

Оценка качества информационного поиска в слабо структурированных источниках на основе метаданных и базы знаний

Егошина А.А., Вороной А.С.
Донецкий национальный технический университет
postmaster@iai.donetsk.ua

Abstract

Yegoshina A.A., Voronoy A.S. Improving the efficiency of extracting information from poorly structured sources on the basis of metadata and knowledge base. The Proposed technology of increasing search effectiveness in not structured databases with Web-interface based on meta-data that shows database structure and data semantics. A model of meta-data for such databases and meta-data storage is proposed. Keeping meta-data and knowledge base on different servers allows to shorten request time. Results of experimental research on actual databases for different profiles of user operations were made. Experiments show that with real number of user requests to databases proposed technology has good response time and decreasing servers load. The more responses number grows, better proposed technology works.

Общая постановка проблемы

Основной проблемой при работе с распределенными и слабо структурированными информационными ресурсами является сложность точной формулировки запроса - подбора ключевых слов, которые предстоит искать в документах или базах данных.

В последнее время активное развитие получило направление в информационных технологиях, использующее стандарт метаданных, который позволяет пользователям совершать поиск в большом количестве таблиц баз данных и уверенно определять местонахождение интересующей информации.

Метаданные определяют ортогональный основному уровню описания информации (который формируется такими понятиями, как классы, типы данных и др.) уровень описания свойств [1]. Использование метаданных, в особенности семантических, позволяет эффективно решать такие задачи работы со знаниями как поиск, категоризация и рекомендация знаний.

Целесообразным представляется также применение методов и средств, разработанных в области искусственного интеллекта, а именно онтологий, которые позволяют производить автоматизированную обработку семантики информации.

Организация базы знаний для семантического поиска на основе онтологий в web-ориентированных реляционных базах данных

Для повышения эффективности информационного поиска в web-ориентированных реляционных базах данных в работе [2] предлагается подход к извлечению

информации из слабо структурированных источников на основе метаданных и базы знаний.

Использование базы знаний, как некоторого унифицированного интерфейса для решения задач над множественными неструктурированными источниками информации, освобождает пользователя от необходимости находить релевантные источники, задавать запросы к каждому из них по отдельности и вручную сопоставлять информацию из них.

В связи со значительным объемом ресурсов (таблиц базы данных) и их слабой структурированности в работе [3] предлагается хранить метаданные отдельно от ресурса в хранилище метаданных на отдельном сервере. Для организации базы знаний используется онтологический подход, который позволяет отразить семантику ресурса.

Первоначально база знаний содержала только онтологии, на основе которых для пользователей формировались ответы в виде html-фрагментов. Однако, в процессе эксплуатации системы обнаружилось, что такой подход требует значительных вычислительных ресурсов, т.е увеличивается нагрузка на сервер.

Поэтому было принято решение о расширении базы знаний шаблонами html-фрагментов, что незначительно увеличивает затраты памяти, но значительно уменьшает нагрузку сервера и снижает временные затраты на обработку запроса пользователя.

В работе [3] разработана структура БЗ, состоящей из двух компонентов: хранилища онтологий (SearchIndex) и хранилища html-шаблонов., как это показано на рисунке 1.

Для проведения исследований использовалась база данных компании Prometheus Research [4], которая предоставляет предприятиям и организациям, проводящим биомедицинские

исследования, средства управления web-ориентированными базами данных.

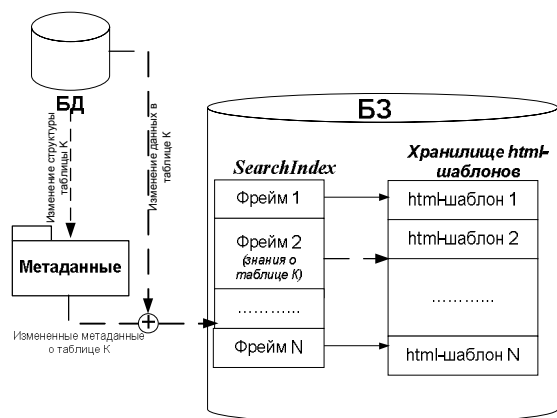


Рис. 1. Структура базы знаний

Формальная модель html-шаблона может быть представлена следующим образом:

$$F_{html} = \langle id_c, title_colum, link, type_data, V \rangle,$$

где id_c – код столбца;

$title_colum$ – название столбца;

$type_data$ – тип данных, содержащихся в столбце;

$link$ – ссылка на таблицу, содержащую столбец id_c ;

$title_table$ – название таблицы, содержащей данный столбец;

$V = \{v_1, v_2, \dots, v_s\}$ – множество значений v_i в столбце id_c .

Основной компонент базы знаний – это хранилище онтологий SearchIndex. Для оптимизации полнотекстового поиска в хранилище онтологий был разработан специальный язык представления онтологий в SearchIndex.

