Полнотекстовые электронные библиотеки с сервисами автоматического наполнения и высокоточного поиска

Full-text E-libraries with Services of Automatic Acquisition and High-precision Search

Повнотекстові електронні бібліотеки із сервісом автоматичного наповнення та пошуку високої точності

Г. С. Осипов, И. В. Смирнов, И. В. Соченков, И. А. Тихомиров
Учреждение Российской академии наук
Институт системного анализа РАН, Москва, Россия

Gennady Osipov, Ivan Smirnov, Ilya Sochenkov, and Ilya Tikhomirov
Institute for Systems Analysis,
Russian Academy of Sciences, Moscow, Russia

Г. С. Осипов, І. В. Смирнов, І. В. Соченков, І. О. Тихомиров
Заклад Російської академії наук
Інститут системного аналізу РАН, Москва, Росія

Представлены программные средства полнотекстовых электронных библиотек с сервисами автоматического наполнения электронных коллекций документами из сетевых ресурсов и сервисами высокоточного поиска документов в электронной библиотеке.

The paper presents software for full-text e-libraries providing services of automatic acquisition of documents from network resources and high-precision document retrieval in the e-library.

Описано програмні засоби повнотекстових електронних бібліотек із сервісами автоматичного наповнення електронних колекцій документами із мережевих ресурсів і сервісами пошуку документів високої точності в електронній бібліотеці.

Введение

Сегодня работа в большинстве областей человеческой деятельности, так или иначе, связана с поиском и обработкой информации. За последние тысячелетия накоплены большие объёмы информации, которые особенно быстро нарастают в последнее время в связи со стремительным развитием науки. Реальность такова, что владение информацией по определенному вопросу даёт неоспоримые преимущества в принятии важных решений, а также обеспечивает эффективность научных исследований и образовательной деятельности. Современный ритм жизненной активности требует сокращения времени на поиск информации, которую необходимо ещё обработать, в то время как рост объёмов информации, как правило, снижает её доступность.

Человеческие знания, литературные произведения и другие виды результатов человеческой деятельности традиционно представлены в форме текстов на естественных языках в книгах, статьях журналов, трудах конференций и проч. Научно-техническая эволюция привела к тому, что эта информация стала храниться в электронном виде, а развитие сети интернет – универсальной среды коммуникации, обеспечило распространение информации и широкий доступ к ней. Казалось бы, несмотря на свой бурный рост, информация стала доступна «в любое время в любом месте» (в том числе за счет поисковых машин Интернет), однако быстрый поиск нужной информации по какой-либо теме остаётся непростой задачей.

Человеческие знания разделяются на узконаправленные области, по каждой из которых существует множество электронных источников информации, включая специализированные журналы, сборники трудов научных конференций и другие информационные ресурсы. Известно, например, что большинство научных и научно-популярных изданий имеют в Интернет свои сайты, на которых в свободном доступе присутствуют электронные копии публикаций. Но получить требуемую информацию по заданной теме с помощью традиционных поисковых машин в Интернет сложно: в

ответ на запрос выдаётся много нерелевантной, недостоверной информации, часто из источников по другим темам.

Электронные библиотеки с тематическими коллекциями полнотекстовых документов являются эффективным инструментом для поиска информации специалистами в разных областях, особенно в научной и образовательной среде. Они позволяют получать достоверную информацию из проверенных источников в одной определенной области человеческой деятельности, исключая ненужную информацию, что отличает их от традиционных поисковых машин. При создании таких полнотекстовых электронных библиотек возникает проблема наполнения тематических коллекций: занесение документов в электронную библиотеку вручную трудоёмко, поэтому полнота охвата темы не обеспечивается. В то же время, при работе с большими объёмами полнотекстовых документов возникают проблемы точности поиска информации.

В Институте системного анализа РАН разработаны программные средства полнотекстовых электронных библиотек с интеллектуальными сервисами автоматического наполнения и высокоточного поиска, речь о которых пойдет далее.

Автоматическое наполнение полнотекстовых электронных библиотек из сетевых ресурсов

Как уже было сказано, в сети Интернет и локальных сетях передачи данных находится огромное количестве информации по различным областям деятельности человека. Эта информация является подходящим материалом для наполнения тематических коллекций полнотекстовых электронных библиотек.

Для автоматического наполнения электронной библиотеки из сетевых ресурсов нами был разработан кроулер, который обходит web-сайты по гипертекстовым ссылкам и загружает электронные документы, в общем случае любого формата, в библиотеку. Таким образом, для каждой тематической коллекции электронной библиотеки задаётся набор сетевых и локальных ресурсов (обычно это web-сайты), из которых необходимо загружать документы для пополнения коллекции.

