, 1, 107–113.

[7] E. Dijkstra. 1959. A Note on Two Problems in Connexion with Graphs", Numerische mathematik, vol. 1, no. 1, 269–271.

[8] M. L. Fredman and R. E. Tarjan. 1984. Fibonacci Heaps And Their Uses In Improved Network Optimization Algorithms. In Proceedings of the 25th Annual Symposium on Foundations of Computer Science, 1984 (SFCS '84). IEEE Computer Society, Washington, DC, USA, 338–346.

[9] H. Herodotou, H. Lim, G. Luo, N. Borisov, L. Dong, F. Cetin, and S. Babu. Starfish: A Self-tuning System for Big Data Analytics. In Proc. of 5th Conf. on Innovative Data Systems Research (CIDR), 2011.

[10] InfiniteGraph: The Distributed Graph Database. Whitepaper. http://www.objectivity.com/wp-content/uploads/Objectivity_WP_IG_Distr_Benchmark.pdf. Accessed 23/05/2014.

[11] LDBC SocialNet Benchmark: Data Generation. https://github.com/ldbc/ldbc_socialnet_bm/wiki/Data-Generation#graph-generation. Accessed 23/05/2014.

[12] The Neo Database — A Technology Introduction (20061123). http://dist.neo4j.org/neo-technology-introduction.pdf. Accessed 23/05/2014.

[13] Christopher Olston, Benjamin Reed, Utkarsh Srivastava, Ravi Kumar, and Andrew Tomkins. 2008. Pig latin: a not-so-foreign language for data processing. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (SIGMOD '08). ACM, New York, NY, USA, 1099–1110.

[14] Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, and Theo Vassilakis. 2010. Dremel: interactive analysis of web-scale datasets. Proc. VLDB Endow. 3, 1–2 (September 2010), 330–339.

[15] Paul W. Olsen, Alan G. Labouseur, Jeong-Hyon Hwang. "Efficient Top-k Closeness Centrality

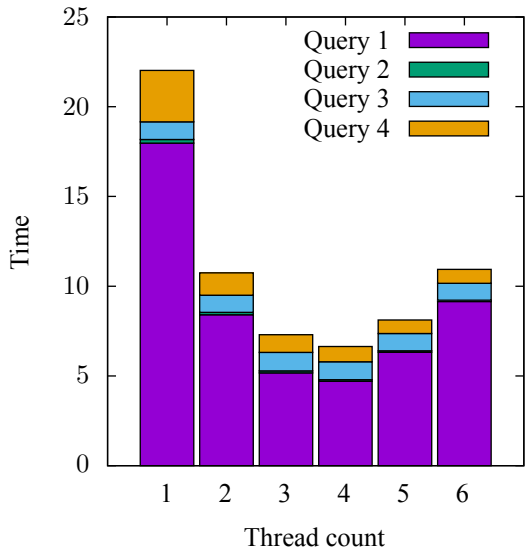



Figure 1: Performance scalability (without pre-treatment phase).

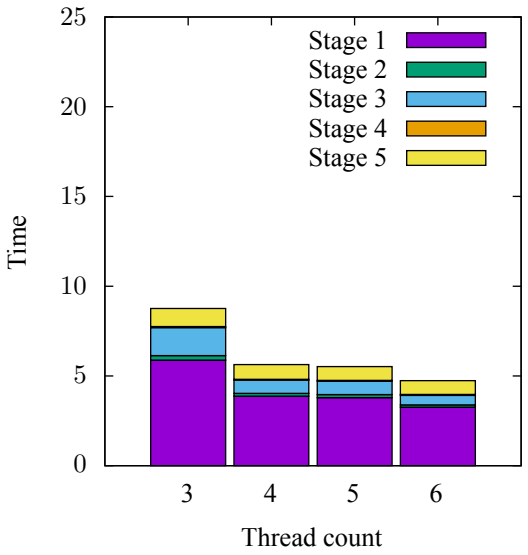



Figure 2: Performance scalability and effects of query reordering (pre-treatment phase).

Search". In Proceedings of the Data Engineering (ICDE), 2014 IEEE 30th International Conference, p 197-207, Chicago, IL, USA, 2014.

[16] Social Network Benchmark (SNB) Task Force Progress Report http://www.ldbc.eu:8090/download/attachments/4325436 /LDBC_SNB_Report_Nov2013.pdf. Accessed 23/05/14.

[17] Signal/Collect Documentation (website). http://uzh.github.io/signal-collect/documentation.html. Accessed 23/05/14.

[18] ACM SIGMOD Programming Contest: an opportunity to study distinguished aspects of database systems and software engineering. Kirill K. Smirnov, Georgiy A. Chernishev. 2012. Компьютерные инструменты в образовании, 6(2012), 22–25, ISSN: 2071-2340, url:http://ipo.spb.ru/journal/index.php?article/1541/ (in Russian).

[19] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff, and Raghotham Murthy. 2009. Hive: a warehousing solution over a map-reduce framework. Proc. VLDB Endow. 2, 2 (August 2009), 1626–1629.

## 8 Appendix: SNB Description

Let's briefly survey the SNB benchmark which was used during the contest and in the experimental section of this paper.

**The purpose.** In order to provide efficient evaluation for a variety of algorithms, tools, frameworks for social network data management tasks, a standard benchmark, called Social Network Benchmark (SNB) [16] was developed. This benchmark allows not only efficient, but also a repeatable evaluation for a variety of scenarios: online transactions, business intelligence and graph analytics. Authors of the benchmark tried to make it as realistic as possible.

**Covered systems.** This benchmark covers several types of systems: graph DBMS and graph programming frameworks, RDF database systems, relational and NoSQL database systems.

**Data schema.** The general data schema of the benchmark is presented on Figure 3 (illustration taken from [16]). It is called Social Intelligence Benchmark Data Schema. The schema uses UML notation to describe entities, attributes and their relationships of different cardinalities. The schema defines the result of the benchmark's data generator. Essentially it is a set of tables linked via primary-foreign key relationships.

The schema defines some social network and its most characteristic features:

1. users and their personal details, tags and likes;
2. relations between users (follows and knows);
3. textual content: posts and comment trees.

**Generator and its output: technical details.** This benchmark is essentially a synthetic data generator, which is implemented using MapReduce programming model. The generator is dictionary-based and is capable of generating correlated values. The result of the generator is the set CSV files, where each file contains records of the corresponding table.

