УДК 004.3

А.А.Егошина, А.С.Вороной.
Государственный университет информатики и искусственного интеллекта
postmaster@suiai.edu.ua

Повышение эффективности извлечения информации из слабо структурированных источников 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. Предложена модель метаданных для рассматриваемых баз данных и структура хранилища метаданных. Хранение метаданных и базы знаний на отдельном сервере позволяет сократить время обработки запросов. Приведены результаты экспериментальных исследований в производственной среде работы с базами данных для различных профилей операций пользователей. Результаты экспериментов показали, что при реальных значениях числа запросов к базам данных предлагаемая технология обеспечивает комфортное для пользователей время обработки запросов и сокращение нагрузки на серверы системы. С ростом числа запросов преимущества предложенной технологии возрастают.
***Ключевые слова**: метаданные, база знаний, база данных, поиск, ресурс.*

Общая постановка проблемы

Непрерывный рост количества электронных документов и их доступности, а также увеличение объемов баз данных информационных систем (ИС) наряду со слабой структурированностью информационных фондов осложняет управление информацией и работу пользователей с ней.

За последние несколько лет активное развитие получило направление в информационных технологиях, занимающееся проблемами учета семантики в рамках информационных систем. ИС, использующие семантические технологии, отличаются от традиционных ИС следующими особенностями: использований знаний предметной области, в которой проводится обработка информации; знания предметной области выражаются явно, в виде моделей; модель выражает смысл терминов (понятий) предметной области через связи между ними.

Для решения проблемы совершенствования доступа к растущему объему информации и информационным услугам необходимо проводить обработку на семантическом уровне с использованием хорошо структурированного и постоянно применяемого стандарта метаданных. Такой стандарт позволит пользователям совершать поиск в большом количестве таблиц баз данных и уверенно определять местонахождение интересующей информации.

Информационный поиск на основе использования метаданных

Метаданные определяют ортогональный основному уровню описания информации (который формируется такими понятиями, как классы, типы данных и др.) уровень описания свойств [1]. Метаданные также могут описывать схему информационного источника, например реляционную, объектную или слабоструктурированную.

Контекстные метаданные описывают связь объекта с другими объектами системы, а контентные метаданные описывают содержимое объекта (т. е., имеющиеся в объекте знания).

Семантические метаданные (англ. semantic metadata) — дополнительный слой информации, позволяющий автоматизировать установление связей между различными частями контента. Семантические метаданные позволяют обогатить имеющийся контент принципиально новыми связями, которые невозможно установить при стандартной кластеризации.

Использование метаданных, в особенности контентных (семантических), позволяет эффективно решать такие задачи работы со знаниями, как поиск, категоризация и рекомендация знаний.

Для поиска необходимой информации в настоящее время в основном применяется поиск по ключевым словам. Пользователю приходится проходить по нескольким десяткам ссылок и изучать огромное количество информации, основная часть которой не является релевантной. Гораздо удобнее иметь поиск, выдающий однозначный результат, который содержит в

структурированном виде всю информацию об искомом объекте, отсортированную по убыванию релевантности.

Дополнение запросов шаблонами, логическими операторами, специальными словами для поиска в метаданных и другими средствами уточнения картину принципиально не меняет. Более ценными оказываются качественная поддержка морфологии различных языков и учет расстояния между словами. Однако при обработке значительных массивов данных, и они не позволяют радикально снизить «поисковый шум» – ведь простое наличие искомых слов далеко не всегда корректно отражает тему документа [2].

Поиск по метаданным в отличие от детального поиска по информационным ресурсам является более эффективным, поскольку метаданные предоставляют информацию, входящую в ресурс.

Технология поиска в слабо структурированных источниках информации с использованием базы знаний

При наличии больших объемов данных одно даже простое изменение может неблагоприятно повлиять на системы всего предприятия и сделать нерабочими модели данных, отчеты или привести к несовместимости данных. Визуальное представление зависимостей данных упрощает задачи проведения интеграции, анализа, управления данными, получая серьезное сокращение затрат времени при выполнении аудита.

Инициативы по интеграции данных обычно наталкиваются на отсутствие знаний о данных как таковых. Поиск и идентификация требуемых данных, так же как и определение их местоположения, могут занимать до 70% времени всего интеграционного проекта [3].

Использование базы знаний в гетерогенных информационных источниках предоставляет гибкий подход к интеграции информации.

База знаний является промежуточным слоем, расположенным между информационными источниками (базами данных с метаданными) и потребителями информации, как это показано на рисунке 1.