Знания представляются в виде пар :

$$\langle search_tag \rangle \langle Value \rangle,$$

где $search_tag$ – тип слота (записи);

$Value$ – значение слота.

Так количество слотов во фрейме не превышает двадцати, то для кодирования типов слотов достаточно трех символов: «!», «?», «>», которые никогда не встретятся в начале реальных исходных данных.

Как показали проведенные аналитиками Prometheus Research исследования, если ключевое слово из запроса встречается в названии столбца, то количество релевантной информации будет максимальной, и наоборот, если ключевое слово встретилось в значении, т.е. только один раз, то – минимальное.

Аналогичным образом с учетом результатов данных исследований были сформированы приоритеты для каждого типа слота: приоритет 1 - «?!>»; приоритет 2 - «?!>»; приоритет 3 - «?!>».

Экспериментальные исследования

С целью определения эффективности использования предложенного в работе подхода были проведены экспериментальные исследования на серверах компании Prometheus Research, результаты которых приведены в работе [2].

К основным характеристикам эксперимента относятся: количество пользователей, одновременно работающих с системой; типы выполняемых пользователями операций; количество обрабатываемых документов в индексе.

Для тестирования использовалась конфигурация, в состав которой входил один сервер индексирования (Xeon MP 3.0, 16 Gb RAM, 1Tb HDD) и один сервер баз данных (Xeon MP 2.2 8Gb RAM, 1Tb HDD). В ходе тестирования был произведен обход около 500 элементов (таблиц базы данных), размер которых составляет от 10 килобайт (КБ) до 100 МБ.

Результаты экспериментов показали, что при больших значениях числа пользовательских запросов к базам данных предлагаемая в данной работе технология обеспечивает комфортное для пользователей время обработки запросов (в случае 1000 запросов в секунду время обработки сокращается в 3 раза).

Кроме того, сокращается нагрузка на серверы системы (на сервер индексирования - в 10,84 раза, на сервер баз данных - в 9,26 раза), что позволяет снизить требования к аппаратному обеспечению. С ростом числа запросов преимущества предложенной технологии возрастают.

Для оценки качества информационного поиска по предложенной методике использовались критерии полноты R (отношение количества найденных при поиске релевантных значений к общему количеству значений, релевантных запросу) и точности P (отношение количества попавших в результат значений, релевантных запросу, к общему количеству выбранных значений).

Тестирование системы по данным критериям проводилось на секции онтологии «Аутоиммунные заболевания» по следующим группам:

- 1) группа 1 «Ревматоидные заболевания» содержит 180 концептов;
- 2) группа 2 «Полиэндокринные заболевания» - 127 концептов;
- 3) группа 3 «Аутоиммунные заболевания печени» - 85 концептов.

Одной из характеристик онтологии является выразительность, которая определяется степенью детальности описания вводимых в онтологии понятий (концептов).

Соответствие результатов поиска запросу определялось следующими категориями экспертов:

1) Viewer – эксперты с наиболее ограниченным уровнем доступа (медицинские сестры, сотрудники социальных служб);

2) DataEntry – эксперты, которым разрешен ввод и корректировка данных (лаборанты, рядовые научные сотрудники);

3) RИAccess – ведущие сотрудники лабораторий и научно-исследовательских центров, подписавшие соглашение о неразглашении информации, хранящейся в личных медицинских картах;

4) FullAccess – аналитики (инженеры по знаниям) Prometheus Research.

Результаты экспериментальных исследования для определения влияния выразительности онтологии на полноту и точность поиска по перечисленным выше группам приведены на рисунках 2 – 5.

Здесь $R1, P1$ – значения полноты и точности результатов поиска без использования метаданных и БЗ, а $R2, P2$ – с метаданными и БЗ.

