УДК 004.3

А.А. Егошина, канд. техн. наук, доцент,
А.С. Вороной, ассистент
Донецкий национальный технический университет, г. Донецк, Украина
kafedra.sii@gmail.com

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

Предлагается технология повышения эффективности поиска в слабо структурированных базах данных с Web-интерфейсом на основе метаданных, представляющих структуру баз данных и базы знаний, описывающей семантику хранимых данных. Проведена оценка качества информационного поиска в слабо структурированных источниках на основе метаданных и базы знаний для различных категорий пользователей с использованием критериев полноты и точности. Результаты экспериментов показали, что при реальных значениях числа пользовательских запросов к базам данных предлагаемая технология обеспечивает комфортное для пользователей время обработки запросов, сокращение нагрузки на серверы системы и приемлемые для всех категорий пользователей полноту и точность. С ростом числа запросов преимущества предложенной технологии возрастают.

Ключевые слова: метаданные, база знаний, база данных, поиск, точность, полнота

Общая постановка проблемы

Основной проблемой при работе с распределенными и слабоструктурированными 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.

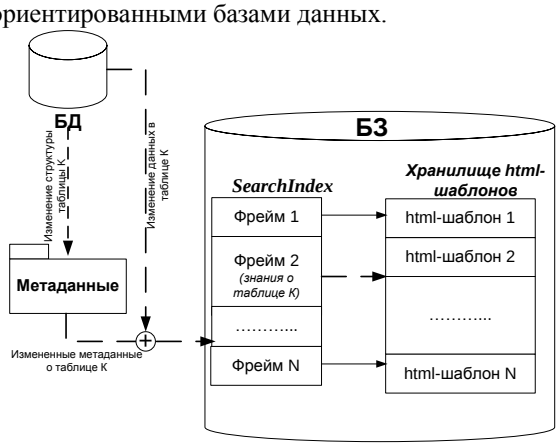

Рисунок 1 - Структура базы знаний

Формальная модель html-шаблона может быть представлена следующим образом:

$$F_{html} = < id\_c, title\_colum, link, type\_data, V >,$$

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;
$V=\{v_1, v_2, \ldots v_s\}$ – 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 :

<*search_tag*> <*Value*>,

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 опрашиваемых документов в индексе.

Для тестирования использовалось 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 «Аутоимунные 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) PHIAccess – 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.

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;

3) 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;

4) вследствие того, что эксперты категорий PHIAccess 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 онтологии.

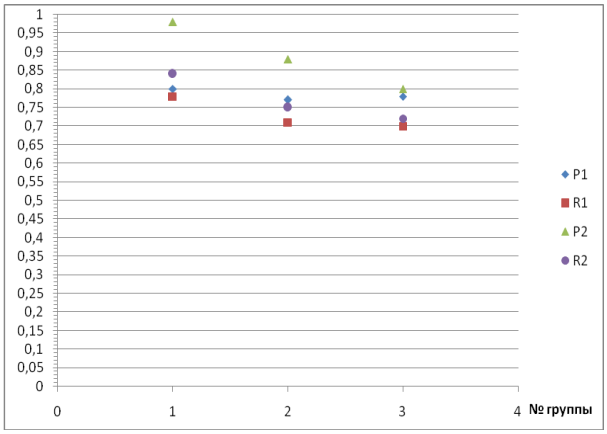

Рисунок 2 - Оценка точности и полноты результатов поиска категорией экспертов Viewer

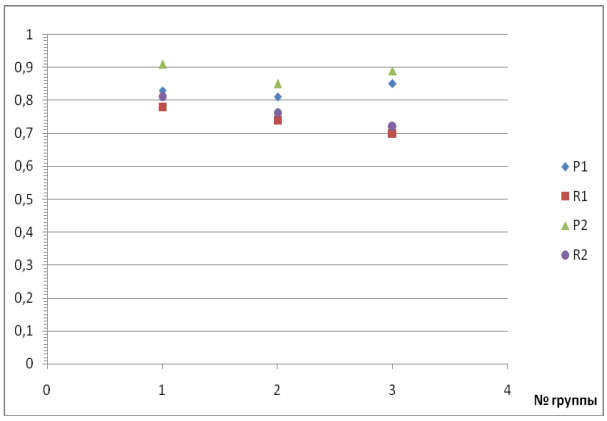

Рисунок 3 - Оценка точности и полноты результатов поиска категорией экспертов DataEntry

