УДК 002.513.5

ИНФОРМАЦИОННЫЙ ПОИСК В ЭЛЕКТРОННЫХ БИБЛИОТЕКАХ

Попов А.В., Шабанов В.С.,
научный руководитель д-р физ.-мат. наук Добронец Б.С.
Сибирский федеральный университет

Введение

Широкое распространение персональных компьютеров и быстрое развитие телекоммуникационных сетей привели к интенсивному росту электронных библиотек. В настоящее время многие статьи и книги публикуются непосредственно в электронном формате, а бумажные источники информации активно оцифровываются. Одной из наиболее актуальных потребностей, которые возникают при работе с большим количеством электронных источников, очевидно, является быстрый поиск с предоставлением релевантных и точных результатов, т.е. необходимо найти не только сам документ, но и конкретную страницу или абзац в этом документе.

Особую ценность представляет собой контекстный поиск, который учитывает тематику текста при выдаче результатов на поисковый запрос. Это в большей степени важно при поиске научной литературы, так как многие термины имеют различное значение в различных науках и релевантность результатов существенно зависит от определения контекста, в котором необходимо найти информацию.

На текущий момент электронная литература представлена во множестве различных форматах, что усложняет процесс обработки. Наибольшую популярность приобрели форматы PDF и DJVU. Особенностью данных форматов является их инвариантность при изменении средств просмотра и операционной системы у пользователя. Помимо текстовой информации PDF и DjVu могут содержать и графические элементы, что очень удобно, в частности для технической литературы. В наше время разработано несколько систем управления электронными библиотеками.

1. Существующие информационно-поисковые системы

Из этого множества библиотечных систем по функциональности, известности, использованию и русской локализации, выделяются три: DSpace, EPrints и Greenstone. Все эти системы широко применяются по всему миру и в России в частности. Так например библиотека Уральского государственного университета использует DSpase, библиотека правительства республики Марий Эл - Greenstone.

Каждая из вышеперечисленных систем разрабатывается уже не один год, для их усовершенствования работают десятки программистов. Информационно поисковые системы вошли в нашу повседневную жизнь и сложно представить использование персонального компьютера без их помощи. Однако каждая из перечисленных систем имеет как свои достоинства, так и недостатки. Ни в одной из рассмотренных АБИС не предусмотрена поддержка формата DjVu, который очень популярен в настоящее время и число книг данного формата (особенно технической литературы) продолжает постоянно увеличиваться. АБИС EPrints и DSpace не являются кроссплатформенными, для их установки нужна Unix-подобная Ос, и доступ к библиотеке со стороны конечного пользователя осуществляется по сети либо через интернет по средствам интернет браузера. Для домашней библиотеки и небольшой общественной библиотеки это является существенным недостатком, поскольку около 90 процентов компьютеров в мире находится под управление ОС Windows.

Отдельный интерес представляла бы система для работы с электронными библиотеками, которая позволяла бы обеспечить совместную работу в некотором научном сообществе, где у участников имеются общая литература.

2. Информационный поиск

В настоящее время работу большинства поисковых систем можно упрощенно представить в виде следующих шагов (рис1):

1. Пользователь формирует поисковый запрос.
2. Запрос передается поисковой системе.
3. Система производит поиск ресурса в базе данных на основе этого запроса.
4. Отображение списка результатов, наиболее подходящего данному поисковому запросу.

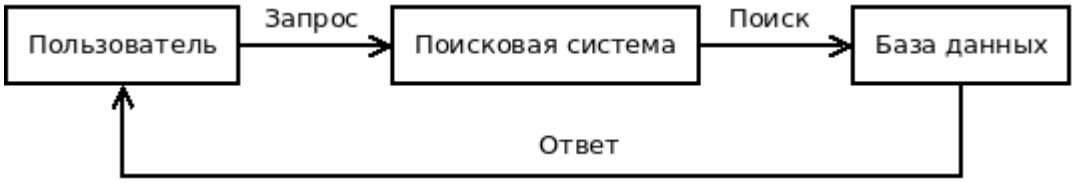

Рис 1. Схема поиска

Существует три основных подхода к организации поиска, каждый из которых обладает своими преимуществами и недостатками:

1. Булева модель. Относительно проста в реализации, позволяет обрабатывать большие объемы данных. Отсутствие контекстных операторов. Невысокая эффективность поиска.

2. Векторно-пространственная модель. В классическом виде малопригодна для обработки больших объемов данных в связи с применением массивов высокой размерности.

3. Вероятностная модель. Характеризуется низкой вычислительной масштабируемостью и необходимостью обучения системы.

Помимо приведенных выше, существуют и другие подходы, но их эффективность пока не достигла того уровня, который бы способствовал их распространению.

3.1. Модель

Содержание документов и запросов в информационно-поисковых системах обычно описывается некоторыми наборами терминов, представляющих собой отдельные слова или словосочетания.

Часто для характеристики терминов используются их веса, которые отражают предполагаемую важность каждого из терминов.

