Практический опыт применения поисковых технологий для библиотечных фондов

Набатчиков Дмитрий Евгеньевич, ведущий руководитель проектов, Корпорация ЭЛАР, dnabatchikov@elar.ru

В докладе рассматриваются потребности библиотек по использованию новых технологий поиска по электронным каталогам и коллекциям, а также практический опыт Корпорации ЭЛАР по внедрению в российских библиотеках поисковых технологий, таких как интеграция библиотечных ресурсов, полнотекстовый поиск, семантический поиск, визуальный поиск, статистика использования фондов. Особое место в докладе уделено методам создания тематических словарей (тезаурусов).

Ключевые слова: электронный каталог; полнотекстовый поиск; интеллектуальный поиск; тезаурус; electronic catalogue; full text search; text mining; thesaurus.

1. Введение

Все библиотеки стремятся предоставлять читателям максимум информации, содержащейся в фондах библиотек. Приветствуются любые методы облегчения работы читателей с этими фондами. Бумажные каталоги и краткие аннотации – это то, что облегчало работу в библиотеке до 21 века. Электронные каталоги, оцифрованные издания, поиск по тексту оцифрованных изданий – это то, что облегчает работу в 21 веке.

Но нужно идти дальше, и новые технологии не стоят на месте: анализ текста изданий, активные подсказки читателю, нахождение ответов на вопросы за долю секунды, а в идеале и «угадывание» желаний читателей. Ну и, конечно, всё это должно быть в удобном приятном интерфейсе.

2. Практический опыт

В течение 2012-2013 гг. был выполнен ряд проектов, предоставляющих читателям новые возможности поиска информации. Самые значимые из этих проектов – для Российской государственной библиотеки (РГБ) и Государственной публичной исторической библиотеки (ГПИБ). В РГБ – система поиска по библиотеке диссертаций, в ГПИБ – единый электронный каталог.

Основным инструментом повышения качества поиска информации в библиотеке мы видим систему поиска, интегрированную с библиотечными системами, которые есть в библиотеке (иногда их больше одной), любыми другими электронными ресурсами библиотеки (включая веб-сайт), с внешними подписными изданиями, и с возможностью интеграции с другими информационными системами для библиотек на межотраслевом/межрегиональном уровне. Общая схема приведена на рисунке 1.

Какими же функциями обладает система поиска, которые позволяют помочь читателю и сильнее раскрыть информационный потенциал, заложенный в электронных ресурсах библиотеки?

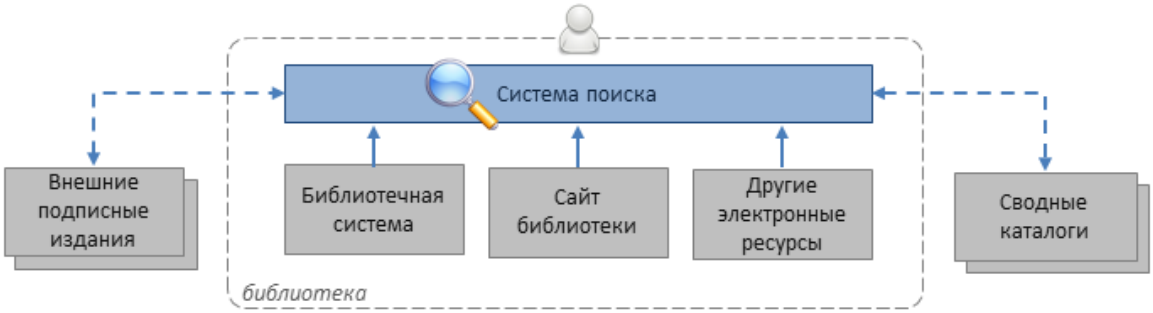

Рисунок 1

2.1. Мгновенный поиск по большим массивам

Основным компонентом системы поиска является модуль полнотекстового поиска (full text search), выполняющий предварительное индексирование электронных ресурсов (аналогично системам интернет-поиска). Электронные ресурсы библиотеки могут включать в себя книги, периодические издания, изображения, видео, аудио и пр. Интеграция с внешними подписными изданиями может строиться как на основе предоставления электронных ресурсов, так и перенаправления поисковых запросов в режиме реального времени (федеративный поиск, federated search).

