# MemoPIM: управление личной информацией и знаниями с использованием семантических технологий

© А. А. Бездушный © А. Н. Бездушный © В. А. Серебряков
ВЦ им. А.А. Дородницына РАН
andrey.bezdushny@gmail.com anb@ccas.ru serebr@ccas.ru

**Аннотация**

В ходе повседневной деятельности, человек сталкивается со все большими объемами информации, значительная часть которой хранится в цифровом формате. В настоящей работе, на основе библиографических исследований, рассматриваются требования к системе MemoPIM, совмещающей функциональные возможности управления личной информацией и знаниями. Рассматриваются вопросы ведения контекста работы, автоматизированной категоризации данных и выявления связей между ресурсами. Знания предлагается представлять в RDF-формате, используя OWL-онтологии для описания их структуры.

## 1 Введение

В ходе своей разноплановой деятельности человек регулярно сталкивается с необходимостью изучения новой информации, значительная часть которой хранится в цифровом формате и распределена по различным источникам. В процессе систематизации и изучения информации вырабатывается собственное понимание сведений, формируются связи между новой информацией и изученной ранее. В ходе этого процесса формируются личные знания, которые человек в дальнейшем использует при решении различных задач профессиональной или повседневной деятельности. Со временем часть знаний неизбежно забывается, что порождает необходимость переноса знаний в цифровой формат и организации эффективной работы с ними.

Вопросы организации личных знаний и ведения работы с ними рассматриваются в рамках задач управления личной информацией (Personal Information Management) и управления личными знаниями (Personal Knowledge Management). *Управление личной информацией* изучает вопросы работы с информационными ресурсами, такими как электронные и бумажные документы, веб-страницы, почтовые сообщения. В задаче *управления личными знаниями* рассматриваются вопросы формирования знаний, сопровождения и оперативной работы с ними. Знания накапливаются человеком в ходе его повседневной деятельности. Они могут быть связаны с информационными ресурсами или произвольными понятиями, фактами или идеями, важными для человека. На настоящее время большинство существующих систем подобного назначения ориентированы либо только на управление информацией, либо только на управление знаниями, хотя, как правило, между знаниями и информацией существует тесная взаимосвязь. В настоящей работе рассматривается система MemoPIM, объединяющая функции управления информацией и знаниями. Система MemoPIM расширяет прототип системы, представленной в работе [1].

## 2 Системы управления личной информацией и знаниями

В данном разделе рассматриваются используемые в настоящее время подходы к управлению личной информацией и управлению личными знаниями.

Популярным направлением исследований, в области *управления личной информацией*, является Semantic Desktop [2] – подход к организации данных на персональном компьютере. В соответствии с этим подходом, все виды информации, хранимые в цифровом формате, – файл, email-сообщение или событие календаря, рассматриваются как RDF-ресурсы с уникальным идентификатором. Совокупность этих ресурсов образует *информационное пространство* пользователя. Системы управления личной информацией, осуществляют сбор сведений из различных источников, объединяют их и предоставляют интерфейсы работы с ним. Данный подход применяется в работах [2- 7]. Далее, в самом общем объеме, перечислены основные требования, которые предъявляются в этих работах к системам управления личной информацией, включая наличие следующих механизмов:

- предоставления OWL-онтологии, определяющей структуру ресурсов информационного пространства, а также возможности ее расширения пользователем;
- загрузки в систему и представления в RDF-формате информации, сформированной

**Труды XVII Международной конференции DAMDID/RCDL'2015 «Аналитика и управление данными в областях с интенсивным использованием данных», Обнинск, 13-16 октября 2015**

сторонними приложениями (почтовые сообщения, календари, контакты);
- проставления связей между только загруженными и существовавшими ранее ресурсами (например, между человеком из списка контактов и отправителем электронного письма);
- предоставления интерфейсов, навигации и поиска по информационному пространству;
- поддержки совместной работы с ресурсами (комментирования, обсуждений).

*Системы управления личными знаниями*, организуют работу со знаниями, которые человек накапливает в ходе своей повседневной деятельности. Для автоматизации работы с этими знаниями, они сначала должны быть переведены в цифровой формат. В результате этого процесса формируются так называемые *компьютеризированные знания* [8], которые являются лишь «тенью» настоящих знаний, хранящихся в голове человека. Полноценный перенос всех знаний в цифровой формат является трудоемким процессом и, в большинстве случаев, не оправдывающим ресурсозатраты на его реализацию. По этой причине в работе [9] предлагается переносить в цифровой формат только ту часть знаний, которая необходима человеку для извлечения нужной информации из памяти в нужный момент времени. Компьютеризированные таким образом знания – это набор взаимосвязанных ключей (подсказок), увидев которые, человек вспомнит, с какой целью они создавались, и какие реальные знания с ними ассоциированы. В качестве таких ключей могут выступать электронные документы, заметки, картинки и другие виды информации. В связи с этим, важной функцией системы управления знаниями, является помощь пользователю в запоминании ключей и последующем их извлечении.

Можно выделить два подхода к представлению знаний и работе с ними – *графический и гипертекстовый*. В графических способах записи, знания обычно представляются в виде графа, вершинами которого являются ресурсы базы знаний, а ребрами – связи между ними. В гипертекстовом подходе, каждый ресурс базы знаний имеет собственное детальное представление, из которого выводятся все его атрибуты и связи с другими ресурсами. Переход между ресурсами осуществляется с помощью ссылок.