**The benchmark and the contest.** The organizers of the contest used only the dataset generator, but not

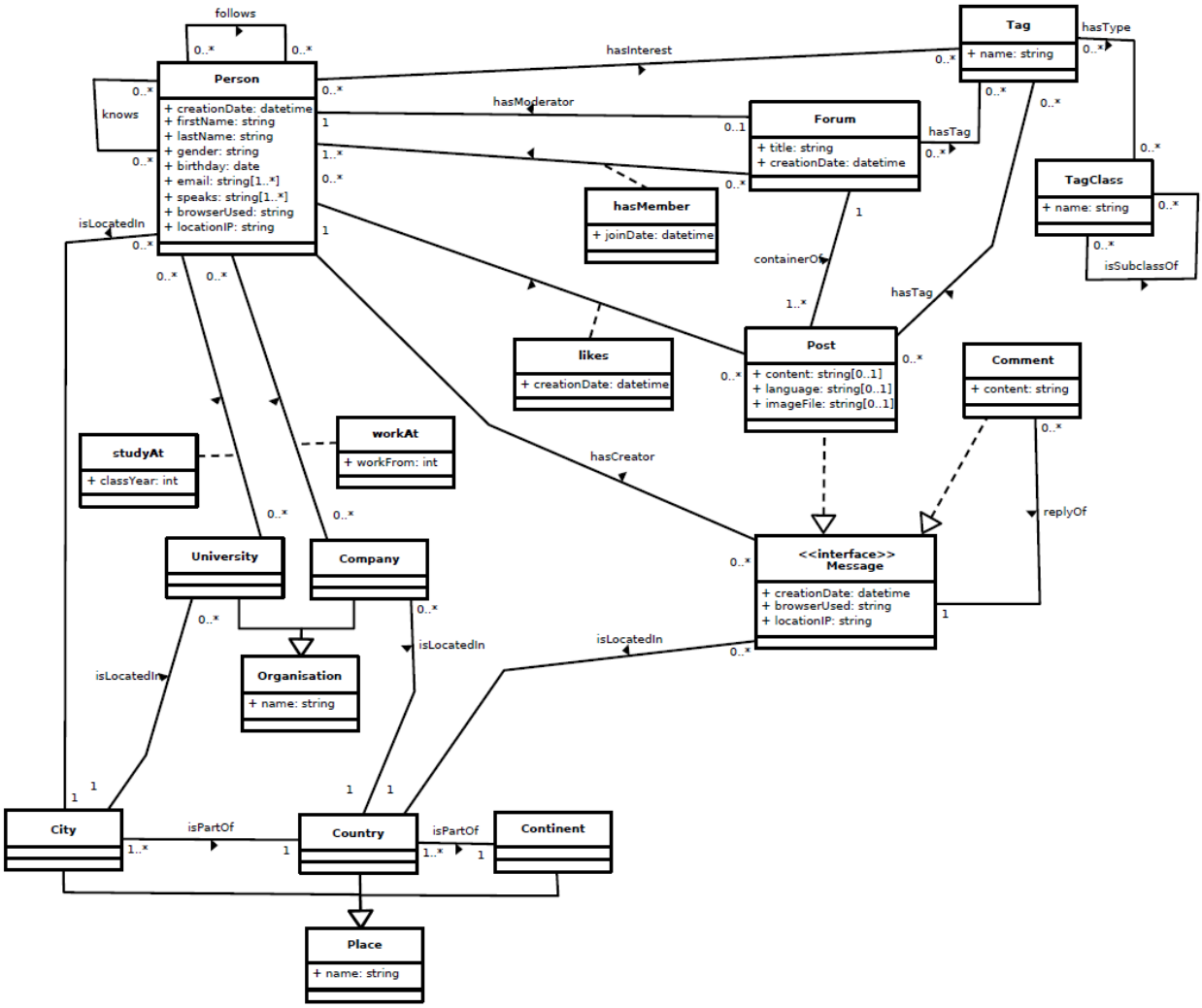



Figure 3: Social Intelligence Benchmark Data Schema

queries. Instead, they proposed four stand-alone types of queries, which we described earlier.

The dataset generator provided four types of graph workloads: small (1k vertices), medium (10k vertices), large (100k vertices) and huge (1M vertices). The last one would be used for the final evaluation by the contest organizers.

Unfortunately, only the first two datasets were fully released to the public. The third one was discussed in the mailing list, where some of the generator parameters for this dataset were disclosed. However, no queries are known. In this paper we use the largest available (on the current date) dataset — the medium one for the experimental evaluation. All of the queries are known at the start of the processing, contestants are not required to process them in a specific order.

# Тематические модели: учет сходства между униграммами и биграммами

© М. А. Нокель
МГУ им. М. В. Ломоносова, Москва
mnokel@gmail.com

## Аннотация

В статье представлены результаты экспериментов по добавлению сходства между униграммами и биграммами в тематические модели. Вначале изучается возможность применения ассоциативных мер для выбора и последующего включения биграмм в тематические модели. Затем предлагается модификация оригинального алгоритма PLSA, учитывающая похожие униграммы и биграммы, начинающиеся с одних и тех же букв. И в конце статьи предлагается новый итеративный алгоритм без учителя, показывающий, как темы сами могут выбирать себе наиболее подходящие биграммы. В качестве текстовой коллекции была взята подборка статей из электронных банковских журналов на русском языке. Эксперименты показывают значительное улучшение качества тематических моделей по всем целевым метрикам.

## 1 Введение

*Вероятностные тематические модели* (далее просто *тематические модели*) – одно из современных приложений машинного обучения к анализу текстов. Тематические модели предназначены для описания текстов с точки зрения их тем. Они определяют, к каким темам относится каждый документ в текстовой коллекции и какие слова образуют каждую такую тему. При этом темы представляются в виде дискретных распределений на множестве слов, а документы – в виде дискретных распределений на множестве тем [1]. Пользователям темы предоставляются, как правило, в виде некоторых списков часто встречающихся рядом друг с другом слов, упорядоченных по убыванию степени принадлежности им.