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.

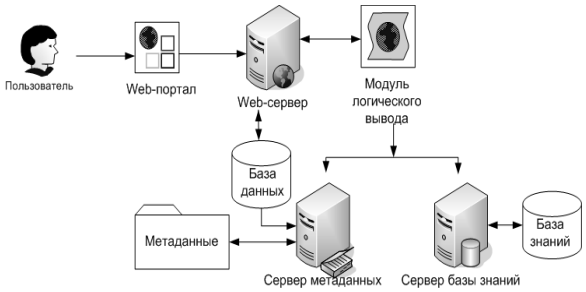

Рисунок 1 – Технология поиска в слабоструктурированных источниках информации с использованием базы знаний и метаданных

Метаданные описывают ресурсы с помощью маленьких простых пакетов информации, которые несложно найти, и к которым есть доступ для большего количества пользователей. Метаданные значительно улучшают степень структурирования поисковых задач в рамках объема данных в наличии.

В данной работе метаданные определяются как тройка следующего вида:

$$META = \{T, S, C\},$$

где T – совокупность сведений о таблицах базы данных:

$$T = \{id\_t, Sch, title\_t, \{Inf\_t\}\},$$

где *id_t* – код таблицы;
Sch – схема таблицы (тип группировки по некоторым общим признакам);
title_t – название таблицы;
{Inf_t} – множество вспомогательной информации, например, имя обработчика таблицы, коды таблиц – „родителей” и т.д.
S - совокупность сведений о схемах таблиц:

$$S = \{id\_s, schema - name\},$$

где *id_s* – код схемы;
schema-name – название схемы.
C - совокупность сведений о столбцах всех таблиц базы данных:

$$C = \{id\_c, type\_data, title\_table, title\_colum, \{Inf\_c\}\},$$

где *id_c* – код столбца;
type_data – тип данных, содержащихся в столбце;
title_table – название таблицы, содержащей данный столбец;
title_colum – название столбца;
{Inf_c} – дополнительная информация о столбце.

При значительном объеме ресурсов (таблиц базы данных) и их слабой структурированности будем хранить метаданные отдельно от ресурса в хранилище метаданных на отдельном сервере, как показано на рисунке 2.

Хранилища метаданных используются для хранения и управления метаданными. Такой механизм предоставляет больше гибкости, так как метаданные можно сделать открытыми в разных

установках и синтаксисах, которые со временем можно легко модифицировать. Глобальные изменения и поправки можно осуществлять после изначального создания.

Схема хранилища метаданных приведена на рисунке 2.

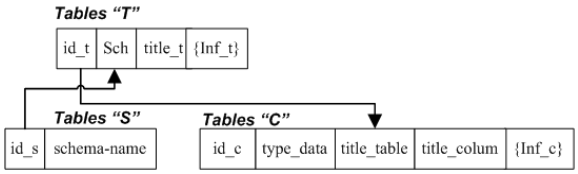

Рисунок 2 – Схема хранилища метаданных

Создание и управление предлагаемыми в данной работе метаданными позволяет решить следующие задачи.

1. Поддержка интеграции информационной системы. Схемы и интеграция данных зависят от метаданных, описывающих структуру и смысл отдельных источников данных и целевых систем.

2. Поддержка анализа и проектирования новых приложений. Метаданные повышают контролируемость и надежность процесса разработки приложений, обеспечивая информацию о смысле данных, их структуре и источниках. Кроме этого, метаданные, касающиеся решений по проектированию приложений, можно использовать повторно.

3. Повышение гибкости информационной системы и возможности повторного использования существующих программных модулей. Это возможно только для активного и полуактивного использования метаданных. Быстро изменяющиеся семантические аспекты явным образом хранятся в виде метаданных вне прикладных программ. Систему можно расширить и адаптировать без всяких трудностей. Данный подход также дает возможность повторного использования «фрагментов кода»;

4. Автоматизация административных процессов. Метаданные управляют запуском различных процессов (например, загрузки и обновления). Информация об их исполнении (журналы доступа, количество добавленных записей и т.п.) также содержится в репозитории, легко доступном администратору;

5. Эффективное извлечение информации и повышение качества данных, определяемого:

– согласованностью (является ли представление данных однородным, нет ли дубликатов, данных с пересекающимися или конфликтующими определениями);
– полнотой (все ли данные присутствуют);
– точностью (совпадением хранимых и фактических значений);
– своевременностью (актуально ли хранимое значение).