Среди наиболее распространенных способов графической записи знаний можно выделить карты памяти (Mind Map [10]) и карты понятий (Concept Map [11]). В работе [12] в качестве интерфейса для работы со знаниями предлагается использовать тематические карты (Topic Maps – ISO-стандарт графического представления знаний). В работе [13] описывается система iMapping, предоставляющая интерфейс для графического описания знаний. Основной особенностью iMapping является использование масштабируемых пользовательских интерфейсов для представления знаний. Все ресурсы базы знаний располагаются на неограниченной плоскости, а их содержимое становится доступно по мере увеличения масштаба интерфейса. Система Docear [14] позволяет использовать карты памяти при ведении и анализе библиографической информации.

В работе [9] используется гипертекстовый подход к работе со знаниями. Авторы описывают модель Conceptual data structures (CDS), в основе которой лежит представление знаний в виде троек «объект-предикат-субъект». В этой работе также описывается система «Hypertext Knowledge Workbench», предоставляющая гипертекстовый интерфейс для создания и управления знаниями. В работах [15] и [16] рассматриваются вопросы проектирования модели организации знаний, согласующейся с устройством памяти человека, а также вопросы применения механизма распространения активации для ведения контекста работы пользователя. Другим способом гипертекстового представления знаний является подход семантических Wiki [17-18]. Семантические Wiki расширяют стандартные Wiki, позволяя задавать семантику связей между страницами.

Большинство рассмотренных систем поддерживают либо только управление знаниями, либо только управление информацией, хотя личные знания человека часто связаны с информационными ресурсами, хранимыми на его компьютере или в сети. Также мало внимания уделяется вопросам анализа информации и знаний, автоматизации действий, выполняемых пользователем. На основании проведенного анализа были определены ключевые функции системы MemoPIM, объединяющей управление личной информацией и знаниями.

## 3 Архитектура системы MemoPIM

Предлагаемое автором решение развивает идеи Semantic Desktop, дополняя функции системы управления информацией механизмами, обеспечивающими работы со знаниями. Помимо информации, изначально хранящейся в цифровом формате (электронные документы, email-сообщения), система поддерживает работу с личными знаниями человека. Пополняя сведения, выгруженные из внешних источников дополнительными метаданными, и проставляя связи между ресурсами, пользователь формирует цифровое представление личных знаний. На рисунке 1 приведена архитектура системы.

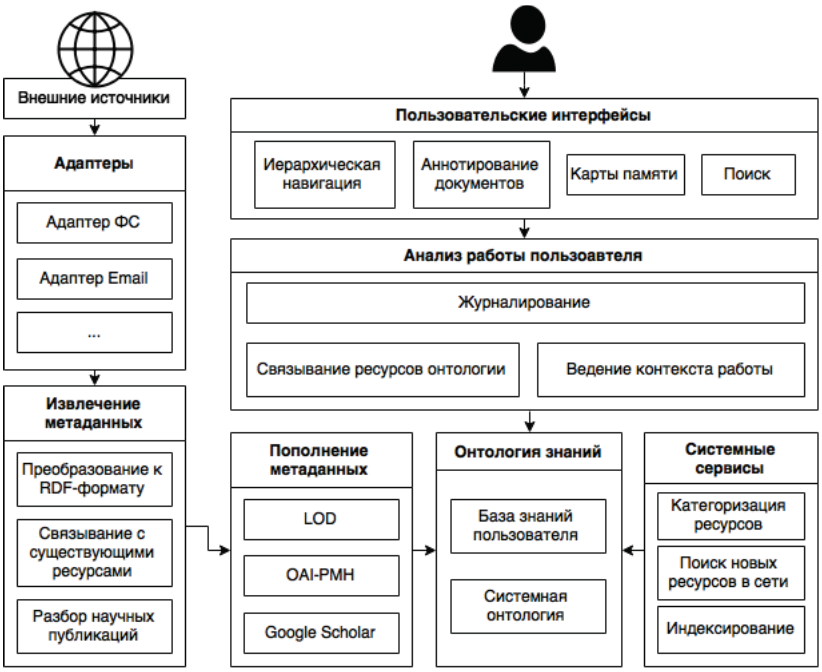


Рис. 1. Архитектура системы MemoPIM

Рассмотрим основные модули, представленные на схеме. В основе системы лежит база знаний пользователя. Структура базы знаний задается системной онтологией и ее пользовательскими расширениями. Более подробно системная онтология будет рассмотрена в дальнейших разделах. База знаний пополняется либо вручную пользователем, с помощью интерфейсов системы, либо автоматизировано, из внешних источников.

Внешними источниками для пополнения базы знаний пользователя являются файловая система (ФС), почтовый сервер и другие. В фоновом режиме адаптеры производят синхронизацию ресурсов базы знаний с внешними источниками.

После выгрузки информации из внешних источников она приводится к RDF-формату, в соответствии с системной онтологией. Далее, загруженные данные связываются с существующими ресурсами, например, отправитель письма ассоциируется с ресурсом, представляющим соответствующую персону в базе знаний. После этого, для некоторых типов ресурсов из их