**Труды 16-й Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" – RCDL-2014, Дубна, Россия, 13-16 октября 2014 г.**

С момента своего появления тематические модели достигли значительных успехов в задачах информационного поиска [2], разрешении морфологической неоднозначности [3], многодокументного аннотирования [4], кластеризации и категоризации документов [5]. Также они успешно применяются в выявлении трендов в научных публикациях и новостных потоках [6], обработке аудио- и видео-сигналов [7] и других задачах. Самыми известными представителями являются латентное размещение Дирихле (LDA) [1], использующее априорное распределение Дирихле, и метод вероятностного латентного семантического анализа (PLSA) [8], не связанный ни с какими параметрическими априорными распределениями.

В работах [9] и [10] было показано, что использование тематических моделей в задаче извлечения однословных терминов способно значительно улучшить качество извлечения последних из текстов предметных областей. Поэтому актуальной является и проблема улучшения качества самих тематических моделей за счет использования некоторой лингвистической информации, чему и посвящена данная работа.

Одним из главных недостатков тематических моделей является использование модели "мешка слов", в которой каждый документ рассматривается как набор встречающихся в нем слов. Данная модель не учитывает порядок слов и основывается на гипотезе независимости появлений слов в документах друг от друга. На данный момент проведено множество исследований, посвященных изучению вопроса добавления словосочетаний, n-грамм и многословных терминов в тематические модели. Однако часто это приводит к ухудшению качества модели в связи с увеличением размера словаря или к значительному усложнению модели [12], [13], [14].

В статье предлагается новый подход, позволяющий учесть взаимосвязь между похожими словами (в частности, однокоренными) в тематических моделях (такими, как *банк – банковский – банкир*, *кредит – кредитный – кредитовать – кредитование*). На основании данного метода в статье описывается и новый подход к добавлению биграмм в тематические модели, который рассматривает биграммы уже не как "черные ящики", а

учитывает взаимосвязь между ними и униграммами, основанную на их внутренней структуре. Предлагаемый алгоритм улучшает качество тематических моделей по двум целевым метрикам: перплексии и согласованности тем [15].

Все эксперименты, описанные в статье, проведены на основе алгоритма PLSA и его модификаций на коллекции текстов банковской тематики на русском языке, взятых из электронных журналов.

Статья организована следующим образом. В разделе 2 рассматриваются близкие работы. В разделе 3 описывается текстовая коллекция, использующаяся в экспериментах, все стадии её предобработки и метрики, применяемые для оценивания качества работы тематических моделей. В разделе 4 проводится обширный анализ ассоциативных мер для выбора и последующего включения биграмм в тематические модели. В разделе 5 предлагается новый алгоритм, позволяющий учесть сходство между униграммами и биграммами в тематических моделях. В разделе 6 предлагается еще один новый итеративный алгоритм, использующий тот факт, что темы могут сами выбирать себе наиболее подходящие биграммы. И в последнем разделе приводятся выводы.

## 2 Близкие работы

### 2.1 Тематические модели

На сегодняшний день разработано достаточно много различных тематических моделей. Исторически одними из первых появились модели, основанные на традиционных методах кластеризации текстов [11]. При этом после окончания работы алгоритма кластеризации каждый получившийся кластер рассматривается как отдельная тема для вычисления вероятностей входящих в него слов по следующей формуле:

$$P(w|t) = \frac{f(w|t)}{\sum\limits_{w} f(w|t)}$$

где $f(w|t)$ – частотность слова $w$ в теме $t$.

Естественным ограничением таких моделей является отнесение каждого документа лишь к одной теме.

В последнее время появились вероятностные механизмы нахождения тем в документах, рассматривающие каждый документ в виде смеси тем, а каждую тему в виде некоторого вероятностного распределения над словами. Вероятностные модели порождают слова по следующему правилу:

$$P(w|d) = \sum_{t} P(w|t)P(t|d)$$

где $P(t|d)$ и $P(w|t)$ – распределение тем по документам и слов по темам, а $P(w|d)$ – наблюдаемое распределение слов по документам.

Согласно данной модели коллекция $D$ – это выборка наблюдений $(d, w)$, генерируемых Алгоритмом 1.

**Algorithm 1:** Порождение коллекции текстов с помощью тематической модели

**Input**: распределения $P(w|t)$ и $P(t|d)$
**Output**: коллекция $D = \{(d, w)\}$

```
for d ∈ D do
    Задать длину n_d документа d
    for i = 1, ..., n_d do
        Выбрать тему t из P(t|d)
        Выбрать слово w из P(w|t)
        Добавить в D пару (d, w)
```

Самыми известными представителями данной категории являются метод вероятностного латентного семантического анализа (PLSA) [8] и латентное размещение Дирихле (LDA) [1].

### 2.2 Словосочетания в тематических моделях

Все описанные в прошлом разделе алгоритмы работают только со словами, основываясь на гипотезе о независимости слов друг от друга – модели “мешка слов”. Идея же использования словосочетаний в тематических моделях сама по себе не нова. На данный момент существуют 2 подхода к решению данной проблемы: создание унифицированной вероятностной модели и предварительное извлечение словосочетаний и n-грамм для их последующего добавления в тематические модели.

Большинство исследований на данный момент посвящено первому подходу. Так, первая попытка выйти за пределы модели “мешка слов” была предпринята в работе [12], где была представлена Биграммная Тематическая Модель. В этой модели вероятности слов зависят от вероятностей непосредственно предшествующих им слов. Модель словосочетаний LDA расширяет Биграммную Тематическую Модель за счет введения дополнительных переменных, способных генерировать и униграммы, и биграммы. В работе [14] представлена Тематическая N-граммная Модель, усложняющая предыдущие для обеспечения возможности формирования биграмм в зависимости от контекста. В работе [16] предложена тематическая модель Слово-Символ, выходящая за рамки использовавшегося ранее предположения о том, что тема каждой n-граммы определяется в зависимости от тем слов, составляющих данное словосочетание. Эта модель оказалась наиболее пригодной для китайского языка. В работе [17] устанавливается связь между LDA и вероятностными контекстно-свободными грамматиками и предлагаются две но-

вые вероятностные модели, сочетающие в себе идеи из LDA и вероятностных контекстно-свободных грамматик для добавления словосочетаний и имен собственных в тематические модели.

Несмотря на то, что все описанные выше модели имеют теоретически элегантное обоснование, у них очень большая вычислительная сложность, что ведёт к неприменимости на реальных данных. Так, например, вычислительная сложность Биграммной Тематической Модели равна $O(W^2T)$, в то время как для LDA она равна $O(WT)$, для PLSA – $O(WT+DT)$, где $W$ – размер словаря, $D$ – количество документов в коллекции и $T$ – число тем. Поэтому такие модели представляют в основном чисто теоретический интерес.

