Программные средства создания и наполнения полнотекстовых электронных библиотек

© Г.И. Назаренко[1], В.А. Плотникова[1], И.В. Смирнов[2], И.В. Соченков[2], И.А. Тихомиров[2]

[1]Медицинский центр Банка России

[2]Учреждение Российской академии наук Институт системного анализа РАН, г. Москва
ivs@isa.ru

Аннотация

В работе представлены программные средства полнотекстовых электронных библиотек с сервисами автоматического наполнения документами, автоматического определения полей метаданных документов и высокоточного семантического поиска информации. Указанные программные средства позволяют быстро сформировать тематические коллекции электронных документов из различных сетевых источников и обеспечивают высокорелевантные результаты поиска документов.

1 Введение

Полнотекстовые электронные библиотеки являются эффективным инструментом для поиска информации в научной и образовательной среде. Научные знания делятся на узконаправленные области, по каждой из которых существует множество электронных источников информации, включая специализированные журналы, сборники трудов научных конференций и другие информационные ресурсы. Электронные библиотеки с тематическими коллекциями полнотекстовых документов позволяют получать достоверную информацию в одной определенной области человеческой деятельности, исключая недостоверную и рекламную информацию, что отличает их от традиционных поисковых машин. В связи с этим актуально создание тематических коллекций электронных документов и объединение их в электронные библиотеки с сервисами полнотекстового поиска.

Известно, что большинство научных и научно-популярных изданий размещают в свободном доступе электронные версии публикаций. Это даёт возможность автоматически наполнять электронные коллекции документами из ресурсов интернета. При этом могут быть отобраны только достоверные проверенные ресурсы, соответствующие тематике электронной коллекции. В то же время при работе с большими объёмами полнотекстовых документов возникают задачи обеспечения точности поиска информации.

Институтом системного анализа РАН совместно с Медицинским центром Банка России созданы программные средства полнотекстовых электронных библиотек (ПС ПЭБ), и медицинская электронная библиотека (МЭБ), функционирующая на их основе. ПС ПЭБ, обеспечивает автоматическое наполнение коллекций электронной библиотеки документами из различных источников и высокоточный поиск документов в электронной библиотеке.

2 Автоматическое наполнение полнотекстовой электронной библиотеки

Электронная библиотека состоит из коллекций электронных документов. Каждая коллекция содержит документы по одной теме. Документы загружаются автоматически из сетевых ресурсов (интернет, интранет) или заносятся вручную. Как уже было отмечено, перспективным является способ автоматического наполнения, когда для каждой коллекции задаётся набор сетевых и локальных ресурсов, из которых необходимо автоматически загружать документы (обычно это веб-сайты) для пополнения коллекции.

Для автоматического наполнения электронной библиотеки из сетевых ресурсов разработан программный модуль – краулер, который обходит веб-сайты по гипертекстовым ссылкам и загружает электронные документы в библиотеку.

Сложность автоматического наполнения коллекций электронной библиотеки электронными документами из источников в интернете заключается в том, что на сайтах содержится много сопутствующей информации – новости, содержание выпусков журналов, контактная информация и проч., в то время как в электронную библиотеку должны попасть только целевые документы. Для решения этой задачи в краулере реализована специальная процедура, которая на основании HTML-структуры и других характеристик страниц сайта определяет, какие документы необходимо загружать в библиотеку, а какие нет. Эта процедура использует шаблоны на основе регулярных выражений и XPath-выражений, которые применяются к URL и к DOM деревьям HTML-документов соответственно.

Шаблоны формируются вручную на основе анализа структуры сайта и выделения подразделов, содержащих целевые документы, и хранятся в конфигурационном файле, создаваемом отдельно для каждого сайта.

Подключение нового источника загрузки документов заключается в создании конфигурационного файла, что занимает от 10 до 30 минут в зависимости от сложности структуры сайта. Для каждого сетевого ресурса задаётся периодичность обхода, что позволяет автоматически пополнять коллекции новыми публикуемыми документами и поддерживать их в актуальном состоянии.