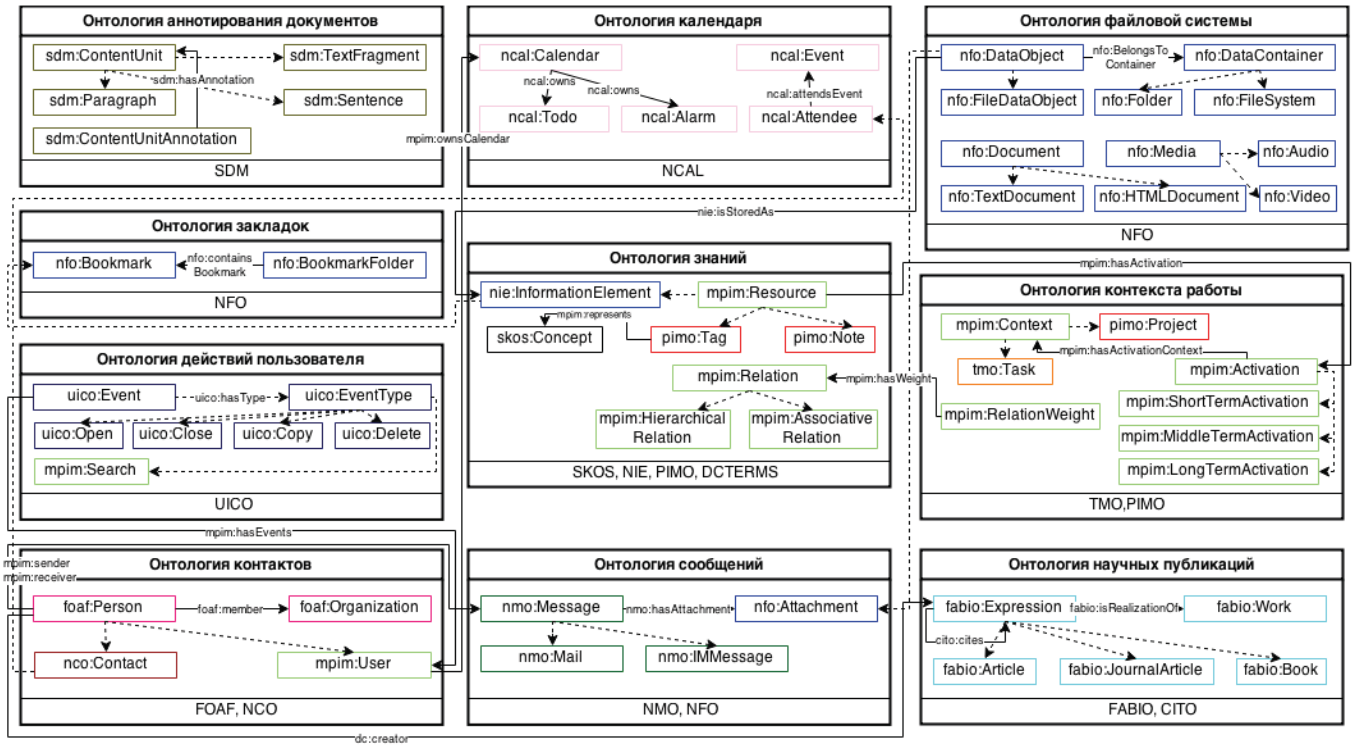


Рис. 2. Системная онтология

содержимого, выделяются дополнительные метаданные. Например, из текста научной публикации выделяются название, авторы, список литературы. Более подробно процесс выделения указанных метаданных из текстов публикаций рассмотрен в работе [19].

Следующим этапом является пополнение созданного ресурса дополнительными метаданными из внешних источников. Например, для научных публикаций, на основании названия публикации, осуществляется поиск полной библиографической информации. В случае успеха найденная информация переводится в RDF-формат и сохраняется в базе знаний пользователя. Более детально модуль, реализующий такое пополнение, рассматривается в работе [19].

Для работы с базой знаний системой предоставляются иерархические и графические пользовательские интерфейсы. В режиме иерархического представления пользователю выводится дерево ресурсов, соединенных иерархическими связями. Просмотр полной информации о ресурсе доступен с помощью его детального представления. В графическом интерфейсе работы, пользователь может создавать карты памяти, узлами которых являются ресурсы базы знаний. Интерфейс аннотирования позволяет размечать и комментировать текстовое содержимое ресурсов базы знаний. Интерфейс поиска позволяет осуществлять полнотекстовый поиск по базе знаний, а также поиск по ресурсам с учетом их структуры.

По мере работы пользователя системой ведется журналирование выполняемых им действий. Записи из журналов в дальнейшем используются для ведения контекста пользователя и поиска новых связей между ресурсами. Более подробно эти модули рассмотрены в следующих разделах.

Системные сервисы производят фоновый анализ базы знаний пользователя. Сервис автоматизированной категоризации отвечает за предоставление пользователю рекомендации по присвоению ресурсам тегов. Сервис поиска новых ресурсов в сети отвечает за предоставление пользователю потенциально интересных ему ресурсов, на основании анализа тех, которые хранятся в базе знаний. Например, на основании имеющихся в базе знаний публикаций, в соответствии с рекомендациями Google Scholar, системой предлагаются другие публикации, потенциально интересные пользователю. Сервис индексирования отвечает за ведение индексов, необходимых для эффективного поиска информации в базе знаний.

## 4 Системная онтология

Системная онтология описывает основные понятия и классы, используемые различными модулями системы. Онтология состоит из десяти частей, разделенных по предметным областям. Используются классы из следующих онтологий: Simple Knowledge Organization System (SKOS)[1], Dublin Core Terms (DCTERMS)[2], Friend of a Friend (FOAF)[3], NEPOMUK Information Element Ontology (NIE)[4], NEPOMUK File Ontology (NFO)[5], NEPOMUK Contact Ontology (NCO)[6], NEPOMUK Message Ontology (NMO)[7], NEPOMUK Calendar Ontology (NCAL)[8], Personal Information Model Ontology (PIMO)[9], Task Model Ontology (TMO)[10], FRBR-aligned Bibliographic Ontology (FaBiO)[11], Citation Typing Ontology (CiTO)[12], User Interaction Context Ontology (UICO) [20]. На рисунке 2 представлены ключевые классы системной онтологии и связи между ними. Пунктирными линиями обозначено наследование, сплошными – связи между классами. В начале названий всех классов стоит пространство имен, соответствующее одной из перечисленных выше онтологий. Классы, созданные в рамках данной работы, отмечены пространством имен “mpim”.