Алгоритм, предложенный в работе [18], относится ко второму типу методов, добавляющих словосочетания в тематические модели. На этапе предобработки авторы извлекают биграммы с помощью $t$-теста и заменяют отдельные униграммы лучшими по данной мере биграммами. При этом используются 2 метрики оценивания качества полученных тем: перплексия и согласованность тем [15]. В статье показано, что добавление биграмм в тематические модели приводит к ухудшению перплексии и к улучшению согласованности тем.

Данная работа также относится ко второму типу методов и отличается от работы [18] в том, что описываемый здесь подход учитывает внутреннюю структуру биграмм и взаимосвязь между ними и составляющими их униграммами, что приводит к улучшению обоих показателей: и перплексии, и согласованности тем.

Идея использования априорных лингвистических знаний в тематических моделях сама по себе не нова. Так, в работе [19] предметно-ориентированные знания представляются в виде Must-Link и Cannot-Link примитивов с помощью априорного леса Дирихле. Эти примитивы отвечают за то, чтобы слова порождались одними и теми же или, наоборот, разными темами. Однако позднее было замечено, что данный метод может привести к экспоненциальному росту при кодировании Cannot-Link примитивов, и потому его сложно применять с большим количеством ограничений [20]. Другой способ включения подобных знаний представлен в работе [21], где был предложен частично обучаемый с учителем EM-алгоритм для группировки выражений в некоторые заданные пользователем категории. Для обеспечения наилучшей инициализации EM-алгоритма предложенный в статье метод использует априорное знание о том, что синонимы и выражения, имеющие одинаковые слова, должны, скорее всего, относиться к одним и тем же группам. Данная работа отличается от приведённых выше тем, что в ней сходства между униграммами и биграммами добавляются в тематическую модель естественным образом путем подсчета их совместной встречаемости в документах коллекции. Предлагаемый подход никак не увеличивает вычислительную сложность оригинального алгоритма PLSA.

## 3 Текстовая коллекция и методы оценивания качества тематических моделей

### 3.1 Текстовая коллекция и предобработка

В экспериментах, описанных в данной статье, использовалась текстовая коллекция из 10422 статей на русском языке, взятых из некоторых электронных банковских журналов (таких, как Аудитор, РБК, Банковский журнал и др.). В данных документах содержится почти 15.5 млн слов.

На этапе предобработки был проведен морфологический анализ документов. В экспериментах рассматривались только *существительные*, *прилагательные*, *глаголы* и *наречия*, поскольку служебные слова не играют значительной роли в определении тем. Кроме того, из рассмотрения исключались слова, встретившиеся менее 5 раз во всей текстовой коллекции.

На этапе предобработки из документов также извлекались биграммы в формах *сущ. + сущ. в родительном падеже* и *прил. + сущ.* В экспериментах рассматривались только такие биграммы, поскольку темы, как правило, задаются именными группами.

### 3.2 Методы оценивания качества тематических моделей

Для оценивания качества полученных тем в статье рассматриваются две метрики.

Во-первых, использовалась *перплексия*, являющаяся стандартным критерием качества тематических моделей [22]. Эта мера несоответствия модели $p(w|d)$ словам $w$, наблюдаемым в документах коллекции, определяется через логарифм правдоподобия:

$$Perplexity(D) = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw}\ln p(w|d)\right)$$

где $n$ – число всех рассматриваемых слов в текстовой коллекции, $D$ – множество всех документов в коллекции, $n_{dw}$ – частота слова $w$ в документе $d$, $p(w|d)$ – вероятность появления слова $w$ в документе $d$.

Чем меньше значение перплексии, тем лучше модель предсказывает появление слов $w$ в документах коллекции $D$. Поскольку известно, что перплексия, вычисленная на той же самой обучающей коллекции документов, склонна к переобучению и может давать оптимистически заниженные

значения [1], в данной статье используется стандартный метод вычисления контрольной перплексии, описанный в работе [24]. Коллекция документов изначально разбивалась на 2 части: обучающую $D$, по которой строилась модель, и контрольную $D'$, по которой вычислялась данная метрика. Хотя на данный момент существует множество исследований, утверждающих, что перплексию нельзя применять для оценивания качества тематических моделей [23], данная метрика попрежнему широко используется для сравнения различных тематических моделей.

В то же время неоднократно предпринимались попытки предложить способ автоматического оценивания качества тематических моделей, никак не связанного с перплексией и коррелирующего с мнениями экспертов. Данная постановка задачи является очень сложной, поскольку эксперты могут достаточно сильно расходиться во мнениях. Однако в недавних работах [15], [25] было показано, что возможно автоматически оценивать *согласованность тем*, основываясь на семантике слов с точностью, почти совпадающей с экспертами. Предложенная метрика измеряет интерпретируемость тем, основываясь на способах оценивания экспертом [15]. Поскольку темы, как правило, предоставляются экспертам для проверки в виде первых топ-$N$ слов, согласованность тем оценивает то, насколько данные слова соответствуют рассматриваемой теме. Newman в работе [15] предложил использовать автоматический способ вычисления данной метрики исходя из меры взаимной информации:

$$TC\text{-}PMI(t) = \sum_{j=2}^{10} \sum_{i=1}^{j-1} \log \frac{P(w_j, w_i)}{P(w_j)P(w_i)}$$

где $(w_1, w_2, \ldots, w_{10})$ – топ-10 слов в рассматриваемой теме $t$, $P(w_i)$ и $P(w_j)$ – вероятности униграмм $w_i$ и $w_j$ соответственно, а $P(w_j, w_i)$ – вероятность биграммы $(w_j, w_i)$. Итоговая мера согласованности тем вычисляется усреднением $TC\text{-}PMI(t)$ по всем темам $t$.

Данная метрика показывает очень высокую корреляцию с оценками экспертов [15]. Предложенная метрика рассматривает только первые топ-10 слов в каждой теме, поскольку они, как правило, предоставляют достаточно информации для формирования предмета темы и отличительных черт одной темы от другой. Согласованность тем становится все более широко используемым способом оценивания качества тематических моделей наряду с перплексией. Так, в работе [26] также было показано, что данная метрика очень сильно коррелирует с оценками экспертом. А в работе [27] она просто используется для оценки качества полученных тем.