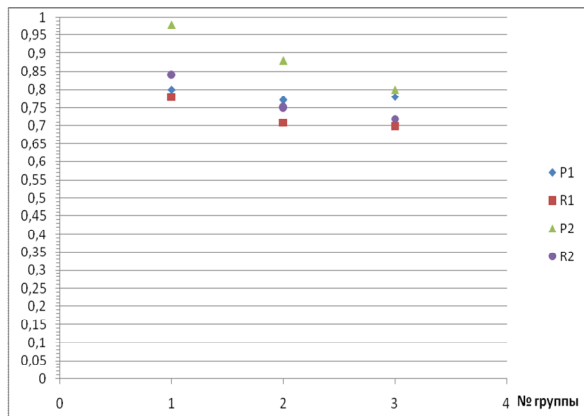


Рис. 2. Оценка точности и полноты результатов поиска категорией экспертов Viewer

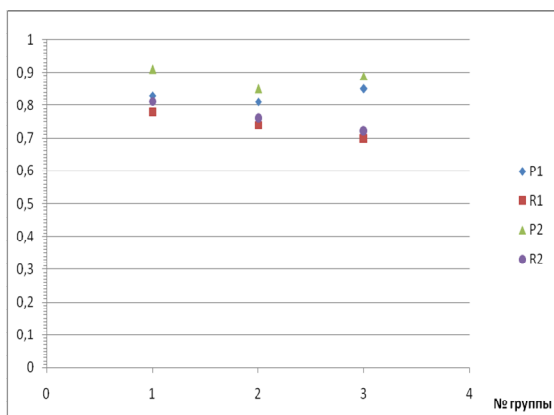


Рис. 3. Оценка точности и полноты результатов поиска категорией экспертов DataEntry

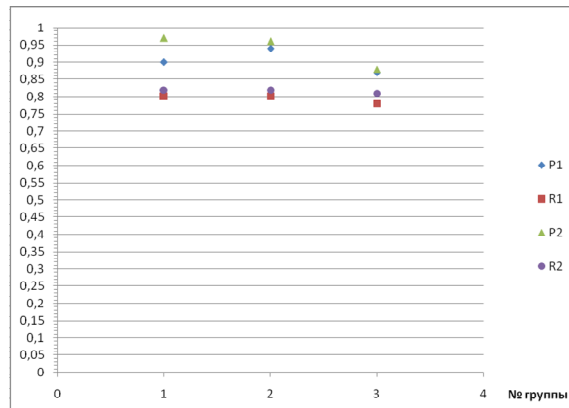


Рис. 4. Оценка точности и полноты результатов поиска категорией экспертов RИAccess

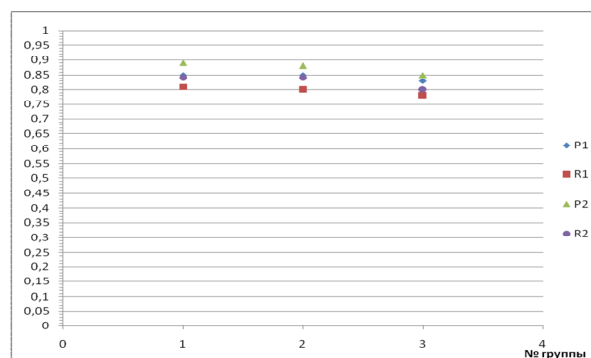


Рис. 5. Оценка точности и полноты результатов поиска категорией экспертов FullAccess

Результаты проведенных экспериментальных исследований позволили сделать следующие выводы:

1) при больших значениях числа пользовательских запросов к базам данных предлагаемая технология сокращает нагрузку на серверы системы (на сервер индексирования - в 10,84 раза, на сервер баз данных - в 9,26 раза), что позволяет снизить требования к аппаратному обеспечению;

2) значения критерия полноты увеличиваются для первой группы в среднем на 0,095, для второй – 0,047, для третьей – 0,02, что подтверждает положительное влияние большей выразительности онтологии для первой группы;

наилучшие результаты полноты и точности при использовании предложенного подхода были выявлены экспертами категорий с наибольшим числом участников Viewer и DataEntry, которые являются основными пользователями системы;

3) вследствие того, что эксперты категорий RИAccess и FullAccess имеют доступ к большему количеству информации, в том числе и к метаданным, и их запросы более точно и правильно сформулированы, преимущества внедрения предлагаемой технологии для этих экспертов менее значительны и практически не

отличаются для разных групп онтологий, т.е. слабо зависят от выразительности онтологий.

Заключение

В настоящей работе проведена оценка качества информационного поиска в слабо структурированных источниках на основе метаданных и базы знаний. Данная структура и схема хранилища метаданных сокращает нагрузку на серверы системы, что позволяет снизить требования к аппаратному обеспечению.

Предложенная структура базы знаний, содержащая не только онтологии предметной области, но и html-шаблоны, значительно снижает временные затраты на обработку запроса пользователя.

Использование представленной технологии информационного поиска с использованием базы знаний и метаданных, улучшает критерии полноты и точности поиска для различных категорий пользователей.

Литература

1. <http://www.w3.org/Metadata/>
2. Егошина А.А., Вороной А.С. Повышение эффективности извлечения информации из слабо структурированных источников на основе метаданных и базы знаний // Збірник наукових праць ДонНТУ. Серія «Інформатика, кібернетика і обчислювальна техніка». – Донецьк: ДонНТУ, 2011. - № 13(185). С. 44-47.
3. Егошина А.А., Вороной А.С. Организация базы знаний для семантического поиска на основе онтологий в web-ориентированных реляционных базах данных // Искусственный интеллект. – 2011. – № 1. С. 277-283.
4. <http://www.prometheusresearch.com>