В основе системных онтологий лежит *онтология знаний*, в которой определяется класс Resource, являющийся родительским для всех остальных классов. Для структурирования знаний вводятся ассоциативные/иерархические связи между ресурсами и теги. Выделение иерархических связей необходимо для поддержания иерархического режима представления знаний.

## 5 Карты памяти

Карты памяти (Mind Maps [10]) это диаграммы, обычно используемые для представления идей, мыслей, задач или других видов знаний. В центре карты памяти располагается узел, представляющий основную идею. По радиусу от него распространяются ветви, образуя древовидную структуру, в вершинах которой находятся различные идеи, слова или понятия. Карты памяти часто используются для представления знаний при обучении, для классификации идей или при проведении «мозгового штурма».

Можно выделить несколько направлений исследований, связанных с картами памяти. В работе [14] на основании содержимого карт памяти, пользователю предоставляются рекомендации по потенциально интересным ему научным публикациям. В работе [20] рассматриваются вопросы использования онтологий при работе с картами памяти.

---

[1] http://www.w3.org/TR/skos-reference/
[2] http://dublincore.org/documents/dcmi-terms/
[3] http://xmlns.com/foaf/spec/
[4] http://www.semanticdesktop.org/ontologies/2007/01/19/nie
[5] http://www.semanticdesktop.org/ontologies/2007/03/22/nfo
[6] http://www.semanticdesktop.org/ontologies/2007/03/22/nco
[7] http://www.semanticdesktop.org/ontologies/2007/03/22/nmo
[8] http://www.semanticdesktop.org/ontologies/2007/04/02/ncal
[9] http://www.semanticdesktop.org/ontologies/2007/11/01/pimo
[10] http://www.semanticdesktop.org/ontologies/2008/05/20/tmo
[11] http://purl.org/spar/fabio
[12] http://purl.org/spar/cito

В рамках системы MemoPIM карты памяти применяются для изучения, анализа информации, регистрации новых знаний. Карты памяти с одной стороны являются визуально удобной для восприятия человека формой записи знаний, а с другой – легко могут быть преобразованы в машинно-читаемую форму – RDF-данные. В работе [20] к недостаткам карт памяти относят отсутствие формальной семантики, что приводит к двусмысленной интерпретации знаний, усложняет их совместное использование. В разрабатываемой системе узлами карты памяти являются ресурсы базы знаний, а ребрами – связи между ними. Таким образом, знания, формализованные в карте памяти, описываются системной онтологией и ее пользовательскими расширениями.

## 6 Ведение контекста работы пользователя

Люди, как правило, испытывают трудности при одновременной работе в разных сферах предметной деятельности, что связано с необходимостью регулярных переключений между текущими задачами. Чтобы возобновить работу в другой сфере, человеку необходимо вернуться в состояние, в котором он был на момент переключения, вспомнить, о чем он думал, и в каком направлении планировал развивать мысли. Поэтому, важной задачей системы управления личной информацией, является ведение контекста работы пользователя. По мере выполнения действий, система должна отслеживать ресурсы, с которыми работает пользователь, запоминать их и упрощать дальнейший доступ к ним.

Для задачи ведения контекста можно руководствоваться принципами работы с*емантической (долгосрочной) памяти человека.* Семантическую память можно рассматривать в виде графа, узлами которого являются понятия, а ребрами – связи между ними (Рис. 3). В процессе размышления о некотором понятии, человек вспоминает также и о других понятиях, связанных с ним в этом графе.


Рис. 3. Граф понятий семантической памяти

Общепринятым принципом извлечения информации из долгосрочной памяти считается механизм *распространения активации* (Spreading Activation [21]). В соответствии с ним, получение информации из памяти начинается с внешнего стимула, провоцирующего активацию одного из понятий, находящихся в долгосрочной памяти. Например, если человек видит пожарную машину, то происходит активация соответствующего понятия в памяти. Далее, активация распространяется по ребрам графа понятий, при этом большее значение активации получают элементы, связанные ребрами с наибольшим весом, т.е. элементы более близкие по смыслу. Например, при активации термина «пожарная машина», для дальнейшей работы становятся доступны такие понятия как «красный», «огонь», «дом».

В рамках проектируемой системы, механизм распространения активации может использоваться для ведения текущего контекста работы пользователя. В таком случае, внешними стимулами являются действия пользователя (например, просмотр редактирование или поиск ресурсов), а в роли графа понятий выступают ресурсы базы знаний и имеющиеся между ними связи. По мере работы с системой, формируется группа ресурсов, обладающих наибольшим значением активации. Такие ресурсы представляют собой *текущий контекст работы пользователя* – наиболее актуальные для него сведения на конкретный момент времени.

### Алгоритм распространения активации

Стандартный алгоритм распространения активации описан в работе [22]. В данном разделе рассматриваются основные факторы, оказывающие влияние на результаты работа алгоритма.

При применении алгоритма распространения активации, важно, чтобы между ресурсами, которые часто используются совместно, существовала связь в графе. Частично данный вопрос решается использованием тегов, однако этого может быть недостаточно. За формирование дополнительных связей отвечает отдельный модуль системы, который будет рассмотрен в разделе 8.

