Принципы создания многоязычной электронной библиотеки для крупного информационного центра

© В.Н. Захаров © Ю.В. Никитин © Ал-др А. Хорошилов © Ал-ей А. Хорошилов

Федеральный исследовательский центр «Информатика и управление» РАН, Москва, Россия

vzakharov@ipiran.ru yuri.v.nikitin@gmail.com khoroshilov@mail.ru a.a.horoshilov@mail.ru

Аннотация. Описан подход к созданию многоязычной электронной библиотеки для крупного информационного центра. Показано, как организовать процесс формализации документов на разных языках таким образом, чтобы поиск был максимально эффективен и позволял пользователю получать результаты независимо от языка запроса и документов, содержащихся в базе данных. Исследование эффективности предложенного подхода показало достаточно высокие результаты, позволяющие применять его в промышленных информационных системах.

Ключевые слова: многоязычная электронная библиотека, многоязычный поиск, информационный поиск, автоматизированная обработка текстов, формализованное описание текста, смысловая структура, лингвистическое программное обеспечение, декларативные средства.

The Principles of Creating a Multilingual Electronic Library for a Large Information Center

© V.N. Zakharov © Yu.V. Nikitin © Al-dr A. Khoroshilov © Al-ey A. Khoroshilov

Federal Research Center Computer Science and Control of the Russian Academy of Sciences, Moscow, Russia

vzakharov@ipiran.ru yuri.v.nikitin@gmail.com khoroshilov@mail.ru a.a.horoshilov@mail.ru

Abstract. This paper describes the approach to creating a multilingual electronic library for a large information center. The authors show how to organize the process of formalizing documents in different languages in such a way that the search is most effective and allows the user to receive results regardless of the query language and documents contained in the database. The study of the effectiveness of the proposed approach has shown quite good results, allowing it to be used in industrial information systems.

Keywords: multilingual electronic library, multilingual search, information retrieval, automated text processing, formal description of text, semantic structure, linguistic software, declarative means.

1 Введение

В нашей стране в настоящее время функционирует множество организаций, каждый день имеющих дело с огромным объемом документов. Многие из этих организаций из-за специфики своей деятельности получают и обрабатывают документы на нескольких языках. К таким организациям можно отнести, например, предприятия авиационно-косми-ческой отрасли, для которых стоит важнейшая задача соответствия международным стандартам; всевозможные научные организации, для которых жизненно необходимо быть в курсе последних исследований и разработок; организации, обеспечивающие безопасность государства, для получения актуальной политической и технической информации и т. д. Соответственно, для решения различных задач дальнейшего эффективного использования получаемых постоянно документов необходима их предварительная автоматическая обработка, позволяющая свести к минимуму трудозатраты обслуживающего персонала. В настоящее время множество информационных систем имеет достаточно полный функционал работы с русскоязычными текстами, но, к сожалению, все эти системы имеют довольно скромные возможности при работе с разноязычными массивами документов, а задача сравнения текстов, выявления документов-дубликатов и заимствований в отечественных информационных системах в настоящий момент решена только для документов, написанных на одном языке. В то же время потребность в таких системах достаточно велика, и задача требует скорейшего решения.

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

2 Существующие подходы к организации электронных библиотек

Задача хранения и организации доступа к большим коллекциям документов стоит уже достаточно давно. За это время было разработано множество решений, которые в разной степени удовлетворяют требованиям, предъявляемым современными пользователями. Далее приведем некоторые данные о развитии такого программного обеспечения в настоящее время.

В работе [1] авторы провели серьезное сравнение свободно распространяемых технологий для организации электронных библиотек, существующих в настоящее время. Были протестированы системы OJS, ePubTK, DPubS, GAPWorks, Ambra, e-Journal. Сделан вывод, что практически все решения поддерживают общепринятые стандарты в области интеграции и обмена данными и имеют широкие возможности по генерации различных метаданных в зависимости от потребностей пользователя. Но, к сожалению, большинство из рассматриваемых продуктов более не развивается. Понятно, что такие системы позволяют решать стандартный набор задач и используются для небольших электронных библиотек.

При росте объемов документов становится важно решить задачу повышения эффективности поиска. Для этого многие ученые разрабатывают новые механизмы, одним из которых стал семантический поиск. В работе [2] авторы предлагают новый метод поиска, основанный на использовании модели S-тег. Особенностью данного метода является то, что индексируется не весь текст, а только его значимые части в зависимости от задачи, при этом за счет изменения размера значимой части можно контролировать точность и полноту.