Экспериментальные исследования

С целью определения эффективности использования предложенного в работе подхода были проведены экспериментальные исследования на серверах компании Prometheus Research [4]. Основополагающим программным обеспечением компании является инновационная технология HTSQL, предоставляющая предприятиям и организациям, проводящим биомедицинские исследования, средства управления web-ориентированными базами данных в режиме реального времени.

Тестирование этого сценария должно помочь в оценке того, как изменение различных факторов и наличие метаданных влияет на производительность системы Prometheus Research Informatics Services And Solutions.

К основным характеристикам эксперимента относятся: количество пользователей, одновременно работающих с системой; типы выполняемых пользователями операций; количество опрашиваемых 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).

В таблице 1 представлены примеры элементов, обход которых был осуществлен.

Размер элементов составляет от 10 КБ до 100 МБ.

Таблица 1. Характеристики элементов

Название элемента	Число столбцов	Число строк
ssc_med_hx_birth_defects	96	808
ssc_med_hx_diet_medication_sleep_puberty	221	800
ssc_med_hx_language_disorders	139	806
ssc_med_hx_v2_labor_delivery_birth_feeding	130	802
ssc_med_hx_v2_medication_drugs_mother	225	810
ssc_med_hx_v2_pregnancy_history	247	790
ssc_med_hx_v2_psychiatric_disorders	182	806

В таблице 2 показано использование дискового пространства.

Таблица 2. Использование дискового пространства

Размер индекса на сервере индексирования	100 МБ
Размер базы данных поиска	2 ГБ

Следует отметить, что экспериментальные исследования проводились в производственной среде, где задержка сети и реакция хранилищ, обход которых осуществлялся, влияли на скорость обхода. Скорость обхода, измеряемая количеством таблиц в секунду, могла быть значительно выше в исключительно тестовой среде или в средах с более высокой пропускной способностью и лучшей реакцией хранилищ, обход которых осуществлялся.

В таблице 3 представлены результаты тестирования для различных профилей операций пользователей на основе оборудования и профиля использования, указанного выше. Таблица содержит следующие столбцы: № - номер эксперимента; з/с - запросов в секунду; %ЦПСИ - процент использования ресурсов центрального процессора сервера индексирования; %ЦПСБД - процент использования ресурсов центрального процессора сервера баз данных; о/с - средняя скорость операции (поиска, извлечения, записи) в секунду на диск сервера баз данных.

Таблица 3. Результаты экспериментов

№	з/с	%ЦПСИ	%ЦПСБД	о/с
1	1	0,038	0,016	0,78
2	10	0,56	0,17	1,019
3	50	0,87	0,48	1,098
4	100	5,2	6,7	5,2
5	500	11	10,27	7,8
6	1000	36,4	28,3	9,2
7	10000	54,2	46,3	11,6

В таблице 4 приведены результаты аналогичных экспериментов, но с использованием метаданных, предложенных в данной работе.

Результаты экспериментов показали, что при реальных значениях числа пользовательских запросов к базам данных (>1000 запросов в сек.) предлагаемая авторами технология обеспечивает комфортное для пользователей время обработки запросов.

Таблица 4. Результаты экспериментов с использованием метаданных

№	з/с	%ЦПСИ	%ЦПСБД	о/с
1	1	0,031	0,01	0,72
2	10	0,5	0,11	1,01
3	50	0,72	0,3	1,09
4	100	1	0,5	1,5
5	500	1,2	0,8	1,7
6	1000	2	1,5	3
7	10000	5	4,7	7

Кроме того, сокращается нагрузка на серверы системы, что позволяет снизить требования к аппаратному обеспечению. С ростом числа запросов преимущества предложенной технологии возрастают.

Заключение

В настоящей работе предложен новый подход к извлечению информации из слабо структурированных источников на основе метаданных и базы знаний.

Использование метаданных и базы знаний в информационной системе позволяет эффективно проводить мониторинг и оценивать качество данных, их содержание, применение в проектах и производительность. Компании и предприятия могут быстро и эффективно создавать собственные отчеты для своих интеграционных нужд.

Литература

1. http://www.w3.org/Metadata/
2. Даниэла Флореску Технологии баз данных для World-Wide Web: Обзор. Системы управления базами данных / Даниэла Флореску, Алон Леви, Альберто Мендельсон. – 1998. – 286 с.
3. Allison Powell James. The impact of database selection on distributed searching. In Proc. of the SIGIR'00, 2000.
4. http://www.prometheusresearch.com

Надійшла до редакції 01.03.2011