Таким образом, уже на этапе наполнения электронной библиотеки производится отсечение нецелевой информации, что впоследствии способствует повышению точности информационного поиска и значительному уменьшению объёмов хранимых данных.

3 Автоматическое определение метаданных документов: авторов, названий, дат публикации

Документы в электронных коллекциях структурированы по метаданным. Это означает, что для каждого документа, как правило, известны авторы, название, дата публикации, источник публикации. Структурированность документов обеспечивает более точный поиск информации и позволяет создавать систематические каталоги по авторам, издательствам, названиям документов.

При автоматическом наполнении электронной библиотеки документами из сетевых ресурсов возникает задача автоматического определения значений метаданных загружаемых документов. Решение этой задачи основывается на анализе структуры целевых документов и промежуточных гипертекстовых страниц, содержащих оглавления выпусков журналов, подборок статей и т. п. Обычно эти страницы имеют регулярную HTML-структуру и представляют собой списки публикуемых документов с указанием авторов, названий и другой информации, включая ссылки на сами документы. Это позволяет краулеру автоматически выделять значения полей документов на основании правил, которые также задаются в конфигурационном файле для каждого отдельного ресурса.

Для выделения значений метаданных документов реализована специальная процедура, которая определяет значения полей двумя путями:

- непосредственно из целевых документов;
- из промежуточных документов в формате HTML, содержащих ссылки и описания целевых документов.

Работа процедуры основана на применении XPath-шаблонов к DOM деревьям HTML-документов.

Автоматическое выделение полей метаданных позволяет организовать в электронных коллекциях не только эффективный полнотекстовый поиск информации, но также такой вид поиска, когда пользователь фокусируется на публикациях за некоторый период времени или осуществляет выборку публикаций определённых авторов.

4 Высокоточный полнотекстовый поиск документов и поиск по метаданным

Поисковые сервисы являются неотъемлемой частью полнотекстовой электронной библиотеки. Известно, что традиционные подходы к поиску информации основываются на статистических характеристиках слов документов (используются, например, TFIDF веса слов), при этом поиск документов сводится к поиску по ключевым словам, в лучшем случае с учётом морфологии языка. Очень часто такой подход даёт слаборелевантные результаты.

В течение последних лет были созданы оригинальные методы информационного поиска, которые объединяют статистические и лингвистические подходы к обработке текстов на естественном языке [4, 5, 7]. В частности, эти методы позволяют выполнять семантический поиск информации, т. е. поиск по смыслу запросов.

Семантический поиск информации основан на лингвистической теории, описывающей законы передачи осмысленной информации в естественном языке [2]. Опираясь на эту теорию, смысл высказываний на естественном языке можно представить с помощью неоднородных семантических сетей, которые позволяют реализовать смысловое сравнение текстов запроса и документов и определить смысловую близость между ними. Для выполнения семантического поиска все документы электронной библиотеки подвергаются морфологическому, синтаксическому и семантическому анализу.

Использование методов семантического поиска в электронной библиотеке обеспечивает высокоточный поиск документов по их полным текстам по запросам на естественном языке [6]. Кроме того, семантический полнотекстовый поиск позволяет находить не только документы в электронной библиотеке, но и непосредственно ответы на интересующие пользователя вопросы по выбранной теме (коллекции). При этом у пользователя сохраняется возможность формулировки запроса в виде набора ключевых слов, если он считает этот вид поиска наиболее подходящим для удовлетворения собственной информационной потребности.

В разработанных ПС ПЭБ существует возможность поиска документов не только по полнотекстовому содержанию, но и по автору, названию, дате публикации и источнику, с которого был получен документ. Авторы и название задаются в поисковом запросе в произвольной форме на естественном языке, при этом результаты поиска по этим полям объединяются логикой «И».