Другим подходом к семантическому поиску, о котором сейчас пишет все большее число авторов, является использование онтологических моделей [3]. Основной идеей данного подхода является использование онтологий предметных областей для аннотирования содержания электронных ресурсов. Авторы работы [4] дополнили онтологический подход добавлением новых операций над онтологиями – проекции и масштабирования – и описали модель их применения для задач информационного поиска.

Еще одним направлением развития поиска в электронных библиотеках является многоязычный поиск. К сожалению, в настоящее время работ по этой тематике не так много. Одно из таких решений было описано в работе [5]. В ней представлено решение задачи двуязычного поиска с помощью тезауруса для двух языков (русского и английского). Похожего мнения придерживаются и многие иностранные исследователи, в том числе, например, в работе [6]. Несколько иной подход предложен в [7]: для решения задачи многоязычного поиска использован инструментарий систем автоматического перевода текстов.

3 Организация многоязычной электронной библиотеки для крупного информационного центра

3.1 Архитектура многоязычной электронной библиотеки

Проанализировав подходы и решения, имеющиеся на сегодня в области разработки современных электронных библиотек, авторами был составлен список требований, которым должна удовлетворять система, функционирующая в крупном информационном центре:

- обеспечение модульной архитектуры с возможностью быстрого включения в систему новых модулей;
- использование средств СУБД, позволяющих максимально эффективно организовать процесс доступа к данным;
- обеспечение возможности оперативного пополнения декларативных средств системы;
- обеспечение максимальной простоты добавления новых языков в систему;
- обеспечение распределенной массово-параллельной лингвистической и статистической обработки загружаемых данных;
- обеспечение масштабируемости на множество узлов обработки без деградации инфраструктуры обработки данных;
- обеспечение всех этапов лингвистической обработки, включающей этапы графематического, морфологического, семантико-синтаксического, концептуального и дистрибутивно-статистического анализа [12];
- обеспечение эффективного многоязычного поиска;
- обеспечение эффективного сравнения смыслового содержания документов, в том числе поиска заимствований и документов-дубликатов [13-15];
- обеспечение поддержки общепринятых стандартов в области интеграции и обмена данными;
- создание наиболее полной и удобной структуры метаданных для хранимых в базе документов;
- обеспечение удобного пользовательского интерфейса, максимально упрощающего доступ пользователя ко всему функционалу электронной библиотеки.

На Рис. 1 представлена предлагаемая авторами архитектурная схема многоязычной электронной библиотеки для крупного информационного центра.

Рисунок 1 Архитектурная схема многоязычной электронной библиотеки для крупного информационного центра

3.2 Процесс формализации документов в многоязычной электронной библиотеке

Основной задачей при выполнении формализации документа является представление смысловой структуры текста в структурированном виде. По мнению авторов, формализованное представление текстового содержания документа должно включать:

- библиографические реквизиты (например, информационный источник, рубрика, автор, наименование и дата публикации и т. п.);
- аннотацию или реферат документа;
- список ключевых выражений;
- список значимых объектов (персоны, организации, территории, наименования товаров, географические объекты, бренды, и т. д.);

При этом для создания многоязычной системы данная информация должна содержаться на всех поддерживаемых языках. Также каждому документу должна соответствовать следующая информация:

- содержащиеся в документе формулы, параметры с их числовыми значениями и т. д.;
- классификация документа по смысловому содержанию – отнесение его к той или иной рубрике и кластеризация [11] (группировка) текстов публикаций по темам;
- ссылки на связанные документы (цитаты, заимствования, документы-дубликаты, близкие по смыслу документы) [8–10].

3.3 Организация многоязычного поиска

В ходе исследования авторами была разработана двухступенчатая процедура поиска, которая может быть использована для поиска в многоязычном массиве информации. На первом этапе запрос был преобразован в его унифицированное семантическое представление, на втором этапе производился поиск в базе данных стандартными средствами. Рассмотрим каждый из этапов подробнее.

3.3.1 Метод трансформации поискового запроса в его унифицированное семантическое представление

Разработанный авторами метод трансформации поискового запроса в его унифицированное семантическое представление основан на использовании многоязычного словаря унифицированных формализованных представлений наименований понятий [16]. В данном исследовании словарь был сформирован для трех языков (русского, английского и немецкого), но в этот словарь могут быть добавлены эквиваленты на других языках при наличии переводных словарей схожих объемов. Также для работы метода необходимы процедуры морфологического, семантико-синтаксического и концептуального анализа для каждого языка, который содержится в словаре унифицированных формализованных представлений наименований понятий. При выполнении этих условий трансформация поискового запроса сводится к следующему алгоритму (Алгоритм 1):

Шаг 1. Определяется язык обрабатываемого запроса.