В соответствии с подходом, изложенным в работе [25], в данной статье вероятности униграмм и биграмм вычисляются путем деления количества документов, в которых встретилась та или иная униграмма или биграмма, на число всех документов в коллекции. Другой вариант вычисления меры согласованности тем на основе логарифма от условной вероятности ($TC\text{-}LCP$), предложенный в работе [25], не рассматривается, поскольку в работе [18] было показано, что этот вариант работает значительно хуже, чем $TC\text{-}PMI$.

## 4 Добавление биграмм в тематические модели

На первом этапе экспериментов исследовалось, может ли улучшиться качество тематической модели путем добавления в неё биграмм в качестве отдельных элементов словаря. Для этой цели были извлечены все биграммы, встретившиеся в коллекции, с частотностью не меньше 5. Для последующего упорядочения извлечённых биграмм применялись *ассоциативные меры* – математические критерии, определяющие силу связи между составными частями фраз, основываясь на частотах встречаемости отдельных слов и словосочетаний целиком. В экспериментах были использованы следующие 15 ассоциативных мер: *Взаимная Информация (MI)* [28], *Дополненная Взаимная Информация (Дополненная MI)* [29], *Кубическая Взаимная Информация (Кубическая MI)* [30], *Нормализованная Взаимная Информация (Нормализованная MI)* [31], *Настоящая Взаимная Информация (Настоящая MI)*, *Коэффициент Dice (DC)* [32], *Модифицированный Коэффициент Dice (Модифицированный DC)* [33], *T-Score*, *Симметричная Условная Вероятность* [34], *Коэффициент Простого Соответствия*, *Коэффициент Kulczinsky*, *Коэффициент Yula* [30], *Хи-Квадрат*, *Отношение логарифмического правдоподобия* [35] и *Лексическая Связность* [36].

В соответствии с результатами [18] в тематические модели добавлялись топ-1000 биграмм для каждой ассоциативной меры. Так, в каждом эксперименте к словарю в качестве отдельных элементов добавлялись топ-1000 биграмм, и в каждом документе, содержащем любые из добавляемых словосочетаний, из частот образующих их униграмм вычитались частоты биграмм, а сами словосочетания добавлялись в его разреженное представление. Отдельно следует отметить, что во всех экспериментах число топиков фиксировалось равным 100.

Хотя эксперименты были проведены для всех 15 упомянутых выше ассоциативных мер, в таблице 1 представлены только наиболее характерные результаты добавления топ-1000 биграмм наряду с результатом оригинального алгоритма PLSA без добавления биграмм (значения, выделенные полужирным шрифтом, соответствуют улучшению по

одному из критериев).

| Ассоциативная мера | Перплексия | TC-PMI |
|---|---|---|
| *Оригинальный PLSA* | 1694 | 86.4 |
| *MI* | **1683** | 79.2 |
| *Настоящая MI* | 2162 | **110.7** |
| *Кубическая MI* | 2000 | **95** |
| *DC* | 1777 | **89.6** |
| *Модифицированный DC* | 2134 | **94.1** |
| *T-Score* | 2189 | **104.9** |
| *Лексическая Связность* | 1928 | **101.3** |
| *Хи-Квадрат* | 1763 | **89.6** |

Таблица 1: Результаты добавления биграмм в тематическую модель

Как видно, добавление топ-1000 биграмм, упорядоченных по той или иной ассоциативной мере, как правило, приводит к увеличению размера словаря и, следовательно, ухудшению перплексии, в то время как согласованность тем становится лучше. Эти выводы полностью согласуются с результатами, описанными в работе [18]. Однако, используя некоторые ассоциативные меры (например, Взаимную Информацию), можно получить немного лучше перплексию, но чуть хуже согласованность тем, что обусловлено добавлением нестандартных и низкочастотных биграмм.

## 5 Добавление схожих униграмм и биграмм в тематические модели

### 5.1 Добавление схожих униграмм в тематические модели

Оригинальные тематические модели (PLSA и LDA) используют модель "мешка слов", предполагающую независимость слов друг от друга. Однако в документах есть много слов, связанных между собой по смыслу – в частности, однокоренные слова, например: *банк – банковский – банкир*, *кредит – кредитный – кредитовать – кредитование* и др. Поэтому на следующем этапе экспериментов исследовалась возможность учета в тематических моделях подобных похожих слов – а именно, слов, начинающихся с одних и тех же букв.

Для данной цели был модифицирован оригинальный алгоритм PLSA. При описании проведённой модификации будет использоваться описание алгоритма PLSA, представленное в работе [37], и следующие обозначения:

- $D$ – коллекция документов;
- $T$ – множество полученных тем;
- $W$ – словарь (множество уникальных слов в коллекции документов $D$);
- $\Phi = \{\phi_{wt} = p(w|t)\}$ – распределение слов $w$ по темам $t$;
- $\Theta = \{\theta_{td} = p(t|d)\}$ – распределение тем $t$ по документам $d$;
- $S = \{S_w\}$ – множество похожих слов, где $S_w$ – множество слов, похожих на $w$;
- $n_{dw}$ и $n_{ds}$ – частотности слов $w$ и $s$ в документе $d$;
- $\hat{n}_{wt}$ – оценка частотности слова $w$ в теме $t$;
- $\hat{n}_{td}$ – оценка частотности темы $t$ в документе $d$;
- $\hat{n}_t$ – оценка частотности темы $t$ в коллекции документов $D$.

Псевдокод алгоритма PLSA-SIM представлен в Алгоритме 2. Единственная модификация оригинального алгоритма PLSA касается строчки 6, где в рассмотрение добавляются предварительно вычисленные множества похожих слов (в оригинальном алгоритме данная строчка отсутствует, а в строчке 9 вместо $f_{dw}$ используется $n_{dw}$). Тем самым вес подобных слов увеличивается в каждом документе коллекции.

