Совершенствование системы поиска в Электронной библиотеке Республики Карелия

© А.Г. Марахтанов

Петрозаводский государственный университет
marahtanov@petrsu.ru

Аннотация

Рассматриваются подходы к совершенствованию поисковых механизмов в Электронной библиотеке Республики Карелия. Обсуждаются вопросы индексирования хранящихся документов, применения механизмов поиска и ранжирования согласно релевантности запросу документов на основе векторной модели информационного поиска, реализации поиска по авторам, географическим объектам, персоналиям. Ожидается, что внедрение новой системы поиска приведет к повышению эффективности использования ресурсов электронной библиотеки в учебной, научной и практической деятельности посетителей библиотеки.

В 2004 году Региональным центром новых информационных технологий Петрозаводского государственного университета (ПетрГУ) была разработана и до настоящего времени активно используется Электронная библиотека Республики Карелия (ЭБ РК). Данная информационная система содержит более 1500 полнотекстовых изданий по различным областям знаний, многие из которых используются в учебном процессе студентами и сотрудниками ПетрГУ [1].

Каждое издание в ЭБ РК снабжено набором полей – метаданных, содержащих краткое описание, а также информацию об авторах данного ресурса, его названии, годе и месте издания, ключевых словах и т. п. Эти поля формируются в соответствии со стандартом метаописаний Dublin Core. На основании полей метаданных осуществляется поиск ресурсов в библиотеке.

Несмотря на востребованность ЭБ РК в учебной, научной и практической деятельности, выявлен целый ряд проблем с организацией поиска и навигации в ней, что понижает эффективность использования библиотеки в целом [2].

Предложен целый ряд мер, направленных на совершенствование поисковых и навигационных механизмов, в т. ч.:

- реализация полнотекстового поиска в ЭБ РК;
- увеличение скорости обработки поискового запроса;
- ранжирование результатов поиска по релевантности запросу пользователя;
- реализация новых видов поиска, таких, как поиск по авторам, географическому объекту, терминам тезауруса и т. п.

Механизм полнотекстового поиска позволит находить все произведения (и место в тексте), содержащие упоминание какого-либо персонажа, географического места, названия или термина, даже если они не встречаются в метаописании ресурса. Возможность осуществления такого вида поиска будет полезна, например, для изучения Республики Карелия (ее истории, фольклора), в исследовательской деятельности, при построении генеалогических деревьев и т. п.

Реализацию полнотекстового поиска в ЭБ РК затрудняют следующие обстоятельства:

- ресурсы, представленные в библиотеке, хранятся в различных форматах: наборы jpeg-изображений, .doc и .rtf файлы, .pdf-файлы, html-страницы, текстовые файлы (.txt);
- в библиотеке представлены издания на различных языках: русский (в том числе книги на древнерусском языке, изданные до 1917 года), английский, языки финно-угорской группы (финский, карельский, вепсский);
- файлы изданий могут храниться на различных серверах (сервера ПетрГУ, Национальной библиотеки РК, ссылки на сайты, хранящиеся на других региональных серверах).

Очевидно, что для реализации полнотекстового поиска необходимы извлечение текстовой информации из файлов изданий (другими словами – извлечение из файлов набора слов), а также последующая обработка и хранение этой текстовой информации с целью сокращения времени на дальнейшее обслуживание поисковых запросов пользователей.

Способ извлечения массивов текста для дальнейшей обработки зависит от формата файла, в котором представлен ресурс.

- .txt файлы; открытие файла в режиме чтения позволяет работать с ним, как с массивом текста;
- .html файлы; считывание содержимого файла, вырезание всех html-тегов, замена специальных символов (html-мнемоник);
- .doc и .rtf файлы; использование специальных программ, позволяющих консольно преобразовывать файлы, подготовленные процессором Micro-

Труды 12[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

soft Word, в txt или html-формат; существуют решения этой проблемы, такие, как интерфейс Docvert (http://holloway.co.nz/docvert), которые позволяют осуществлять подобные преобразования на Linux-серверах (ЭБ РК функционирует на платформе с OS OpenSuSE под управлением веб-сервера Apache);

- .jpeg файлы; использование программ оптического распознавания (OCR); существует большое количество подобных программ, однако не все они подходят для распознавания ресурсов из ЭБ РК; некоторые плохо работают с текстами на русском (и, особенно, древнерусском) языке; некоторые не позволяют использовать консольный вызов и т. п.; рассматривалось большое число OCR-программ (Tesseract от Google, OmniPage, ABBYY FineReader и пр.), наилучшие результаты были получены при использовании ABBYY FineReader; для исследования использовался серверный вариант программы (под OS типа Linux);
- .pdf файлы; преобразование pdf-файла в набор jpeg-изображений и последующее распознавание каждого из них.

Массивы текста, извлеченные из файлов каждого из изданий, могли бы непосредственно использоваться для поиска, однако такой вид поиска (его называют «прямой поиск») будет медленным и неэффективным, ведь для каждого запроса необходимо будет осуществлять поиск подстрок, соответствующих запросу пользователя, во всем массиве текстов. Для сокращения времени поиска осуществляется построение так называемого «обратного индекса», то есть списка всех встречающихся слов, с указанием изданий (и позиций в издании), где слово встретилось. Поиск по такому индексу осуществляется значительно быстрее прямого.