Шаг 2. С помощью процедуры концептуального анализа (для выявленного языка) определяется совокупность значимых наименований понятий с указанием местоположений этих понятий в тексте запроса.

Шаг 3. Каждое наименование понятия запроса приводится к нормальной форме с помощью процедуры автоматической пословной нормализации.

Шаг 4. Каждое нормализованное наименование понятия ищется в многоязычном словаре унифицированных формализованных представлений наименований понятий, после чего ему присваивается номер из этого словаря. Пример словаря приведен в Таблице 1.

Таблица 1 Фрагмент многоязычного словаря унифицированных формализованных представлений наименований понятий

№ n/n	Значения на русском языке	Эквиваленты на английском языке	Эквиваленты на немецком языке
…	…	…	
816437	нефтехранилище / нефтесклад / хранилище	oil reservoir / oil storage / petroleum storage / tank farm	öllager / erdöllager / tanklager
816438	нефть / каустобиолит / петролеум / черный золото	mineral oil / naphtha / oil / petrol / petroleum / rock-oil	öl / caustobiolith / petroleum / petrol
816439	нефтяник / нефтедобытчик	oilman / oil-industry worker	ölproduzent / ölhändler
…	…	…	

Схема работы данного алгоритма отображена на Рис. 2.

Рисунок 2 Схема работы алгоритма трансформации поискового запроса в его унифицированное семантическое представление

3.3.2 Процесс поиска в многоязычных массивах, основанный на использовании метода трансформации поискового запроса

Далее рассмотрим алгоритм поиска документов в многоязычных массивах с использованием стандартных средств СУБД (Алгоритм 2):

Шаг 1. На вход поступает поисковый запрос, после чего он обрабатывается с помощью алгоритма 1.

Шаг 2. Средствами СУБД производится поиск наименований понятий запросов в многоязычном массиве (при поиске сравниваются не сами наименования понятий, а их номера в многоязычном словаре унифицированных формализованных представлений наименований понятий).

Шаг 3. Запускается процедура ранжирования результатов поиска, полученных с помощью стандартных средств СУБД. Процедура ранжирования зависит от типа поиска.

Шаг 4. Выдача результатов поиска пользователю.

На Рис. 3 представлена общая схема работы программного модуля, в котором реализованы описанные алгоритмы.

Целью эксперимента являлась проверка работоспособности предложенных методов поиска информации в многоязычном массиве, установление их эффективности [8], а также возможности их использования в промышленных информационных системах. Эксперимент проводился на основе разработанного авторами программного комплекса. В качестве исходных данных для эксперимента был взят массив текстов по тематике «Информационные технологии» (182641 текст).

Рисунок 3 Общая схема работы программного модуля поиска текстовой информации в многоязычных массивах

3.3.3 Эксперимент по проверке разработанного метода поиска в многоязычном массиве

Эксперимент проводился в несколько этапов:

1. На первом этапе тексты документов, приготовленные для эксперимента, были загружены в систему и обработаны при помощи *алгоритма 1*, изложенного в разделе 3.3.1. Все результаты обработки были занесены в базу данных программного комплекса.

2. На втором этапе из загруженных текстов было выбрано 35000 предложений и 90000 случайных наименований понятий. При этом был создан контрольный массив, где содержались все адреса предложений и наименований понятий в текстах документов коллекции.

3. На третьем этапе выбранные предложения и наименования понятий были переведены на английский и немецкий язык с помощью системы перевода Google Переводчик (https://translate.google.ru/).

4. На четвёртом этапе был произведен поиск каждого из переведенных на третьем этапе предложений и наименований понятий в русскоязычном массиве документов. Для этого использовался *алгоритм 2*, изложенный в разделе 3.3.2. После этого информация об адресах найденных соответствий сопоставлялась с информацией, полученной в п. 2.