**Algorithm 2:** PLSA-SIM алгоритм: PLSA с похожими словами

**Input**: коллекция документов $D$,
количество тем $|T|$,
начальные приближения $\Phi$ и $\Theta$,
множества похожих слов $S$

**Output**: распределения $\Phi$ и $\Theta$

1 **while** *не выполнится критерий остановки* **do**
2  **for** $d \in D$, $w \in W$, $t \in T$ **do**
3  $\hat{n}_{wt} = 0$, $\hat{n}_{td} = 0$, $\hat{n}_t = 0$
4  **for** $d \in D$, $w \in W$ **do**
5  $Z = \sum_t \phi_{wt}\theta_{td}$,
6  $f_{dw} = n_{dw} + \sum_{s \in S_w} n_{ds}$
7  **for** $t \in T$ **do**
8  **if** $\phi_{wt}\theta_{td} > 0$ **then**
9  $\delta = f_{dw}\phi_{wt}\theta_{td}/Z$
10  $\hat{n}_{wt} = \hat{n}_{wt} + \delta$
11  $\hat{n}_{td} = \hat{n}_d + \delta$
12  $\hat{n}_t = \hat{n}_t + \delta$
13  **for** $w \in W$, $t \in T$ **do**
14  $\phi_{wt} = \hat{n}_{wt}/\hat{n}_t$
15  **for** $d \in D$, $t \in T$ **do**
16  $\theta_{td} = \hat{n}_{td}/\hat{n}_t$

Поскольку в русском языке достаточно богатая морфология, а темы в основном задаются именными группами, в качестве потенциальных кандидатов в похожие слова рассматривались только существительные и прилагательные. В таблице 2 представлены результаты добавления похожих слов в тематические модели наряду с оригинальным алгоритмом PLSA (значения, выделенные полужирным шрифтом, соответствуют лучшим значениям по одному из критериев).

| Число одинаковых букв | Перплексия | TC-PMI |
|---|---|---|
| *0 букв (PLSA)* | 1694 | 86.4 |
| *2 буквы* | 1852 | 187.2 |
| *3 буквы* | 1565 | 432.9 |
| *4 буквы* | **1434** | **2432.3** |
| *5 букв* | 1620 | **2445.3** |
| *6 букв* | 1610 | 1310.85 |

Таблица 2: Результаты экспериментов по добавлению похожих униграмм в тематическую модель

Как видно, наилучшие результаты показывает модель, рассматривающая в качестве похожих слова, начинающиеся с 4 одинаковых букв. Однако в русском языке есть множество приставок длины в 4 буквы и больше. Учитывая это, был составлен список из 43 наиболее широко использующихся таких приставок (*анти-*, *гипер-*, *пере-* и др.) и введён дополнительный критерий: если слова начинаются на одну и ту же приставку, то они считаются похожими, если следующая буква после приставки также совпадает. Данный критерий позволил еще больше снизить перплексию до **1376** и оставить согласованность тем примерно на лучшем уровне – **2250**. В дальнейших экспериментах, описываемых в данной статье, было решено использовать именно эти 2 критерия.

Следует отметить, что в результате добавления знаний о похожести слов в тематические модели такие слова с большей вероятностью окажутся в топ-10 в полученных темах. Тем самым происходит неявная максимизация меры *TC-PMI*, поскольку похожие слова склонны встречаться в одних и тех же документах. Поэтому было принято решение модифицировать данную метрику для учета не всех топ-10 слов, а только топ-10 непохожих слов в темах (в дальнейшем в статье данная метрика будет обозначаться как *TC-PMI-nSIM*). В таблице 3 подытожены результаты добавления похожих слов в тематические модели с использованием описанных выше критериев и введённой новой метрики:

| Алгоритм | Перплексия | TC-PMI-nSIM |
|---|---|---|
| *Исходный PLSA* | 1694 | 78.3 |
| *PLSA-SIM* | **1376** | **87.8** |

Таблица 3: Результаты наилучших способов добавления похожих слов в тематическую модель

Как видно, модифицированная версия алгоритма PLSA-SIM показывает результаты лучше оригинального алгоритма PLSA по обоим целевым метрикам. В таблице 4 представлены топ-5 слов, взятых из двух случайно выбранных тем для оригинального и модифицированного алгоритмов.

| PLSA алгоритм | | PLSA-SIM алгоритм | |
|---|---|---|---|
| Бумага | Документ | Аудитор | Правый |
| Ценный | Электронный | Аудиторский | Право |
| Акция | Форма | Аудитор | Правило |
| Рынок | Организация | Аудируемый | Акция |
| Облигация | Подпись | Проверка | Акционер |

Таблица 4: Топ-5 слов, взятых из тем, полученных с помощью алгоритмов PLSA и PLSA-SIM

### 5.2 Добавление схожих биграмм в тематические модели

Для применения подхода, представленного в разделе 5.1 к топ-1000 биграммам, упорядоченными в соответствии с различными ассоциативными мерами, описанными в разделе 4, было решено ввести дополнительный критерий схожести биграмм и униграмм. Биграмма $(w_1, w_2)$ считается похожей на униграмму $w_3$, если выполнен один из следующих критериев:

- слово $w_3$ похоже на $w_1$ или $w_2$ в соответствии с критериями, описанными в разделе 5.1;
- слово $w_3$ совпадает с $w_1$ или $w_2$ и длина $w_3$ больше трех букв.

Хотя эксперименты были проведены для всех ассоциативных мер, описанных в разделе 4, в таблице 5 представлены только наиболее характерные результаты интеграции биграмм и добавлению похожести униграмм и биграмм наряду с результатами алгоритмов PLSA и PLSA-SIM (значения, выделенные полужирным шрифтом, соответствуют лучшим значениям по одному из критериев).

| Алгоритм | Перплексия | TC-PMI-nSIM |
|---|---|---|
| *PLSA* | 1694 | 78.3 |
| *PLSA-SIM* | 1376 | 87.8 |
| *PLSA-SIM + MI* | 1411 | 106.2 |
| *PLSA-SIM + Настоящая MI* | 1204 | **177.8** |
| *PLSA-SIM + Кубическая MI* | 1186 | 151.7 |
| *PLSA-SIM + DC* | 1288 | 99 |
| *PLSA-SIM + Модифицированный DC* | **1163** | 156.2 |
| *PLSA-SIM + T-Score* | 1222 | 171.5 |
| *PLSA-SIM + Лексическая связность* | 1208 | 125.6 |
| *PLSA-SIM + Хи-квадрат* | 1346 | 122.9 |

Таблица 5: Результаты добавления похожих униграмм и биграмм в тематическую модель

Как видно, добавление в тематическую модель похожих униграмм и топ-1000 биграмм, упорядоченных в соответствии с большинством ассоциа-