Таким образом, средства полнотекстового семантического поиска и поиска по метаданным документов повышают эффективность поиска необходимой информации в полнотекстовой электронной библиотеке.

5 Архитектура ПС ПЭБ

На рис. 1 приведена схема архитектуры ПС ПЭБ, поясняющая принципы взаимодействия компонен-

тов системы с пользователем и внешними информационными ресурсами.

ПС ПЭБ состоит из следующих функциональных подсистем:

- ✓ наполнения и актуализации коллекций электронных документов;
- ✓ индексации электронных документов;
- ✓ информационного поиска;
- ✓ интерфейса пользователя.

Система является многопользовательской; все подсистемы имеют распределённую многокомпонентную внутреннюю структуру. Распределённость компонентов ПС ПЭБ позволяет масштабировать их для создания обширных тематических коллекций (до десятков миллионов электронных документов) путём введения в её состав дополнительных серверов.

5.1 Подсистема наполнения и актуализации коллекций электронных документов

Функции подсистемы наполнения и актуализации коллекций электронных документов схожи с аналогичными функциями веб-краулеров поисковых систем интернета.

Подсистема осуществляет:

- ✓ обход внешних информационных ресурсов (веб-сайтов, хранилищ документов);
- ✓ получение электронных документов;
- ✓ помещение полученных документов в коллекцию,
- ✓ преобразование документов к внутреннему представлению (поддерживаются все распространённые форматы текстовых документов, имеется возможность добавления поддержки новых форматов).

В подсистеме реализованы описанные выше функции фильтрации целевых электронных документов, содержащих тематическую информацию, от промежуточных документов, а также выделение значений метаданных документов.

5.2 Подсистема индексации электронных документов

В подсистеме индексации электронных документов реализован комплексный лингвистический анализ, содержащий этапы морфологической, синтаксической и семантической обработки. В результате текст преобразуется во внутреннее представление в соответствии с реляционно-ситуационной моделью представления текста [5, 7]. Результат преобразования сохраняется в хранилище индексатора. Использована структура данных, известная как «обратный индекс», которая модифицирована для эффективной выборки документов с учётом метаданных и семантической информации слов.

Оригинальные версии электронных документов сохраняются в хранилище коллекций.

