

позволит получить информацию о состоянии дорожного движения на основе данных о положении транспортных средств.

В заключение хочется отметить, что разработанный проект работает в полном объеме и в нем отсутствуют перечисленные выше недостатки. Также на данный момент ведется работа по внедрению уведомлений в мобильные приложения.

Список литературы

1. Web-site of system «Your Bus». [Electronic resource]. URL: <http://your-bus.ru/> (available on: 06.03.2015).
2. Loadimpact – Automated and on-demand performance testing for the DevOps-minded [Electronic resource]. URL: <http://loadimpact.com> (available on: 05.03.2015).

УДК 004

ИСПОЛЬЗОВАНИЕ ТЕХНОЛОГИЙ BUSINESS INTELLIGENCE ДЛЯ АНАЛИЗА ДАННЫХ В СФЕРЕ ГОСУДАРСТВЕННЫХ ЗАКУПОК

Н.С. Ибраева, Д.А. Сергеев

*Научный руководитель: А.В. Кудинов, к.т.н., доцент каф. ВТ ИК ТПУ
Национальный исследовательский Томский политехнический университет
E-mail: nsi5@tpu.ru; sergeevda@sibhtc.ru*

Abstract. *The article describes an approach to analysis of the information obtained from the state purchases portal containing a sufficiently large amount of data. Modern methods of data analysis (OLAP and Data Mining) are used in the project. They help to the user finding new, previously unknown knowledge. The article describes the process of extracting the necessary information and methods of its analysis.*

Key words: State purchases, parser, OLAP, Data mining, multidimensional cube, Microsoft SQL Server Analysis Services.

Ключевые слова: Государственные закупки, Синтаксический анализатор, OLAP, Data mining, многомерный куб данных, Microsoft SQL Server Analysis Services.

Введение

Предметом исследований, представленных в данной работе, является автоматизация процесса сбора и анализа данных портала государственных закупок (сокращенно госзакупки). Данные портала содержат развернутую информацию о сделках федерального, регионального и муниципального уровней (результатов проведения тендеров на продажу и покупку товаров и услуг). Целью анализа данных является выявление ранее неизвестных зависимостей между атрибутами закупок и их результатами. Инструментами для проведения анализа являются средства технологий OLAP и Data Mining, позволяющие представить информацию в виде, удобном для анализа и находить ранее скрытые знания, например, зависимости между изменениями данных, ранее не рассматриваемых при других видах анализа.

Постановка задачи

Задача, рассматриваемая в данной работе, может быть декомпозирована на следующие подзадачи:

1. Проанализировать архив закупок. Выявить основные объекты в документах закупок и их характеристики, общую и вариативную часть данных о всех видах госзакупок (общие и отличные атрибуты). Оценить сложность извлечения каждого атрибута выявленного объекта. Проанализировать исходный код страниц электронного ресурса.

2. Сформулировать конкретные задачи для извлекаемых наборов данных.
3. Построить информационную модель на основе выявленных объектов и их атрибутов, установить связи между объектами.
4. Извлечь наборы данных из ресурса, предоставляющего информацию о государственных закупках.
5. Решить различные аналитические задачи при помощи полученной реляционной базы данных на основе технологий OLAP и Data Mining.

Анализ предметной области

Исследуемой областью данного проекта является портал государственных закупок Российской Федерации. Данный портал позволяет оформлять заказы, заключать контракты, отслеживать отзывы пользователей услуг, вести мониторинг имеющихся заказов, организаций, контрактов, жалоб, недобросовестных поставщиков. Основой для исследований являются данные о заказах. Обработанная информация должна предоставлять возможность отслеживать следующие атрибутивные данные о заказе: номер извещения, наименование объекта закупки, способ определения поставщика (подрядчика, исполнителя), наименование организации заказчика, организация, осуществляющая закупку, местонахождение, дата и время начала и окончания подачи заявок, место подачи заявок, дата проведения аукциона в электронной форме, начальная (максимальная) цена контракта, источник финансирования, место доставки товара, выполнения работы или оказания услуги, сроки поставки товара или завершения работы либо график оказания услуг, наименования объектов закупки, члены комиссии, количество объектов, цена объектов, общая стоимость, преимущества участников, наличие обеспечения исполнения контракта, наличие жалоб на заказчика, наличие жалоб на поставщика. В результате анализа предметной области был сделан вывод, что первоначально необходимо извлечь данные из ресурса, разработав специализированный синтаксический анализатор (парсер), и из извлеченных данных сформировать БД. Проектирование реляционной модели БД необходимо для подготовки данных к анализу с помощью технологии OLAP в Microsoft SQL Server Analysis Services. Весь процесс анализа данных схематично представлен на рис. 1.