Время поиска изданий по библиографической информации составляет не более 1 секунды, по тексту изданий – не более 2 секунд.

Удобство использования обеспечивается заменой множества поисковых полей (название, автор, год и пр.) на одно универсальное поле. Всю необходимую информацию можно искать, вводя ее в это поле, например: *пушкин болдинская осень 1830*. Система в этом случае сама определяет автора, год, название.

2.3. Визуализация поиск

Большинство современных поисковых систем позволяют фильтровать полученные результаты поиска, сужая область поиска, например: по месту издания или типу издания. Если на странице с результатами поиска есть список авторов найденных книг и их количество, то, скорее всего, это т.н. фасетный поиск (faceted search).

В нашей системе поиска мы упростили использование фасетов (фильтров), сделав их более наглядными. Они представляют собой интерактивные диаграммы и графики. Например, в библиотеке диссертаций это позволяет:

1) наглядно увидеть, как менялся со временем интерес к теме квантовых компьютеров или любой другой теме;

2) быстро найти диссертации по интересующей теме, которые защищались в двух разных научных направлениях одновременно (например, физика и сельское хозяйство).

2.4. Нечеткий поиск

Нечеткий поиск (fuzzy search) – это поиск информации, в которой есть орфографические ошибки. Это особенно актуально для текстов, полученных методом оптического распознавания (OCR). Наши решения позволяют находить слова с одной и двумя ошибками. Дополнительное применение возможностям нечеткого поиска мы нашли в поиске на старорусском языке (см. ниже).

2.5. Рекомендации

В любых интернет-магазинах можно увидеть рекомендации, вроде «*популярные товары*», «*с этим товаром также покупают*». Разработанная нами система поиска выполняет похожие функции, только для библиотек: «*читайте также*», «*с этой книгой также читают*», «*похожие книги*».

2.6. Использование словарей (тезаурусов)

Для того чтобы наделить систему поиска б***о***льшим интеллектом, нужно создавать функции, использующие семантику языка. Любые такие функции требуют составления тезаурусов (словарей) и правил, это неизбежная пока стадия обучения компьютера человеческим понятиям. В нашей работе мы составляем и используем различные словари для повышения качества поиска в библиотеках.

2.6.1. Двуязычный поиск русский-немецкий

В одном из наших проектов мы научили систему двуязычному поиску на русском и немецком языках. Это была довольно простая реализация, нам не пришлось составлять двуязычный словарь. Поисковый запрос пользователя отправляется во внешний сервис переводов *Яндекс.Перевод* или *Google Translate*, поисковая система выполняет два параллельных запроса (один на русском языке, второй – на немецком) и отображает результаты в едином списке.

2.6.2. Двуязычный поиск русский-старорусский

В варианте двуязычного поиска русский-старорусский применить ни один сервис переводов не получилось, т.к. старорусский не представлен среди вариантов языков. Под старорусским понимается дореформенная (дореволюционная) орфография русского языка, например: статскій совѣтникъ.

Проблема дореформенной орфографии в том, что она менялась на протяжении XVIII-XX веков, поэтому книги того периода содержат разную орфографию.

Тем не менее, перевод слов со старой орфографии в новую в подавляющем большинстве случаев может быть описан строгими правилами, вроде поменять букву *ять* (Ѣ) на современную Е, *фита* (Ѳ) – на современную Ф, убрать твердый знак (Ъ) в конце слов после согласных и т.д. Поэтому одним из вариантов решения проблемы может быть преобразование слов к современной орфогра-

фии во время полнотекстового индексирования массива изданий. Но сохранение исходного написания слова тоже нельзя потерять, поэтому размер полнотекстового индекса будет больше при сохранении обоих вариантов каждого слова (если система поддерживает такую возможность).