Другим дополнением, которое может улучшить расчет активации, является ведение нескольких уровней активации одновременно. Данный подход применяется в работе [16]. В ней авторы предлагают различать три вида активации – *краткосрочную* (STA), *среднесрочную* (MTA) и *долгосрочную* (LTA). При превышении STA, некоторого порогового значения, увеличивается значение MTA. Аналогично увеличение MTA ведет к росту LTA. Каждому виду активации соответствует свое значение коэффициента затухания $d_s$. Значения среднесрочной и долгосрочной активации учитываются при расчете краткосрочной – ресурсы, обладающие среднесрочной или долгосрочной

активацией, оказывают большее влияние на распространение краткосрочной активации.

**Онтология контекста работы и онтология действий пользователя**

Онтология контекста работы, описывает основные классы, необходимые для применения механизма распространения активации. Класс Context представляет контекст работы пользователя, контекстом может быть проект (Project) или задача (Task). В рамках некоторого контекста, каждому ресурсу может быть проставлено значение *активации* (Activation). Активация делится на краткосрочную (ShortTermActivation), среднесрочную (MiddleTermActivation) и долгосрочную (LongTermActivation). Класс RelationWeight позволяет задать вес связи, используемый при расчете активации.

Онтология действий пользователя описывает различные типы пользовательских действий, которые инициируют процесс распространения активации. В ней используются классы онтологии User Interaction Context Ontology (UICO), спроектированной в работе [23].

## 7 Автоматическая категоризация ресурсов

Традиционно выделяют два основных подхода к категоризации ресурсов – с помощью иерархических структур и с помощью тегов. Разрабатываемая система объединяет оба вида представлений, позволяя создавать иерархические связи между тегами.

По мере наполнения базы знаний пользователя ресурсами и соответствующими им тегами, появляется возможность проставлять теги к новым ресурсам автоматически. Эту задачу можно рассматривать как задачу рекомендации возможных тегов для ресурсов. На настоящее время такие рекомендации являются весьма популярным направлением исследований. Обычно рекомендации тегов используются в системах, предоставляющих возможность ведения фолксономий – пользовательских таксономий, например CiteULike[13], Bibsonomy[14], Delicious[15].

В данной работе, для формирования рекомендаций предлагается анализировать связь между словами, входящих в содержимое документов, и тегами, присвоенными этим документам. Схожие подходы рассматриваются в работах [24] и [25]. Поскольку в большинстве случаев теги и ресурсы в базах знаний различных пользователей отличаются, в качестве источника для рекомендаций, автором было решено использовать только теги и ресурсы из базы знаний самого пользователя.

[13] http://www.citeulike.org/
[14] http://www.bibsonomy.org/
[15] https://delicious.com/

Перед формированием рекомендаций из каждого слова выделяется его основа (производится стемминг слов), а также отбрасываются стоп-слова. Далее, для всех оставшихся слов, высчитывается связь между словом и тегом, по формуле:

$$WT(w,t) = \sum_{r \in Res(w,t)} (\alpha * \delta_t(w,r) + \beta * \delta_c(w,r) + \gamma * \delta_a(w,r)) * Tfidf(w,r)$$

Значение этой функции определяет, насколько слово характеризует ресурс.

В представленной выше формуле: r – ресурс, $\mathrm{Res(w,t)}$ – множество ресурсов, отмеченных тегом $t$ и содержащих слово $\mathrm{w}$;

$$\delta_t(w,r) = \begin{cases} 1, w \in T(r) \\ 0, \text{иначе} \end{cases}, \quad \delta_c(w,r) = \begin{cases} 1, w \in C(r) \\ 0, \text{иначе} \end{cases};$$

$$\delta_a(w,r) = \begin{cases} 1, w \in A(r) \\ 0, \text{иначе} \end{cases};$$

$T(r)$, $C(r)$, $A(r)$ – множества слов входящих в название, содержание и аннотацию ресурса соответственно. При апробации алгоритма в качестве аннотации использовался текст, содержащийся на первой странице документа.

$\alpha, \beta, \gamma$ – коэффициенты определяющие вклад каждого из слагаемых.

$Tfidf$ – мера информативности слова, которая вычисляется по формуле:

$$Tfidf(w,r) = \frac{n_w}{N} * log \frac{|R|}{|R_w|}.$$

В последней формуле $\mathrm{n_w}$ – количество вхождений слова $\mathrm{w}$ в текст; $\mathrm{N}$ – общее количество слов в тексте; $|\mathrm{R}|$ - общее количество ресурсов; $|\mathrm{R_w}|$ - количество ресурсов содержащих слово $\mathrm{w}$.

При подборе тегов для рекомендации, рейтинг тега считает по формуле:

$$RT(r,t) = \sum_{w \in r} \mathrm{Tfidf(w,r)}^x * \mathrm{WT(w,t)}.$$

Здесь $x$ – коэффициент, определяющий влияние информативности слова на его вклад в итоговое значение.

Таким образом, влияние слова на рейтинг тега складывается из двух факторов:

1. насколько это слово характеризует ресурс;
2. насколько часто слово встречается в ресурсах, помеченных тегом, а также насколько оно характеризует эти ресурсы.

Для оценки эффективности представленного алгоритма, была сформирована тестовая выборка из 381 научной публикации на русском и английском языках. Публикации были разделены на 23 категории.