тивных мер, приводит к улучшению качества получающихся тем по сравнению с алгоритмом PLSA-SIM. В таблице 6 представлены топ-5 униграмм и биграмм, взятых из двух случайно выбранных тем, полученных с помощью алгоритма PLSA-SIM с добавлением топ-1000 биграмм, упорядоченных Модифицированным Коэффициентом Dice (Модифицированным DC), для которого достигаются наилучшее значение перплексии.

| | |
|---|---|
| Инвестиция | Финансовый рынок |
| Инвестор | Финансовая система |
| Инвестирование | Финансовый |
| Иностранный инвестор | Финансовый институт |
| Иностранное инвестирование | Финансовый ресурс |

Таблица 6: Топ-5 униграмм и биграмм, взятых из тем, полученных с помощью PLSA-SIM с биграммами, упорядоченными Модифицированным DC

## 6 Итеративный алгоритм для выбора наиболее подходящих биграмм

На последнем этапе экспериментов было сделано предположение, что темы могут сами выбирать себе наиболее подходящие биграммы. Для проверки данной гипотезы был предложен новый итеративный алгоритм выбора биграмм исходя из вида верхушек тем.

При описании предлагаемого алгоритма будут использоваться следующие дополнительные обозначения:

- $B$ – множество всех биграмм в коллекции документов $D$;
- $B_A$ – множество биграмм, добавленных в тематическую модель;
- $S_A$ – множество потенциальных кандидатов на похожие слова;
- $(u_1^t, \dots, u_{10}^t)$ – топ-10 униграмм в теме $t$;
- $f(u_1^t, u_2^t)$ – частота биграммы $(u_1^t, u_2^t)$.

Псевдокод предлагаемого алгоритма представлен в Алгоритме 3. На каждой итерации алгоритм добавляет в множество кандидатов в похожие слова топ-10 униграмм из каждой темы. Также в это же множество и в саму тематическую модель добавляются все биграммы, которые могут быть образованы с помощью этих топ-10 униграмм. Было принято решение анализировать только первые топ-10 слов в темах, поскольку одной из целевой метрик является согласованность тем, использующая именно это множество (см. определение метрики в разделе 3). В соответствии с данным алгоритмом темы могут выбирать себе только те биграммы, которые образуются с помощью топ-10 униграмм в темах, а такие биграммы с большей вероятностью могут оказаться наиболее подходящими.

**Algorithm 3:** Итеративный алгоритм

**Input**: коллекция документов $D$,
число тем $|T|$,
множество биграмм $B$
**Output**: полученные темы

```
Запуск оригинального PLSA на коллекции документов D для получения тем T
B_A = ∅
while не выполнится критерий остановки do
    S_A = ∅
    for t ∈ T do
        S_A = S_A ∪ {u_1^t, u_2^t, ..., u_10^t}
        for u_i^t, u_j^t ∈ (u_1^t, u_2^t, ..., u_10^t) do
            if (u_i^t, u_j^t) ∈ B and
            f(u_i^t, u_j^t) > f(u_j^t, u_i^t) then
                B_A = B_A ∪ {(u_i^t, u_j^t)}
    S_A = S_A ∪ B_A
    Запуск PLSA-SIM с множеством похожих слов S_A и с множеством биграмм B_A для получения тем T
```

В таблице 7 представлены первые несколько итераций предложенного итеративного алгоритма наряду с результатами оригинального алгоритма PLSA (в таблице обозначен как нулевая итерация).

| **Итерация** | **Перплексия** | **TC-PMI-nSIM** |
|---|---|---|
| *0 (PLSA)* | 1694 | 78.3 |
| *1* | **936** | **180.5** |
| *2* | 934 | 210.2 |
| *3* | 933 | 230 |
| *4* | 940 | **235.8** |
| *5* | **931** | 193.5 |

Таблица 7: Результаты итеративного алгоритма построения тематической модели

Как видно, после первой итерации наблюдается существенное улучшение качества получаемых тем по обеим целевым метрикам. Однако на следующих итерациях результаты начинают колебаться вокруг примерно тех же самых уровней перплексии и согласованности тем (с незначительным улучшением последней). Поэтому мы считаем, что согласно результатам первой итерации выбор необходимых биграмм и кандидатов в похожие слова самими темами приводит к наилучшим значениям перплексии и согласованности тем. В таблице 8 приведены топ-5 униграмм и биграмм, взятых из двух случайно выбранных тем, полученных после первой итерации предложенного алгоритма.

| | |
|---|---|
| Банковский кредит | Ипотечный банк |
| Банковский сектор | Ипотечный кредит |
| Кредитование | Ипотечное кредитование |
| Кредитная система | Жилищное кредитование |
| Кредит | Ипотека |

Таблица 8: Топ-5 униграмм и биграмм, взятых из тем, полученных с помощью итеративного алгоритма построения тематической модели

## 7 Благодарности

Работа частично поддержана грантом РФФИ 14-07-00383.

## 8 Заключение

В работе представлены эксперименты по добавлению биграмм в тематические модели. Эксперименты, проведённые на русскоязычных статьях из электронных банковских журналов, показывают, что большинство ассоциативных мер упорядочивает биграммы таким образом, что при добавлении верхушки этих списков в тематические модели ухудшается перплексия и улучшается согласованность тем. Затем в статье предлагается новый алгоритм PLSA-SIM, добавляющий схожесть униграмм и биграмм в тематические модели. Проведённые эксперименты показывают значительное улучшение перплексии и согласованности тем для этого алгоритма. В конце статьи предлагается еще один новый итеративный алгоритм, основанный на идее, что темы сами могут выбирать себе наиболее подходящие биграммы и похожие слова. Эксперименты показывают дальнейшее улучшение качества по обеим целевым метрикам.

## Список литературы

[1] D. Blei, A. Ng and M. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, No. 3, pp. 993–1002, 2003.

[2] X. Wei and B. Croft. LDA-based document models for ad-hoc retrieval. In the Proceedings of the 29th International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 178–185, 2006.

[3] J. Boyd-Grabber, D. Blei and X. Zhu. A Topic Model for Word Sense Disambiguation. In the Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Processing, pp. 1024–1033, 2007.

[4] D. Wang, S. Zhu, T. Li, and Y. Gong. Multi-Document Summarization using Sentence-based Topic Models. In the Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp. 297–300, 2009.

[5] S. Zhou, K. Li, and Y. Liu. Text Categorization Based on Topic Model. International Journal of Computational Intelligence Systems, Vol. 2, No. 4, pp. 398–409, 2009.