Для построения обратного индекса из текстового массива необходимы:

- выделение из текста отдельных слов без знаков препинания, переносов и пр. (графематический анализ) [3];
- лемматизация, или преобразование выделенных слов к начальной форме (морфологический анализ) [3];
- занесение слов, а также позиций их вхождений, в базу данных.

Графематический и морфологический виды анализа затем должны применяться и к каждому запросу пользователя.

Кроме полного текста документа в индекс добавляются также термины из метаописаний ресурса (названия, описания, списка ключевых слов и пр.). Связь термина из метаописания и документа помечается отдельно.

Для снижения размера индекса (и, следовательно, увеличения скорости обработки поискового запроса) возможно исключение из списка так называемых стоп-слов – слов, часто встречающихся и не определяющих семантику текста. Лемматизация также способствует снижению размерности.

Поиск по построенному индексу должен реализовываться в соответствии с некоторой выбранной моделью поиска. Примером такой модели может быть векторная модель [4].

Согласно данной модели, все документы коллекции и сам запрос представляются в виде векторов в n-мерном пространстве терминов, где n – общее число слов в индексе.

Значение координаты вектора может определяться различными способами. Чаще всего используют коэффициент TF*IDF, построенный на базе естественного статистического наблюдения: чем больше локальная частота термина в документе (TF) и больше «редкость» (т. е. обратная встречаемость в документах) термина в коллекции (IDF), тем выше вес данного документа по отношению к термину.

В общем случае в качестве TF можно взять отношение числа упоминаний слова в документе к общему количеству слов в документе. Например, если слово «Карелия» встречается 7 раз, а всего издание содержит 1000 слов, тогда TF будет равна 0,007. Для вычисления IDF можно использовать отношение количество книг, содержащих слово, к общему числу книг. Например, «Карелия» встречается в 50 документах из 2000, тогда IDF равна 40. Иногда полученное значение логарифмируют.

При представлении документов и запроса в виде векторов задача поиска релевантности сводится к задаче поиска близости между векторами. Мерой такой близости может служить, к примеру, косинусная метрика, то есть косинус угла между векторами:

$$\cos\gamma = \frac{\sum_i W_{di} W_{qi}}{\left(\sum_i W_{di}^2 \sum_i W_{qi}^2\right)^{1/2}},$$

где W_{qi} – i-я координата вектора запроса, W_{di} – i-я координата вектора документа. Нормирование длин векторов позволит упростить формулу (знаменатель = 1).

При подсчете косинусной меры используются только координаты, соответствующие терминам, общим для запроса и документа, что снижает время расчета.

Чем ближе значение косинуса угла к 1, тем выше релевантность документа относительно запроса пользователя. На основании этого значения можно осуществлять ранжирование найденных документов, то есть выводить их в порядке уменьшения релевантности.

При реализации векторной модели в ЭБ РК предполагается модификация простейшей векторной модели, а именно:

- введение коэффициентов для повышения значимости слова, если оно встретилось в основных полях метаданных (заголовке, аннотации, списке ключевых слов и т. п.);
- нормирование по длинам документа, запроса и полей метаданных при вычислении частоты, для избавления от негативного влияния на результат поиска очень коротких и очень длинных документов.

При практической реализации векторной модели информационного поиска необходимо осуществлять

вычисление векторов документов (и их нормирование) в момент добавления документов в библиотеку, а при обработке поискового запроса – только вычислять вектор запроса и косинусную меру.

Кроме полнотекстового предполагается реализация и других видов поиска: по автору, географическому объекту, персонажу. Для их реализации также будет использован обратный индекс, в котором значимые термины (географические объекты и персонажи) будут маркированы вручную.

Поиск по авторам будет осуществляться на основании совпадения запроса с соответствующим полем метаданных. Предполагается, что будет отображаться список всех авторов изданий библиотеки (по алфавиту). При нажатии на ФИО автора будет отображаться страница со списком его работ.

Существенной проблемой при реализации подобного механизма поиска является отсутствие единого формата при заполнении поля «авторы». Встречается более 10 различных способов написания ФИО автора, например:

- Одоевский Владимир Фёдорович (Фамилия Имя Отчество);
- Бохонский Д. О. (Фамилия И. О.);
- Е.П. Шемилина (И.О. Фамилия);
- Т.Ашер (И.Фамилия);
- Бергштрессер К. (Фамилия И.);
- Boitchenko Larissa (Familia Imja);
- и т. п.

Таким образом, для реализации поиска по авторам, необходимо:

- выбрать единый формат для представления данных об авторе;
- привести уже добавленные в базу авторов строки к единому формату;
- модифицировать интерфейсы ввода и редактирования изданий в ЭБ РК с целью упрощения ввода авторов в едином формате и автоматической проверки введенных модераторами данных.