В ходе тестирования был проведен ряд экспериментов, в каждом из которых для каждой публикации формировался список рекомендованных тегов. Эксперименты отличались между собой –

значениями коэффициентов $x, \alpha, \beta, \gamma$. В таблице 1 приведены результаты для некоторых наборов параметров. По горизонтали указаны порядковые номера, которые получили правильные рекомендации в сформированном списке, по вертикали наборы параметров, при которых эти результаты были получены.

Из результатов экспериментов видно, что более точные результаты получаются для конфигураций в которых словам, входящим в название или аннотацию документа, присваивается дополнительный вес. Лучший результат был получен при наборе параметров (3,1,1,1) – в 83% случаев правильный тег оказывался в числе первых пяти сформированных рекомендаций и в 55% первая рекомендация совпадала с правильной. На рисунке 4 приведен график распределения порядковых номеров правильных рекомендаций в тестируемом наборе.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| (1,1,0,0) | 178 | 43 | 39 | 19 | 19 |
| (1,1,1,0) | 186 | 42 | 42 | 18 | 10 |
| (1,1,1,1) | 196 | 45 | 33 | 22 | 10 |
| (3,1,0,0) | 208 | 38 | 28 | 11 | 22 |
| (3,1,1,0) | 211 | 43 | 29 | 13 | 14 |
| (3,1,1,1) | 210 | 49 | 32 | 12 | 14 |
| (5,1,0,0) | 198 | 52 | 17 | 18 | 16 |
| (5,1,1,1) | 200 | 54 | 17 | 18 | 16 |
| (5,1,1,1) | 201 | 53 | 29 | 15 | 19 |
| (7,1,0,0) | 187 | 62 | 27 | 16 | 15 |
| (7,1,1,0) | 189 | 64 | 28 | 15 | 17 |
| (7,1,1,1) | 194 | 54 | 33 | 18 | 17 |

Таблица 1 Результаты экспериментов

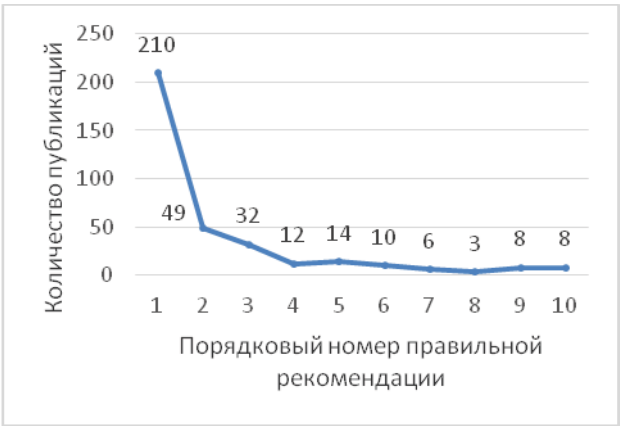


Рис. 4. Распределение порядковых номеров правильных рекомендаций

## 8 Связывание ресурсов базы знаний

В рамках данного модуля проводится анализ базы знаний с целью построения дополнительных связей между ресурсами. Для поиска связанных между собой ресурсов обычно применяются различные меры схожести.

В системе MemoPIM для расчета схожести ресурсов базы знаний, проводится анализ не только самого RDF-графа, но и того, как пользователь строит свою работу с ним. Длина ребра $(a, r, b)$ считается по формуле:

$$d(a,b) = 1 - C(a,b) * \alpha - RFITF(r,a,b) * \beta.$$

Здесь $C(a,b)$ – частота совместного использования ресурсов $a$ и $b$ – чем чаще в журналах работы пользователя два ресурса используются в близкий промежуток времени, тем меньше расстояние между этими ресурсам в RDF-графе. Второй величиной, определяющей близость ресурсов, является функция Relation Frequency - Inverse Triple Frequency ($RFITF$), схожая с функцией $PFITF$, предлагаемой в работе [26]. Коэффициенты $\alpha, \beta$ определяют влияние описанных выше слагаемых на итоговое значение меры схожести.

$$RFITF(r,a,b) = \frac{|T(r,a,b)|}{|T(r,a)|} * \log\left(\frac{|T(r)|}{|T(r,b)|}\right) * \log\left(\frac{|T|}{|T(r)|}\right).$$

Здесь:

$|T(r,a,b)|$ – мощность связи ресурса $a$ с ресурсом $b$ по средствам предиката $r$. Мощность может задаваться для разных предикатов по-разному. Например, для предиката, связывающего ключевое слово с текстовым документом, мощность – это количество вхождений ключевого слова в документе;

$|T(r,a)|$ – количество троек $(a,r,?)$ и $(?,r,a)$;

$|T(r)|$ – количество троек $(?,r,?)$;

$|T|$ – общее количество троек в репозитории пользователя.

Мера схожести между ресурсами определяется по формуле:

$$S(a,b) = \sum_{p \in P} \gamma^{l(p)} * \frac{1}{d(p)}, \quad d(p) = \sum_{a,b \in p} d(a,b),$$

где $\gamma$ – коэффициент затухания; P – множество существующих путей в графе между вершинами A и B; l(p) – число ребер в пути; d(p) – длина пути p.

## 9 Интерфейсы управления знаниями

Системой поддерживаются два вида представления знаний – *иерархическое (рис.5) и графическое (рис.6).* Информация, которая выгружается из внешних источников (файловой системы, почтового сервера), обычно сгруппирована иерархически. При импорте создаются соответствующие RDF-ресурсы, которые

соединяются иерархическими связями. В режиме иерархического представления знаний рабочая область делится на три части. Слева выводится дерево ресурсов, соединенных иерархическими связями. При выделении ресурса, в центральной части рабочей области отображается его детальное представление, содержащее все атрибуты данного ресурса и ссылки на другие ресурсы, связанные с ним. В правой части выводятся ресурсы, представляющие текущий контекст работы пользователя.