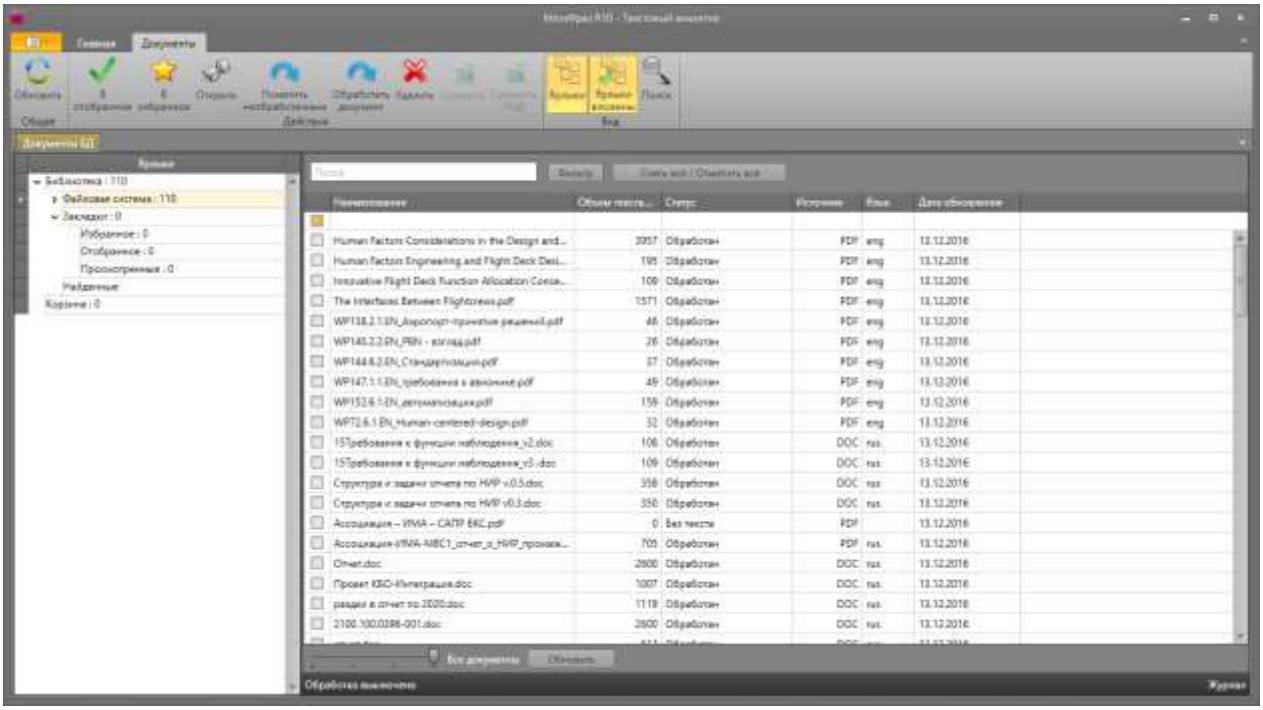

Рисунок 4 Скриншот интерфейса электронной библиотеки MF Text Analyst

5. На пятом этапе с помощью данных, полученных в п. 4, были получены значения полноты, точности и F1-меры. Результаты приведены в таблице 2.

Таблица 2 Значения показателей эффективности метода

	Полнота	**Точность**	**F1-мера**
Поиск наименований понятий	0.88	0.96	0.92
Поиск предложений	0.79	0.99	0.89
Среднее значение	0.84	0.98	0.91

4 Заключение

Идеи, описанные выше, были реализованы в виде программного продукта MF Text Analyst на базе программно-лингвистической платформы MetaFraz R10. Данный программный комплекс предназначен для выполнения следующих простых операций:

- ведение электронной библиотеки научно-технических документов;
- автоматическое формирование формализован ного представления документов;
- семантический поиск, отбор и сравнение документов.

MF Text Analyst позволяет загружать в БД документы в наиболее распространенных форматах (PDF, DOC, DOCX, TXT и др.), а затем извлекать текстовое содержимое и производить все этапы лингвистической обработки. Скриншот интерфейса электронной библиотеки MF Text Analyst представлен на рис. 4.

Также в данном программном продукте в тестовом режиме реализован многоязычный поиск. Его эффективность была проверена на коллекции размером в 182641 документ и показала неплохие для данного этапа исследований результаты. Предложенный авторами метод показал соответствующую аналогам скорость поиска при использовании СУБД RavenDB. Далее для улучшения показателей эффективности необходимо продолжать работу по доработке программного обеспечения, а также пополнять словари новой лексикой. Указанные мероприятия позволят значительно улучшить качество работы разработанных алгоритмов на текстах, относящихся к широкому спектру предметных областей.