[6] L. Bolelli, Ş. Ertekin, C. L. Giles. Topic and Trend Detection in Text Collections Using Latent Dirichlet Allocation. In ECIR Proceedings, Lecture Notes in Computer Science, Vol. 5478, pp. 776–780, 2009.

[7] T. Hyunh, M. Fritz, B. Schiele. Discovery of activity patterns using topic models. In the Proceedings of the 10th international conference on Ubiquitous computing, pp. 10–19, 2008.

[8] T. Hofmann. Probabilistic Latent Semantic Indexing. In the Proceedings of the 22nd Annual International SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57, 1999.

[9] E. Bolshakova, N. Loukachevitch, M. Nokel. Topic Models Can Improve Domain Term Extraction. In ECIR Proceedings, Lecture Notes in Computer Science, Vol. 7814, pp. 684–687, 2013.

[10] M. Nokel, N. Loukachevitch. Application of Topic Models to the Task of Single-Word Term Extraction. In RCDL'2013 Proceedings, pp. 52–60, 2013.

[11] Q. He, K. Chang, E. Lim, A. Banerjee. Keep It Smile with Time: A Reexamination of Probabilistic Topic Detection Models. In the Proceedings of IEEE Transaction Pattern Analysis and Machine Intelligence, vol. 32, issue 10, pp. 1795–1808, 2010.

[12] H. Wallach. Topic Modeling: beyond bag-of-words. In the Proceedings of the 23rd International Conference on Machine Learning, pp. 977–984, 2006.

[13] T. Griffiths, M. Steyvers, and J. Tenenbaum. Topics in semantic representation. Psychological Review, 144, 2, pp. 211–244, 2007.

[14] X. Wang, A. McCallum, and X. Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In the Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, pp. 697–702, 2007.

[15] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic

coherence. In the Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100-108, 2010.

[16] W. Hu, N. Shimizu, H. Sheng. Modeling chinese documents with topical word-character models. In the Proceedings of the 22nd International Conference on Computational Linguistics, pp. 345–352, 2008.

[17] M. Johnson. PCFGs, topic models, adaptor grammars and learning topical collocations and the structure of proper names. In the Proceedings of the 48th Annual Meeting of the ACL, pp. 1148–1157, 2010.

[18] J. H. Lau, T. Baldwin, and D. Newman. On Collocations and Topic Models. In ACM Transactions on Speech and Language Processing, 10 (3), pp. 1–14, 2013.

[19] D. Andrzejewski, X. Zhu, and M. Craven. Incorporating domain knowledge into topic modeling via Dirichlet Forest priors. In the Proceedings of the 26th Annual International Conference on Machine Learning, pp. 25–32, 2009.

[20] B. Liu. Sentiment Analysis and Opinion Mining. Syntheses Lectures on Human Language Technologies. Morgan & Claypool Publishers. 2012

[21] Z. Zhai, B. Liu, H. Xu, and P. Jia. Grouping Product Features Using Semi-Supervised Learning with Soft-Constraints. In the Proceedings of the 23rd International Conference on Computational Linguistics, pp. 1272–1280, 2010.

[22] A. Daud, J. Li, and F. Muhammad. Knowledge discovery through directed probabilistic topic models: a survey. Frontiers of Computer Science in China, 4(2), pp. 280–301, 2010.

[23] J. Chang, J. Boyd-Grabber, C. Wang, S. Gerrich, and D. Blei. Reading tea leaves: How human interpret topic models. In the Proceedings of the 24th Annual Conference on Neural Information Processing Systems, pp. 288–296, 2009.

[24] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In the Proceedings of the International Conference on Uncertainty in Artificial Intelligence, 2009.

[25] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In the Proceedings of EMNLP'2011, pp. 262–272, 2011.

[26] K. Stevens, P. Kegelmeyer, D. Andrzejewski, D. Butter. Exploring topic coherence over many models and many topics. In the Proceedings of EMNLP-CoNLL'12, pp. 952–961, 2012.

[27] D. Andrzejewski and D. Buttier. Latent topic feedback for information retrieval. In the Proceedings of tthe 17th ACM SIGKDD International Conference on Knowledge discovery and data mining, pp. 600–608, 2011.

[28] K. Church and P. Hanks. Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics, vol. 16, pp. 22–29, 1990.

[29] W. Zhang, T. Yoshida, T. Ho, and X. Tang. Augmented Mutual Information for Multi-Word Term Extraction. International Journal of Innovative Computing, Information and Control, 8(2), pp. 543–554, 2008.

[30] B. Daille. Combined Approach for Terminology Extraction: Lexical Statistics and Linguistic Filtering. PhD Dissertation, University of Paris, 1995.

[31] G. Bouma. Normalized Pointwize Mutual Information. In the Proceedings of the Biennal GSCL Conference, pp. 31–40, 2009.

[32] F. Smadja, K. McKeown, and V. Hatzivassiloglou. Translating Collocations for Bilingual Lexicons: A Statistical Approach. Computational Linguistics, 22(1), pp. 1–38, 1996.

[33] M. Kitamura and Y. Matsumoto. Automatic Extraction of Word Sequence Correspondences in Parallel Corpora. In the Proceedings of the 4th Annual Workshop on Very Large Corpora, pp. 79–87, 1996.

[34] J. G. P. Lopes and J. F. Silva. A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units. In the Proceedings of the 6th Meeting on the Mathematics of Language, pp. 369–381, 1999.

[35] T. Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics, 19(1), 1993.

[36] Y. Park, R. Bird, and B. Boguraev. Automatic Glossary Extraction: Beyond Terminology Identification. In the Proceedings of the 19th International Conference on Computational Linguistics, 2002.

[37] K. Vorontsov and A. Potapenko. EM-like algorithms for probabilistic topic modeling. Machine Learning and Data Analysis, vol. 1(6), pp. 657–686, 2013.

**Topic models: taking into account similarity between unigrams and bigrams**

Michael Nokel

The paper presents the results of experimental study of integrating word similarity and bigram collocations into topic models. First of all, we analyze a variety of word association measures in order to integrate top-ranked bigrams into topic models. Then we propose a modification of the original algorithm PLSA, which takes into account similar unigrams and bigrams that start with the same beginning. And at the end we present a novel unsupervised iterative algorithm demonstrating how topics can choose the most relevant bigrams. As a target text collection we took articles from various Russian electronic banking magazines. The experiments demonstrate significant improvement of topic models quality for both collections.