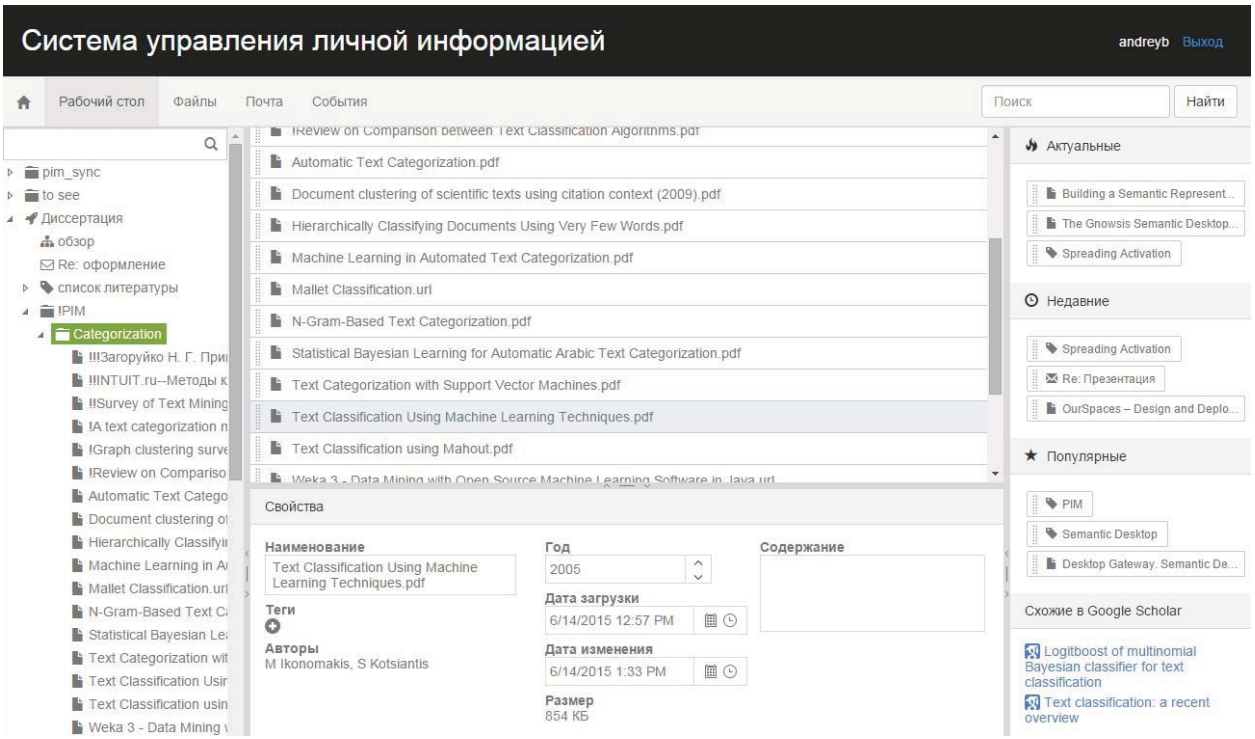


Рис. 5. Иерархический интерфейс работы со знаниями


Рис. 6. Представление знаний в виде карт памяти

Другим режимом работы со знаниями, который предоставляется системой, являются карты памяти. Пользователь может создать диаграмму, узлами которой являются ресурсы базы знаний, а ребрами - связи между ними.

Помимо описанных интерфейсов создания и редактирования знаний, в дальнейшем планируется реализовать интерфейсы аннотирования текстовых документов, просмотра журналов работы с ресурсами.

## 10 Заключение

В данной работе, на основе библиографического исследования, представлены основные требования, предъявляемые к системам управления личной информацией и знаниями. На основании их анализа спроектирована система MemoPIM, расширяющая прототип системы, реализованный в работе [1]. Знания в системе хранятся в RDF-формате. Для формализации структуры RDF-данных, спроектирована OWL-онтология, описывающая основные понятия и классы, используемые различными модулями системы. Для работы со знаниями предоставляются графический и иерархический интерфейсы. К функциям системы также относятся – ведение контекста работы, автоматизированная категоризация данных, связывание ресурсов. По каждой функции проведен обзор существующих решений и способов реализации.

В рамках продолжения работы планируется развитие и апробация прототипа системы MemoPIM. К основным направлениям этой деятельности можно отнести предоставление пользователю возможностей расширения системной онтологии, поддержку интерфейсов создания новых классов и связей между ними. Кроме того, в перспективе пользователю может предоставляться возможность настройки интерфейсов представления ресурсов.

## Литература


[1] Beel J., Langer S., Genzmehr M., Gipp B. Utilizing Mind-Maps for Information Retrieval and User Modelling // User Modeling, Adaptation, and Personalization, Lecture Notes in Computer Science, Vol. 8538, 2014. pp. 301-313.

[2] 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. pp. 731-740.

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

[4] Collins A. M., Loftus E. F. A spreading-activation theory of semantic processing // Psychological review, Vol. 82, No. 6, 1975. P. 407.

[5] Devaurs D., Rath A. S., Lindstaedt S. N. Exploiting the user interaction context for automatic task detection // Applied Artificial Intelligence, Vol. 26, No. 1-2, 2012. pp. 58-80.

