ИЗВЛЕЧЕНИЕ И АНАЛИЗ ДАННЫХ С ПОРТАЛА ГОСУДАРСТВЕННЫХ ЗАКУПОК

Ибраева Н.С., Кудинов А.В.
Национальный Исследовательский Томский Политехнический Университет
nsi5@tpu.ru

Введение

Объектом исследования являются открытые данные портала государственных закупок. Эти данные могут быть проанализированы с целью выявления различных видов мошенничества среди участников торгов. Для того, чтобы использовать современные инструменты и методы автоматизированного анализа открытых данных их сперва необходимо извлечь с портала, а также осуществить их подготовку к анализу. Статья посвящена решению данных задач.

Проанализировав предметную область, была выделена цель работы: создать синтаксический анализатор (далее парсер) и модель базы данных для извлечения данных с ресурса госзакупок для последующего их анализа.

Для достижения цели, обозначенной выше, необходимо решенить следующие задачи:

- провести исследование ресурса госзакупок;
- определить виды мошенничества, выявление которых будет осуществляться системой;
- выделить ключевые данные портала;
- изучить структуру XML документов;
- спроектировать таблицы с атрибутивными данными;
- нормализовать таблицы;
- выполнить связывание таблиц;
- разработать парсер;
- настроить парсер для анализа данных портала;
- произвести тестовое извлечение данных в БД.

Анализ предметной области

Исследуемой областью данного проекта является портал государственных закупок Российской Федерации. Данный портал позволяет оформлять заказы, участвовать в торгах, заключать контракты, отслеживать отзывы пользователей услуг, вести мониторинг имеющихся заказов, организаций, контрактов, жалоб, недобросовестных поставщиков. Основой для исследований являются данные о заказах [1]. Поскольку технические возможности ресурса не совершенны, нередко имеют место случаи мошенничества среди участников конкурсов. Соответственно, есть необходимость в методологиях и технологиях, позволяющим выявлять эти случаи, в том числе «договорные» тендеры, анализировать цены продуктов и услуг в зависимости от времён года, региона и других факторов, а также возможности манипулировать этими параметрами путём создания подставных кандидатов. Таким образом, для обеспечения выявления некоторых видов мошенничества существует необходимость создания методологии извлечения данных и их последующего анализа. Извлечённые данные должны содержать в себе информацию: о заказчиках, поставщиках и их заявках, итогах конкурсов, предметах торгов, результатах торгов, предлагаемые и итоговые цены на продукты или услуги, а также зафиксированое время каждой из операций. Анализ предметной области показал, что первоначально необходимо извлечь данные из ресурса, разработав при этом специализированный синтаксический анализатор. В дальнейшем следует извлечь данные из портала, сформировав реляционную базу данных, необходимую для подготовки данных к анализу.

Процесс извлечения и анализа данных

Для извлечения данных было решено использовать публичный FTP-сервер ресурса госзакупок, имеющего адрес ftp://free:free@zakupki.gov.ru [2].

Для получения первичных данных было принято решение реализовать парсер, используя библиотеки lxml, odbs, а также модули os, zipfile, ftplib, входящие в состав Python 3.4. Для хранения полученной информации было принято решение использовать систему управления базами данных Microsoft SQL Server. Входными данными для приложения, анализирующего сайт, являются FTP-ссылки. В результате обработки парсером получаются выходные данные в реляционной БД. Парсинг страниц осуществляется в несколько этапов: получение архивов данных, находящихся на FTP сайте государственных закупок (ftp.zakupki.gov.ru), извлечение из архивов файлов, которые имеют формат XML, анализ извлечённых файлов и сохранение необходимой информации на сервер.