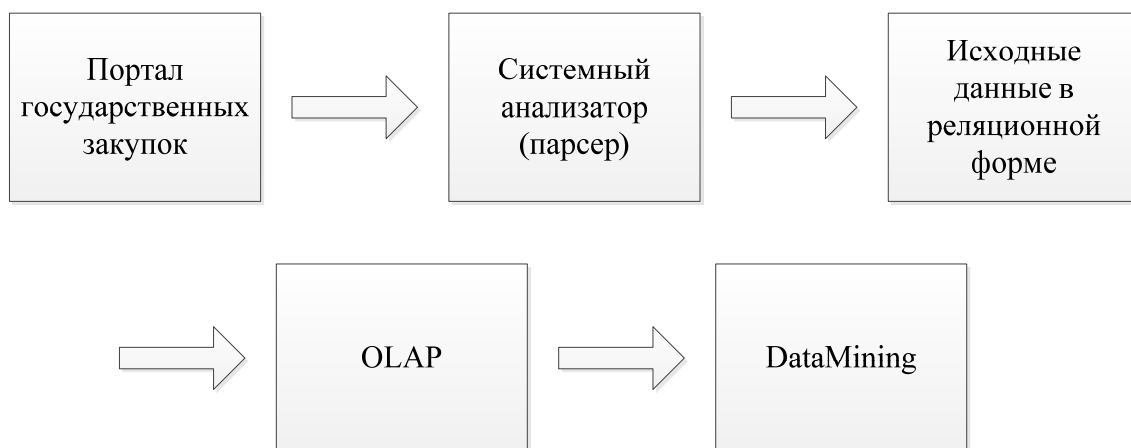


Рис. 1. Анализ данных портала госзакупок

Процесс извлечения и анализа данных

Для разработки парсера было принято решение использовать библиотеку phpQuery, входящую в состав PHP, для получения первичных данных. Библиотека представляет собой сочетание библиотеки jQuery из языка JavaScript и технологий языка PHP. С помощью данной библиотеки реализуется парсер, способный извлекать информацию из необходимого ресурса. Для хранения полученных данных используется система управления базами данных

MySQL. Входными данными для приложения, анализирующего сайт, являются URL-ссылки. В результате обработки парсером должны получиться выходные данные в реляционной БД. Парсинг страниц осуществляется в два этапа: на выходе первого этапа получаются ссылки, ведущие на страницы с необходимым контентом. На втором этапе извлекаются необходимые данные из страниц, ссылки на которые подготовлены на первом этапе. Далее анализ производится с помощью Microsoft SQL Server Analysis Services [1]. OLAP технология позволяет формировать данные в виде, удобном для анализа. Исходные данные представляют в виде многомерных кубов данных. Анализ данных на основе кубов производится с помощью Microsoft SQL Server Analysis Services 2012, который содержит развитые инструменты для создания OLAP-кубов и проведения анализа с их помощью. Сформированный OLAP-куб содержит в себе данные из таблиц, а также агрегатные значения для групп записей из таблиц. После подготовки исходных данных, они обрабатываются алгоритмами Data Mining. Одними из задач, которые можно решать на основе технологии Data Mining, являются ассоциация и кластеризация [2]. Применение ассоциации осуществляется в случае, если несколько событий связаны друг с другом. Располагая информацией о связности данных, можно выявить схожие схемы заключения сделок. Кластеризация позволяет объединять данные в группы по заранее неизвестным признакам. С помощью этого метода средства Data Mining самостоятельно выделяют различные однородные группы данных.

Задачи и технологии, позволяющие их решить

Технологии OLAP и Data Mining позволяют исследовать большие объёмы данных с целью их формирования в виде, удобном для анализа, поиска корреляций, получения хронологии изменения данных. К задачам анализа собранной информации относятся: выявление случаев мошенничества (например, случаи договорных тендеров), сравнительный анализ стоимости услуг и товаров (переплачивает ли заказчик), сравнение цен на продукты и услуги в зависимости от времён года, от региона, анализ тенденции принятий решения комиссии в зависимости от исполнителя, выявление манипуляций путем создания подставных кандидатов, учёт наличия жалоб и недобросовестных поставщиков. К задачам прогнозирования можно отнести анализ шансов на победу в тендере.

Список литературы

1. А.А. Бепреп. Microsoft SQL Server 2005 Analysis Services. OLAP и многомерный анализ данных. – СПб.: БХВ-Петербург, 2007. – 909 с.
2. Data Mining – интеллектуальный анализ данных. В.А. Дюк, Санкт-Петербургский институт информатики и автоматизации РАН [Электронный ресурс]. – режим доступа: URL: <http://www.inftech.webservis.ru/it/database/datamining/ar2.html>, свободный (16.03.15). – Загл. с экрана.