[6] Dong X. L., Halevy A. A platform for personal information management and integration // Proceedings of Conference on Innovative Data Systems Research (CIDR). Asilomar, CA, USA. January 4-7, 2005. pp. 119-130.

[7] F. C. Application of spreading activation techniques in information retrieval // Artificial Intelligence Review, Vol. 11, No. 6, 1997. pp. 453-482.

[8] 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. pp. 235-238.

[9] H. M. The heart of the problem: Knowledge management and knowledge transfer // Proceedings of ENABLE'99. Espoo, Finland. 2-5 June, 1999. Vol. 99. pp. 8-11.

[10] Haller H., Abecker A. iMapping – A Zooming User Interface Approach for Personal and Semantic Knowledge Management // Proceedings of the 21st ACM conference on Hypertext and hypermedia. Toronto, ON, Canada. June 13 - 16, 2010. pp. 119-128.

[11] 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 Conference on Innovative Data Systems Research (CIDR). Asilomar, CA, USA. January 4-7, 2005. pp. 13-26.

[12] Katifori A., Vassilakis C., Dix A. Ontologies and the Brain: Using Spreading Activation through Ontologies to Support Personal Interaction // Cognitive Systems Research, Vol. 11, No. 1, 2010. pp. 25-41.

[13] Křemen P., Mička P., Blaško M. Ontology-driven mindmapping // Proceedings of the 8th International Conference on Semantic Systems. Graz, Austria. September 05 - 07, 2012. pp. 125-132.

[14] Landia N. Content-awareness and graph-based ranking for tag recommendation in folksonomies : дис. – University of Warwick 2.

[15] Lipczak M., Hu Y., Kollet Y. et al. Tag sources for recommendation in collaborative tagging systems // Proceedings of the ECML-PKDD discovery challenge workshop. Bled, Slovenia. 2009. pp. 157-172.

[16] Novak J. D., Cañas A. J. The theory underlying concept maps and how to construct them // URL: http://cmap.ihmc.us/docs/theory-of-concept-maps (дата обращения 22.06.2015)

[17] Oren E., Völkel M., Breslin J. G., Decker S. Semantic wikis for personal knowledge management // Proceedings of the 17th International Conference on Database and Expert Systems Applications (DEXA 2006), Lecture Notes in Computer Science. Krakow, Poland. 2006. Vol. 4080. pp. 509-518.

[18] Pirró G. Reword: Semantic relatedness in the web of data // Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence. Toronto, Ontario, Canada. 2012. pp. 129-135.

[19] Richter J., Völkel M., Haller H. DeepaMehta - A Semantic Desktop // Proceedings of the Semantic Desktop Workshop at ISWC. Galway, Ireland. November 6, 2005.

[20] Sauermann L. G.G.A..K.M.E.A. Semantic desktop 2.0: The gnowsis experience // Proceedings of the 5th International Semantic Web Conference (ISWC). 2006. pp. 887-900.

[21] Schaffert S. IkeWiki. A Semantic Wiki for Collaborative Knowledge Management // Proceedings of the Enabling Technologies: Infrastructure for Collaborative Enterprises. Manchester, UK. 2006. pp. 388-396.

[22] T. B. Use both sides of your brain: new mind-mapping techniques. New York : Plume Books, 1991.

[23] Trullemans S., Signer B. Towards a Conceptual Framework and Metamodel for Context-aware Personal Cross-Media Information Management Systems // Conceptual Modeling, Lecture Notes in Computer Science, Vol. 8824, 2014. pp. 313-320.

[24] Völkel M., Haller H. Conceptual data structures for personal knowledge management // Online Information Review, Vol. 33, No. 2, 2009. pp. 298-315.

[25] Бездушный А. А., Бездушный А. Н., Серебряков В. А. Модель семантического управления личной информацией // Труды 16-й Всероссийской научной конференции «Электронные библиотеки перспективные методы и технологии, электронные коллекции» RCDL 2014. Дубна. 2014. Т. 1. С. 72-19.

[26] Бездушный А.А. Организация и управление личными каталогами научных публикаций с использованием технологий Semantic Web // XV Российская конференция с участием иностранных ученых «Распределенные информационные и вычислительные ресурсы» - DICR-2014. Новосибирск, Россия. 2014.

## MemoPIM: Personal Information and Knowledge Management with Semantic Technologies

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

In the course of daily activities, one is confronted with increasing volumes of information, a large part of which is stored in digital format. In this paper the MemoPIM system is presented. MemoPIM supports management of personal information and personal knowledge and develops ideas of the Semantic Desktop – an approach to the personal information space organization in accordance with principles of Semantic Web and Linked Open Data.

Knowledge and information is stored in RDF repository, based on OWL ontology. Ontology describes classes from different domains, such as file system ontology, publication ontology, calendar ontology etc. Knowledge can be imported from external sources or created manually by user. System provides adapters for following sources: file system, email, calendar, browser bookmarks. Imported information is transformed to RDF in accordance with system ontology. In some cases additional metadata is created, for example author, annotation and title is extracted from scientific articles.

During user work, actions performed by him are logged to knowledge base. These logs are later used for several tasks such as context inference, knowledge base analysis and tag recommendation. Context inference is based on spreading activation algorithm over RDF graph. Activation values for RDF nodes are recalculated after each user action. To create additional relations in RDF data, knowledge base analysis is performed. Another service provided by system is automatic categorization of resources. Categorization is based on tag recommendation algorithm.

System provides two types of user interfaces for knowledge management – based on hierarchical presentation of data and based on mind maps.