На рис. 1 представлена общая модель синтаксического анализатора:

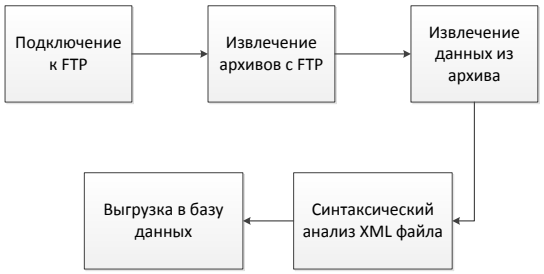

Рис. 4. Общая модель синтаксического анализатора

Для хранения данных необходимо создать реляционную БД. С этой целью была проанализирована структура хранения информация об извещениях, протоколах и контрактах. Данные о них представлены в региональной выгрузке в папках notifications, protocols, contracts.

Начальный этап закупок представлен в виде извещений. С их помощью заказчики информируются о начале торгов. Данные,

хранящиеся в извещениях, имеют ценность для дальнейшего анализа. Структура извещений хранит ключевую информацию о названии закупки, заказчике, дате опубликования, начальной цене, способе размещения заказа (электронный аукцион, открытый конкурс и другие), лотах и объектах закупки [3]. Одна закупка может иметь связь с несколькими поставщиками, поэтому в извещении может быть несколько лотов, т.е. у каждого лота может быть отдельный поставщик. В объектах закупки указывается информация о необходимых товарах или услугах. При их обозначении используется ОКПД (Общероссийский классификатор продукции по видам экономической деятельности).

Протокол представляет собой файлы с итоговыми решениями, принятыми при рассмотрении различных этапов проведения закупки. В файле протокола показаны заявки поставщиков, также в протоколах может быть несколько лотов. В таком случае заявки поставщиков относятся к определённому лоту.

В файлах протоколов имеется информация, необходимая при будущем анализе, такая как: информация о поставщике (раскрывается на последних этапах), дате подписания протокола, предлагаемых ценах и количестве товара. Протокол и заявка могут быть отклонены, что также отражается в базе данных.

Данные из файлов контрактов содержат следующую информацию: этап контракта, тип контракта, данные о победившем поставщике, заказчике, дата подписания, номер изменений. Указывается также информация о необходимых товарах или услугах и об утверждённых ценах на них.

Заключение

В процессе выполнения работы достигнуты следующие результаты: освоены методы предварительного извлечения и обработки данных, для извлечения данных; изучены методы парсинга веб-страниц для извлечения тестовых наборов данных и последующей работы с ними. В результате изучения методов извлечения данных был разработан синтаксический анализатор, предназначенный для выполнения сбора данных с FTP сервера портала государственных закупок. После создания синтаксического анализатора была спроектирована модель базы данных. Данная база является основой для построения кубов данных и дальнейшего их анализа.

По окончанию проектирования БД было произведено извлечение данных с FTP-сервера. В дальнейшем необходимо анализировать полученные данные с помощью технологии OLAP в Microsoft SQL Server Analysis Services.

Используемые источники

1. Ибраева Н.С.; Сергеев Д.А. Использование технологий Business Intelligence для анализа данных в сфере государственных закупок // Технологии Microsoft в теории и практике программирования: сборник трудов XII Всероссийской научно-практической конференции студентов, аспирантов и молодых ученых (Томск, 25–26 марта 2015 г.) / Томский политехнический университет. – Томск: Изд-во Томского политехнического университета, 2015. – 251 с.
2. Разъяснения по процедуре выгрузки сведений об опубликованных документах на FTP-сервер Общероссийского официального сайта [Электронный ресурс]. – режим доступа: URL: http://zakupki.gov.ru/wps/portal/base/topinfo/information, свободный (Дата обращения: 20.10.2015)
3. Интеграция ООС. Описание версии 5.0 [Электронный ресурс]. – режим доступа: URL: http://zakupki.gov.ru/epz/main/public/document/view.html?sectionId=6&pageNo=1&categories=FZ44&_categories=on&_categories=on&_categories=on&_categories=on, свободный (Дата обращения: 20.10.2015)