В кроулере реализована процедура, которая на основании структуры и других характеристик сайта позволяет загружать в библиотеку только целевые документы, отсеивая сопутствующую информацию – новости, содержание выпусков журналов, контактную информацию и прочее. Правила определения целевых документов хранятся в конфигурационном файле, создаваемом отдельно для каждого сайта.

Подключение нового источника загрузки документов заключается в создании конфигурационного файла, что занимает всего несколько 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.

Автоматическое наполнение полнотекстовых электронных библиотек документами из сетевых ресурсов обеспечивает полноту охвата и актуальность материалов по заданным темам.

Автоматическое определение метаописаний электронных документов: авторов, названий, дат публикации

Документы в электронных коллекциях структурированы по метаописаниям. Это означает, что для каждого документа, как правило, должны быть известны авторы, название, дата публикации, источник публикации.

В нашем кроулере реализована специальная процедура автоматического определения метаописаний документов на основе анализа структуры страниц web-сайтов, с которых загружаются документы. Выделение авторов, названий и дат публикации выполняется на основании правил, которые также задаются в конфигурационном файле для каждого отдельного сетевого ресурса. Реализованная процедура автоматического определения метаописаний документов работает с достаточно высокой точностью (95–99%) на сайтах со структурой любой сложности.

Структурированность документов по метаописаниям обеспечивает более точный поиск информации и позволяет создавать систематические каталоги по авторам, издательствам, названиям документов, а также организовать в электронных коллекциях такой вид поиска, когда пользователь фокусируется на публикациях за некоторый период времени или осуществляет выборку публикаций определённых авторов.

Высокоточный семантический поиск документов и поиск по метаописаниям

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

Между тем, авторами разработаны и развиваются оригинальные методы семантического поиска информации, т. е. поиска по смыслу запросов[1, 2]. Семантический поиск основан на лингвистической теории, описывающей законы передачи осмысленной информации в естественном языке [3]. 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 [5]. Кроме того, семантический полнотекстовый поиск позволяет находить не только документы в электронной библиотеке, но и непосредственно ответы на интересующие пользователя вопросы по выбранной теме (в коллекции). При этом у пользователя сохраняется возможность формулировки запроса в виде набора ключевых слов, если он считает этот вид поиска наиболее подходящим для удовлетворения информационной потребности.

Реализована возможность поиска документов не только по полнотекстовому содержанию, но также 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 метаописаниям объединяются логикой «И».

Средства полнотекстового семантического поиска и поиска по метаописаниям документов повышают эффективность поиска необходимой информации в полнотекстовых электронных библиотеках.

Заключение

Разработанные в Институте системного анализа РАН программные средства полнотекстовых электронных библиотек могут быть востребованы в профессиональной деятельности различного вида, прежде всего, в научной и образовательной среде. Эти средства позволяют формировать тематические коллекции электронных документов и автоматически пополнять их из различных источников, включая сетевые ресурсы. 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 метаописаниям обеспечивают высокоточный поиск информации в электронной библиотеке. Перечисленные выше особенности обеспечивают полноту охвата и точность поиска материалов в полнотекстовых электронных библиотеках.

Ещё одной особенностью разработанных программных средств электронных библиотек является их интеграция с системой автоматизации библиотек ИРБИС64. Интеграция состоит в одновременном поиске в коллекциях полнотекстовой электронной библиотеки и в базах данных ИРБИС из единого пользовательского интерфейса. При этом результаты поиска в базах ИРБИС также подвергаются семантической обработке, что повышает точность поиска по библиографическим описаниям.

Разработанные полнотекстовые электронные библиотеки 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 web-интерфейс.
- Демонстрационная версия полнотекстовой электронной библиотеки доступна в сети Интернет по адресу http://elib.isa.ru.

Литература

1. Gennady Osipov, Ivan Smirnov, Ilya Tikhomirov. Application of Linguistic Knowledge to Search Precision Improvement. // Proceedings of 4th International IEEE conference on Intelligent Systems 2008. Volume 2. – P. 17–2 – 17–5.
2. Тихомиров И.А, Смирнов И.В. Интеграция лингвистических и статистических методов поиска в поисковой машине Exactus // Труды международной конференции Диалог'2008. – С. 485–491.
3. Золотова Г. А., Онипенко Н. К., Сидорова М. Ю. Коммуникативная грамматика русского языка. – М. 2004. – 544 с.
4. Осипов Г. С. Приобретение знаний интеллектуальными системами: Основы теории и технологии. – М.: Наука, Физматлит, 1997. – 112 с.
5. 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.