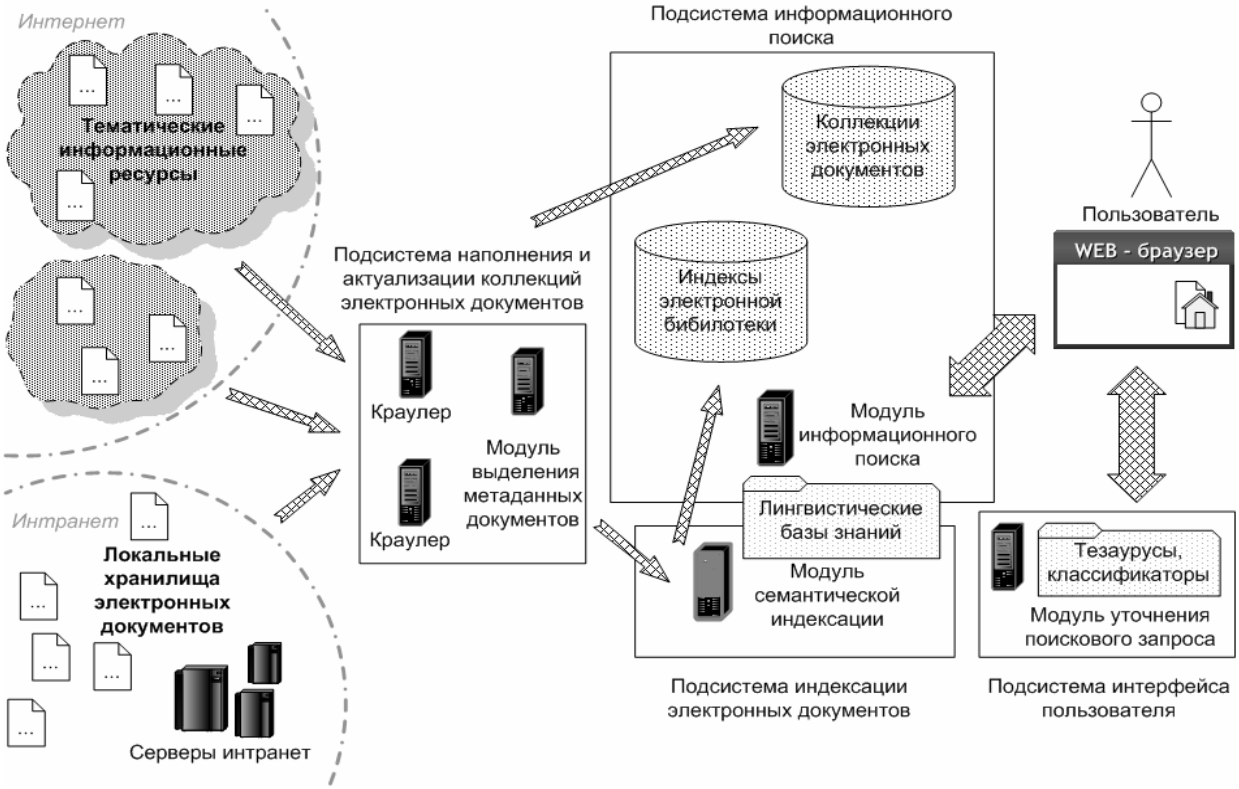

Рис. 1. Архитектура ПС ПЭБ

5.3 Подсистема информационного поиска

На этапе поиска производится лингвистический анализ текста запроса пользователя по схеме, аналогичной анализу текста документа. При обработке запроса рассчитывается релевантность документов запросу на основе статистической и семантической информации [7] с учётом выборки документов, соответствующих заданным значениям метаданных. Реализованный в системе семантико-статистический подход доказал свою эффективность в рамках семинара РОМИП-2008 [6]. Подход даёт возможность поиска ситуаций, описываемых фразами и предложениями на естественном языке, включая вопросно-ответный режим поиска.

5.4 Подсистема интерфейса пользователя

Пользовательский интерфейс ПС ПЭБ представляет собой веб-форму и программный модуль обработки запроса. Интерфейс пользователя содержит элементы, необходимые для ввода текста запроса и выбора значений метаданных. Для текстовых полей ввода предусмотрена интерактивная терминологическая подсказка на основе тезаурусов предметной области (с учётом таксономии терминов, переводом на другие языки).

Результаты поиска предоставляются пользователю в виде списка ссылок на найденные документы с краткими аннотациями. Сохранённые полнотекстовые копии доступны пользователю на этапе просмотра результатов поиска (если это не противоречит лицензионному соглашению об использовании материалов информационного ресурса – источника электронных документов).

6 Развёртывание ПС ПЭБ

Развёртывание электронной библиотеки на основе ПС ПЭБ включает следующие шаги:

1. Установка программных модулей ПС ПЭБ на серверы и их конфигурирование.
2. Определение тематики коллекций, в которых будет производиться поиск и выбор информационных ресурсов (веб-сайтов, хранилищ электронных документов) – доверенных источников информации.
3. Анализ структуры информационных источников, определение множества выделяемых метаданных электронных документов и настройка подсистемы наполнения и актуализации коллекций электронных документов.
4. Настройка подсистемы интерфейса пользователя с учётом выбранных метаданных, по которым будет выполняться поиск, подготовка информационно-справочных ресурсов – тезаурусов предметных областей – и их интеграция в подсистему интерфейса пользователя.