Но всё это будет работать, только если читатель набирает свой запрос, используя современную орфографию. При использовании старой орфографии не будет учитываться морфология слов (другие падежи/склонения/спряжения), т.к. поисковые системы, как правило, не поддерживают дореволюционную морфологию.

Использование современным читателем современной орфографии при работе в библиотеке представляется наиболее вероятным, чем использование дореволюционной орфографии, поэтому описанный выше подход позволяет реализовать поиск старых изданий без использования громоздких тезаурусов, но с ограничениями.

В нашем проекте мы решили не мириться с этими ограничениями и составили собственный «двуязычный» словарь русский-старорусский. Мы проанализировали список слов, содержащихся в нашем массиве изданий, и автоматически составили список старых слов со всеми встречающимися словоформами. Этот список (словарь) используется как словарь синонимов при поиске. С какой бы орфографией читатель ни набирал свой запрос (и с любыми словоформами), будут найдены все требуемые издания. Чтобы не терять даже те словоформы, которые отсутствуют в текстах, применяется нечеткий поиск (см. выше), так как большинство слов со старой орфографией отличаются от современного написания не более чем в двух буквах. Этот подход работает на зафиксированном массиве изданий. При добавлении новых изданий, словарь нужно обновлять. Это выполняется автоматически.

2.6.3. Терминологические тезаурусы

Для системы поиска по диссертациям составлялись тематические тезаурусы, которые использовались для автоматического выделения слов в текстах работ. Найденные термины используются читателями для фасетного поиска, то есть для фильтрации найденных работ по набору используемых терминов. Всего составлено 10 тезаурусов (юриспруденция, экономика, химия, биология и др.), содержащих в общей сложности более 600 тыс. терминов.

Что представляют собой эти тезаурусы? Набор слов и словосочетаний с возможными синонимами, например: болезнь лайма = лайма болезнь = лайм-боррелиоз = клещевой боррелиоз = боррелиоз клещевой. Гипонимы и гиперонимы (отношения общее-частное) в тезарус не вносились.

Как составлялись эти тезаурусы? Была разработана методика, состоящая из нескольких этапов: (1) подбор специализированных справочников, толковых словарей, словарей переводов; (2) составление общего списка слов; (3) очистка тезауруса. Первый этап – это набор организационных мероприятий с участием экспертов в своих областях и юристов для получения исходных словарей, отку-

да берутся термины. Второй этап – автоматическая конвертация в формат нашего тезауруса. Третий этап – самый интересный.

Очистка тезауруса нужна главным образом для того, чтобы убрать из него термины, не представляющие большого интереса для тех, кто работает с массивом диссертаций. Это три основных категории терминов: общеупотребительные слова (даже если они одновременно являются терминами в каких-то областях), широко используемые термины (например, термин «дифференцирование» в технических областях или «переговоры» в политике), обобщающие термины (например, убираем термин «зонд», остаются «атомный зонд», «узкоапертурный зонд» и др.)

Для очистки тезауруса от общеупотребительных слов использовался достаточно большой словарь таких слов. Для очистки тезауруса от широко используемых и обобщающих терминов был проведен частотный анализ *на имеющемся массиве диссертаций*. Решение об удалении каждого термина (не общеупотребительного) принимал эксперт в соответствующей области науки.

Интересным побочным эффектом работы с терминологическими словарями стало определение близости наук друг к другу на основе пересечения их словарей, например, политика и юриспруденция «похожи» на 27%, а политика и химия – всего на 3%.

3. Заключение

Опробованные нами решения для поиска информации в библиотеках позволяют предоставить читателям разные инструменты работы с библиотечными фондами, помогая находить то, что раньше было трудно найти, позволяя сильнее раскрыть информационный потенциал библиотек.

Особенно перспективными представляются любые решения, связанные с семантикой, анализом текстов изданий, так как они позволяют углубиться в смысл написанного текста. А составление всевозможных тезаурусов – это необходимый шаг к пониманию текста компьютером.