В качестве единого формата представления данных об авторе был выбран следующий: «Фамилия Имя Отчество». Они должны записываться всегда в таком порядке, каждое в начальной форме (именительный падеж, единственное число), начинается с заглавной буквы. Разделителем служит одиночный пробел. При отсутствии данных допустимо указание инициалов вместо имени и отчества.

Для приведения уже накопленных данных к единому выбранному формату предполагается, во-первых, исправить явные ошибки, где это возможно, при помощи регулярных выражений (поиск по шаблону и замена) или транслитерации набранных латиницей имен, во-вторых, воспользоваться математическими мерами близости строк для группировки похожих ФИО авторов с целью упрощения дальнейшей ручной модерации.

Одной из наиболее часто используемых в подобных задачах мерой близости строк является расстояние Левенштейна (также известное как редакционное расстояние или дистанция редактирования) [5, 6]. Это мера разницы двух последовательностей символов (строк) относительно минимального количества операций вставки, удаления и замены, необходимых для перевода одной строки в другую.

Одним из минусов использования расстояния Левенштейна в нашей задаче является то, что при перестановке местами слов или частей слов получаются сравнительно большие расстояния. А в нашем случае перестановка слов (например, «Вихавайнен Тимо», «Тимо Вихавайнен») будет встречаться достаточно часто. И даже расстояние Левенштейна – Дамерау, в котором введена дополнительная операция – перестановка соседних символов (при условии, что эти символы являются смежными в обоих строках), не будет вполне адекватной мерой для данных случаев.

Использования расстояния Хемминга также не рекомендуется, поскольку в нашем случае будут сравниваться строки различной длины.

Другим возможным подходом к решению поставленной задачи является использование N-граммных расстояний, основанных на вычислении меры близости по количеству общих подстрок фиксированной длины [7]. Эти подстроки называются N-граммами.

Для любого слова или фразы могут быть построены N-граммы различной длины. Например, фамилии Иванов соответствуют 5 биграмм («Ив», «ва», «ан», «но», «ов»), 4 триграммы («Ива», «ван», «ано», «нов») и т. д. Наличие общих N-грамм повышает оценку близости строк, причем, чем больше порядок совпавшей N-граммы (число N), тем больше должна быть эта оценка.

После приведения списка авторов к единому формату необходимо внедрение новых интерфейсов ввода метаинформации об издании, которые исключали бы возможности некорректного ввода и при этом упрощали бы ввод за счет возможности выбора из уже внесенных в систему авторов.

Практическая реализация подходов, описанных в данной работе, в частности, извлечение текстов из файлов ресурсов библиотеки, их индексирование, программная реализация механизмов поиска на основе векторной модели, возвращающих ранжированный список документов, релевантных запросу, реализация поиска по авторам, географическим объектам, персоналиям существенно улучшат существующие поисковые возможности электронной библиотеки.

В целом подобное совершенствование системы поиска позволит повысить эффективность использования ресурсов электронной библиотеки в учебной, научной и практической деятельности посетителей библиотеки.

Литература

[1] Рузанова Н.С., Насадкина О.Ю., Байтимиров Л.З., Гушкалова А.Г., Марахтанов А.Г. Электронная библиотека Республики Карелия// Труды XIV Всерос. науч.-метод. конф. Телематика'2007 (С.-Петербург, 18 – 21 июня 2007 г.). – С.-Пб., 2007. – Т. 2. – С. 390-391.

[2] Байтимиров Л.З., Власова А.Г., Марахтанов А.Г., Насадкина О.Ю., Фотина Е.В. О проблеме информационного поиска в Электронной библиотеке Республики Карелия// Материалы науч.-метод. конф. «Университеты в образовательном пространстве региона: опыт, традиции и инновации», Петрозаводск, 16 – 17 февраля 2010 г. – Петрозаводск, 2010. – С. 71-74.
[3] Ножов И.М. Морфологическая и синтаксическая обработка текста (модели и программы). – М., 2003. – http://www.aot.ru/ docs/Nozhov/msot.pdf.
[4] Солтон Дж. Динамические библиотечно-поисковые системы. – М.: Мир, 1979.
[5] Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов// Докл. АН СССР. – 1965. – С. 845-848.
[6] Гасфилд Д. Строки, деревья и последовательности в алгоритмах: Информатика и вычислительная биология / Пер. с англ. И.В. Романовского. – С.-Пб.: Невский Диалект; БХВ Петербург, 2003.
[7] Cavnar W.B., Trenkle J.M. N-gram-based text categorization//Proc. of Third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, 11 – 13 April 1994. – UNLV Publications/Reprographics, 1994. – P. 161-175.

Search system development in Digital library of Republic Karelia

A. Marahtanov

In the article approaches to perfection of search mechanisms in Digital library of Republic Karelia are considered. Questions of indexing of stored documents, applications of search engines and ranging according to relevance to inquiry of documents on the basis of vector model of information search, realization of search in authors, geographical objects, a personnel are considered. It is expected that introduction of new system of search will lead to effectivization of resource exploitation of electronic library in educational, scientific and practical activities of visitors of library.