Литература

[1] Елизаров, А.М., Зуев, Д.С, Липачёв, Е.К.: Свободно распространяемые системы управления электронными научными журналами и технологии электронных библиотек. Труды XV Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2013, г. Ярославль, 14–17 октября 2013 года, сс. 227-236 (2013)

[2] Малахов, Д.А., Сидоренко, Ю. А., Атаева, О.М., Серебряков, В.А.: Семантический поиск как средство взаимодействия с электронной библиотекой. Труды XVIII Межд. конф. DAMDID / RCDL'2016 «Аналитика и управление данными в областях с интенсивным

использованием данных», 11–14 октября 2016 года, Ершово, Москва, сс. 85-91 (2016)

[3] Ле Хоай, Тузовский, А.Ф.: Разработка семантических электронных библиотек на основе онтологических моделей. Труды XV Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2013, г. Ярославль, 14 – 17 октября 2013 года, сс. 143-151 (2013)

[4] Голицына, О.Л., Максимов, Н.В., Окропишина, О.В., Строгонов, В.И.: Онтологический подход к идентификации информации в задачах документального поиска: практическое применение. Научно-техническая информация. Серия 2: Информационные процессы и системы, (3), сс. 1-8 (2013)

[5] Добров, Б.В., Лукашевич, Н.В.: Организация двуязычного поиска в университетской информационной системе «Россия». Труды четвертой Всерос. науч. конф. RCDL'2002 «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», г. Дубна, 15–17 октября 2002 г., сс. 148-158 (2002)

[6] Oard, D.: Alternative Approaches for Cross-Language Text Retrieval. Proc. of the AAAI Spring 1997 Symposium on Cross-Language Text and Speech Retrieval (1997)

[7] Cardeñosa, J., Gallardo, C., Toni, A.: Multilingual Cross Language Information Retrieval: A New Approach. Seventh Int. Conf. on Computer Science and Information Technologies, 28 September – 2 October, 2009, Yerevan, Armenia (2009)

[8] Хорошилов, А.А.: Методы автоматического установления смысловой близости документов на основе их концептуального анализа. Труды XV Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2013, г. Ярославль, 14–17 октября 2013 года, сс. 369-376 (2013)

[9] Захаров, В.Н., Хорошилов, Ал-др А., Хорошилов, Ал-ей А.: Метод автоматического выявления неявно выраженных заимствований в научно-технических текстах. Искусственный интеллект и принятие решений, (1), сс. 10-20 (2017)

[10] Захаров, В.Н., Хорошилов, Ал-др. А., Хорошилов, Ал-ей. А.: Метод выявления заимствований в текстах разноязычных документов. Труды XVIII Межд. конф. DAMDID / RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», 11–14 октября 2016 года, Ершово, Москва, сс. 277-282 (2016)

[11] Борзых, А.И., Брагина, Г.А., Хорошилов, А.А.: Методы автоматической кластеризации документов в хранилищах научно-технической информации для решения задачи поиска плагиата в текстах документов. Информатизация и связь, (8), сс. 33-37 (2012)

[12] Дмитришин, А.Н., Калинин, Ю.П., Никитин, Ю.В., Хорошилов, А.А., Хорошилов, А.А.: Технологии автоматической обработки и семантического анализа разноязычных документов в системе мониторинга мирового потока научно-технической информации крупного информационного центра. Информатизация и связь, (1), сс. 49-55 (2017)

[13] Zakharov, V., Khoroshilov, A.: Automatic Assessment of Similarity of the Texts' Thematic Content on The Base of their Formalized Semantic Descriptions Comparison. CEUR Workshop Proceedings. Proc. of the 14th All-Russian Scientific Conf. "Digital libraries: Advanced Methods and Technologies, Digital Collections", Pereslavl-Zalessky, Russia, October 15–18, 934, pp. 143-149 (2012)

[14] Zakharov, V., Khoroshilov, A.: Semantic Methods for Solving a Problem of Automatic Detection of Plagiarism in Structured Scientific and Technical Documents. CEUR Workshop Proceedings. Selected Papers of the 15th All-Russian Scientific Conf. "Digital Libraries: Advanced Methods and Technologies, Digital Collections", Yaroslavl, Russia, October 14–17, 1108, pp. 165-172 (2013)

[15] Khoroshilov, A.A.: Method for Detecting Implicit Plagiarism in Scientific and Technical Texts on the Basis of Their Conceptual Analysis. CEUR Workshop Proceedings. Selected Papers of the XVII Int. Conf. on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2015), Obninsk, Russia, October 13–16, 2015, 1536, pp. 266-372 (2015)

[16] Zakharov, V., Khoroshilov, Alexandr, Khoroshilov, Alexey: A Method of Automatic Plagiarism Detection in Multilingual Documents. CEUR Workshop Proceedings. Selected Papers of the XVIII Int. Conf. on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2016), 1752, pp. 181-186 (2016)