Документ F можно представить как покрытие:

$$\boldsymbol{F} = \bigcup_{\boldsymbol{i}} \boldsymbol{F_i}$$

где $\boldsymbol{F_i}$ – фрагменты документа. Каждому фрагменту $\boldsymbol{F_i}$ сопоставляется набор $\boldsymbol{K_i} = \{(\boldsymbol{K_l}, \boldsymbol{z_l}) | \boldsymbol{l} = \mathbf{1} \ldots \mathbf{ni}\}$, где $\boldsymbol{K_l}$ – ключевое слово и $\boldsymbol{z_l}$ – его значимость. Значения

z_l принадлежат некоторому упорядоченному множеству, например, {0,1,2,3,4,5} или {низкая, средняя, высокая}.

Далее будем предполагать, что выполняются следующие соотношения: если некоторый фрагмент F_i содержится во фрагменте F_j, то набор ключевых слов K_j содержится в K_i. Более того, будем считать, что фрагменту $F_i \cap F_j$ соответствует набор ключевых слов $K_i \cap K_j$. Таким образом, мы определили правило наследования ключевых слов.

Решение о выдаче того или иного документа принимается в результате сравнения наборов терминов, относящихся соответственно к документам и запросам. Вниманию пользователя предлагаются те документы, наборы терминов которых совпадают с наборами терминов запросов. [2]

3.2. Индексация

Для осуществления поиска среди множества электронных документов в настоящее время наиболее эффективной и распространенной является схема с применением индекса документов. Это значит, что перед осуществлением поиска необходимо произвести индексацию документов. При добавлении нового электронного документа информация о нем добавляется в индекс.

Определение 1 *Индексация – это процесс построения поискового индекса.*

При индексации необходимо получить доступ к информации в текстовом виде. Особую сложность на данном шаге представляют собой случаи когда информация представлена в виде изображения. Для ее преобразования в текстовый формат можно использовать метод оптического распознавания, однако результаты будут не всегда точны.

Обычно выделяют три способа индексации:

1. Ручной, при котором человек(индексатор) выделяет наиболее значимые слова и указывает область их действия. Достоинством такого способа является точность созданного индекса, а недостатком – высокая трудоемкость.

2. Автоматический. Программа обрабатывает текст, выделяя ключевые слова и определяя их область действия. В настоящее время существует множество способов выделения ключевых слов. Например, можно использовать частотный анализ, учитывая количество использования слова в тексте. Для исключения попаданий малозначимых слов в индекс примется стоп-словарь, куда заносятся слова, которые следует пропускать во время индексирования. В качестве области действия слов можно использовать типографические области текста, такие как параграф, страница или абзац.

3. Смешанный. Объединяет перечисленные выше способы, позволяя человеку корректировать результаты автоматической индексации. Данный способ является наиболее эффективным т.к. дает возможность охватить хоть сколько много информации и точно выделить основные ключевые слова.

3.3. Представление результатов

Часто в ответ на поисковый запрос система предоставляет достаточно большое множество документов, которые в разной степени подходят пользователю. Способ их отображения во многом связан с удобством пользования системой. Обычно результаты отображаются в виде списка.

Определение 2 *Ранжирование – процесс, при котором поисковая система выстраивает результаты поиска в определённом порядке по принципу наибольшего соответствия конкретному запросу.*

Ранжирование доступно не для всех моделей поиска (например недоступно для булевой). [3] Ранжирование в текстовых и гипертекстовых документах существенно различается. В модели, представленной выше, для ранжирования текстовых документов можно использовать веса ключевых слов. Ранжирование гипертекстовых документов возможно также по свойствам обуславливаемым сетевой структурой.

Поскольку в одной книге может быть несколько страниц, дающих ответ на запрос пользователя, из них необходимо выбрать наиболее релевантную. Чтобы показать наибольшее количество книг, в поисковой выдаче для отдельной книги показывается как правило только одна страница, в редких случаях в поисковой выдаче могут быть показаны несколько страниц из одной книги.

Так же важен способ отображения. Для пользователя было бы удобно видеть результат поиска непосредственно в документе.

Заключение

Развитие контекстного поиска в полнотекстовых электронных библиотеках помогает уменьшить затрачиваемое время на поиск нужной информации, что является несомненным плюсом для потребностей современного общества. Так же необходимо расширить возможность поиска в большем количестве форматов поскольку с появлением новых устройств увеличивается и количество форматов электронных книг. Для улучшения качества поиска может быть полезным разработка предметных тезаурусов и применение поиска по словарю синонимов. [1] Еще одной перспективной ветвью развития системы является разработка приложения для web, тем самым открыв доступ к своей библиотеке из любого места где есть доступ к интернету.

Все это поможет реализовать более быстрый контекстный поиск и сделает использование поисковых систем более удобным для пользователя.

Литература

1. *Когаловский М.Р.* Перспективные технологии информационных систем. ДМК Пресс, Москва, 2003.
2. *Добронец Б.С., Мамедов А.А.* Разработка информационно-поисковых систем для полнотекстовых библиотек / Повышение качества высшего профессионального образования: Материалы Всероссийской научно-методической конференции с международным участием: в 2ч. 2007 Ч. 1. С. 252–256.
3. *Ландэ Д.В, Снарский А.А.* Интернетика: Навигация в сложных сетях: модели и алгоритмы. Книжный дом «ЛИБРОКОМ», Москва, 2009.