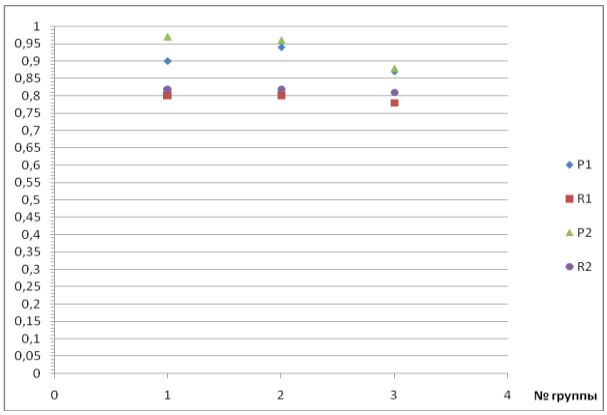

Рисунок 4 - Оценка точности и полноты результатов поиска категорией экспертов PHIAccess

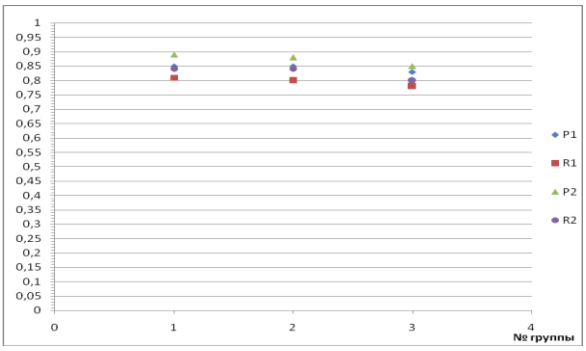

Рисунок 5 - Оценка точности и полноты результатов поиска категорией экспертов FullAccess

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.

Список литературы

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

Надійшла до редакції 20.10.2012

Г.А. ЄГОШИНА, О.С. ВОРОНОЙ
Донецький національний технічний університет

Оцінка якості інформаційного пошуку в слабо структурованих джерелах на основі метаданих і бази знань

Пропонується технологія підвищення ефективності пошуку в слабо структурованих базах даних з Web-інтерфейсом на основі метаданих, що представляють структуру баз даних і бази знань, яка описує семантику даних, що зберігаються. Проведено оцінку якості інформаційного пошуку в слабо структурованих джерелах на основі метаданих і бази знань для різних категорій користувачів з використанням критеріїв повноти і точності. Результати експериментів показали, що при реальних значеннях числа користувальницьких запитів до баз даних запропонована технологія забезпечує комфортний для користувачів час обробки запитів, скорочення навантаження на сервери системи і прийнятні для всіх категорій користувачів повноту і точність. З ростом числа запитів переваги запропонованої технології зростають.

Ключові слова: метадані, база знань, база даних, пошук, точність, повнота

A.A. YEGOSHINA, A.S.VORONOY
Donetsk National Technical University

Evaluation of the Quality of Information Retrieval in Semi Structured Sources Based on Meta-Data and Knowledge Base.

The paper provides a technique allowing a more effective search in semi structured Web-based databases based on metadata representing the structure of the database and knowledge base that describes the semantics of the stored data. We estimated the quality of information retrieval in semi structured sources (based on metadata) and knowledge base for different users using the completeness and accuracy criteria. Results of the experiments showed that with the actual number of user queries' the proposed technique provides a comfort respond time to the users query, reduces the load on the server system and has acceptable for all audiences completeness and accuracy. Growing number of requests increases advantages of the proposed technique.

Keywords: metadata, knowledge base, database, search, accuracy, completeness