Впоследствии в систему могут быть добавлены как новые информационные ресурсы в уже существующие коллекции, так и новые тематические коллекции. Актуализация коллекций производится в фоновом режиме параллельно с обработкой поисковых запросов пользователей.

7 Заключение

Разработанные программные инструментальные средства предназначены для создания и автоматического наполнения полнотекстовых электронных библиотек с сервисами высокоточного информационного поиска. Эти средства позволяют формировать тематические коллекции электронных документов и автоматически пополнять их из различных источников, включая сетевые ресурсы – сайты электронных изданий. Автоматическое выделение авторов, названий и дат публикации загружаемых документов обеспечивает структурированность электронных коллекций и более быстрый доступ к искомым документам. Семантический поиск документов и поиск по полям обеспечивают высокоточный поиск информации в электронной библиотеке.

Представляемые программные средства функционируют в распределенной вычислительной среде с возможностью масштабирования и обладают рядом дополнительных особенностей:

- ✓ работа со всеми распространенными форматами текстовых документов;
- ✓ работа с документами на русском, английском, немецком языках с возможностью поддержки других языков;
- ✓ уточнение поисковых запросов с помощью тезаурусов и словарей;
- ✓ для работы с полнотекстовой электронной библиотекой используется веб-интерфейс;
- ✓ интеграция с системой автоматизации библиотек ИРБИС – поиск по библиографическим описаниям.

С помощью разработанных программных средств была создана медицинская электронная библиотека, включающая коллекцию медицинских журналов (400 тыс. документов), коллекцию клинических руководств (30 тыс. документов) и коллекцию авторефератов (900 документов). Все документы загружались в коллекции автоматически из интернета с выделением полей метаданных. МЭБ внедрена в библиотеке Медицинского центра Банка России и активно используется медицинскими работниками при поиске профильной информации.

Демонстрационная версия полнотекстовой электронной библиотеки доступна в интернете по адресу http://elib.isa.ru. Библиотека содержит коллекции по нанотехнологиям, медицине и генетике (более 30 тыс. документов).

Литература

[1] Абросимов А.Г., Зуев Д.С. Научно-образовательная электронная библиотека ВУЗа // Труды RCDL-2008. – http://rcdl.ru/doc/2008/374_379_paper46.pdf.

[2] Золотова Г.А., Онипенко Н.К., Сидорова М.Ю. Коммуникативная грамматика русского языка. –

M., 2004. – 544 с.
[3] Паринов С.И. Развитие электронных библиотек – путь к открытой науке // Труды RCDL-2009. – http://rcdl.ru/doc/2009/225_234_Invited-2.pdf.
[4] Тихомиров И.А, Смирнов И.В. Интеграция лингвистических и статистических методов поиска в поисковой машине Exactus // Труды межд. конф. Диалог'2008. – С. 485-491.
[5] Осипов Г.С. Приобретение знаний интеллектуальными системами: Основы теории и технологии. – М.: Наука, Физматлит, 1997. – 112 с.
[6] Смирнов И.В., Соченков И.В., Муравьев В.В., Тихомиров И. А. Результаты и перспективы поискового алгоритма Exactus // Труды российского семинара по оценке методов информационного поиска РОМИП'2007-2008. – Санкт-Петербург: НУ ЦСИ, 2008. – С. 66-76.
[7] Osipov G., Smirnov I., Tikhomirov I. Application of linguistic knowledge to search precision improvement // Proc. of 4th Int. IEEE Conf. on Intelligent Systems, 2008. – V. 2. – P. 17-2-17-5.

Software for creation and filling full-text electronic libraries

G.I. Nazarenko, V.A. Plotnikova, I.V. Smirnov, I.V. Sochenkov, I.A. Tikhomirov

The paper presents software tools for full-text electronic libraries with automatic filling with documents, detection of documents' meta-fields and high-precision search in the e-library. The software provides rapid creation of thematic collections of electronic documents from several network resources and high-relevant search results.