УДК 004

ИЗВЛЕЧЕНИЕ И АНАЛИЗ ДАННЫХ О СУДОПРОИЗВОДСТВЕ В г. ТОМСКЕ С ПОМОЩЬЮ ТЕХНОЛОГИЙ OLAP И DATA MINING

К.Б. Щукова, А.А. Хлопонин, Д.М. Паршина
Научный руководитель: А.В. Кудинов, к.т.н., доцент каф. ВТ ИК ТПУ
Национальный исследовательский Томский политехнический университет
E-mail: shchukova_kristina@yahoo.com, alex@diplux.com, sirena13@sibmail.com

Abstracts. *The article is intended to analyze various data obtained from websites of regional and district Tomsk courts via advanced analytic technologies such as OLAP and Data Mining. The process of comparing structure web pages and parsing HTML pages using PHP and C# is considered in details. Near-duplicates and shingling, as well as regular expressions and Levenshtein distance stand for analyzing and comparing texts, sentences and words. Due to these algorithms, the issue relating to extraction of necessary units can be sorted out effectively and quite accurately.*

Key words: the Law field, Data mining, OLAP, Microsoft SQL Server Analysis Service, HTML parser, regular expressions, shingling, text analysis, relational database.
Ключевые слова: судопроизводство, Data Mining, OLAP, Microsoft SQL Server Analysis Service, парсинг HTML-страниц, регулярные выражения, алгоритм шинглов, анализатор текста, реляционная база данных.

Введение. Значимым признаком информационного общества является наличие огромных объёмов разнородных данных в различных предметных областях, что дает возможность решить задачи поиска новых знаний, т. е. получения новых фактов, зависимостей и скрытых корреляций, а также решения ряда аналитических задач, таких как прогнозирование, проверка статистических гипотез, расчёт агрегатных показателей и т. д. В данной статье подробно рассматриваются алгоритмы и технологии для извлечения и анализа данных на примере судопроизводства в г. Томске.

Постановка задачи. Основная задача заключается в извлечении данных с сайтов Томских судов и их анализе с помощью технологий OLAP и Data Mining.

Поставленную задачу можно условно разделить на ряд следующих подзадач:

1. Анализ архивов судебных дел Томских областных, региональных, районных и арбитражных судов: структура архива, документов, судебных дел.
2. Построение информационной модели: выявление основных объектов и их характеристик в части судопроизводства, общей и вариативной части всех видов судебных дел.
3. Анализ HTML-страниц сайтов судов и оценка сложности извлечения из них данных, а также реализация HTML-парсера и анализатора текста для получения наборов данных судопроизводства.
4. Решение различных аналитических задач на полученном наборе данных с помощью технологий OLAP и Data Mining.

Анализ предметной области. Архив судебных актов Томских районных и областных судов состоит из административных, гражданских и уголовных дел. В зависимости от типа судебного решения можно выделить основные информационные объекты: постановления, решения, определения и приговоры. Каждый объект имеет общие атрибуты: номер дела, город, ФИО судьи, дата составления документа, название суда, нормативный акт (статья, часть, название), дата вступления в силу, содержание обвинения/правонарушения, доказательства, тип наказания, ФИО подсудимого. В результате анализа предметной области была построена реляционная модель данных судопроизводства.

Извлечение данных. В качестве инструмента для получения первичных данных была использована библиотека phpQuery, которая представляет собой портированную библиотеку

jQuery из языка JavaScript и технологий с ним связанными, в язык PHP. С помощью библиотеки phpQuery был реализован парсер для получения информации о судебных решениях, нормативных актов и т. д. Входными параметрами парсера является множество URL-ссылок, необходимых для загрузки целевых страниц. На выходе получается набор текстовых файлов, содержащих извлеченный текст, очищенный от тегов. Таким образом, процесс парсинга состоит из двух этапов: парсинг страниц со списком ссылок для перенаправления на основные страницы и парсинг самих страниц с нужной информацией. Следующий этап заключается в разработке анализатора извлеченного текста из HTML-страниц на языке C# для получения нужных значений конкретных атрибутов. Для сравнения текста с заданной категорией был использован алгоритм шинглов. Суть алгоритма заключается в разбиение текста на шинглы – выделенные из текста последовательности слов. Необходимо из сравниваемых текстов выделить подпоследовательности слов, идущих друг за другом по 10 штук. В результате получается набор шинглов в количестве равному количеству слов минус длина шингла плюс один. Кроме того, для сравнения текста с заданной категорией были использованы регулярные выражения и вычисление расстояния Левенштейна. Полученные значения атрибутов информационных объектов были занесены в реляционную базу данных.

Анализ данных. Технологии OLAP и Data Mining позволяют решать аналитические задачи, такие как расчёт статистических данных, задачи интеллектуального анализа, статистическая проверка гипотез, прогнозирование. К примерам задач получения агрегатных данных в судопроизводстве можно отнести следующие: подсчёт количества административных, гражданских и уголовных дел, который провёл судья определенного пола; подсчёт количества приговоров, решений, определений, постановлений, которые вступили в силу за указанный период; подсчёт количества районных, областных судов, где больше всего совершено уголовных, гражданских или административных нарушений; подсчёт процентного соотношения ведения уголовных, административных и гражданских дел в районных и областных судах. Задачи такого типа можно эффективно решать при помощи технологии OLAP.

К задачам интеллектуального анализа относятся: влияние пола судьи на ведение административных дел, влияние условий совершения преступления на степень его тяжести и т. д. С помощью статистических методов можно проверить выдвинутые гипотезы. К задачам прогнозирования можно отнести предсказание результата приговора по уголовному делу, учитывая следующие данные: степень тяжести совершенного уголовного преступления, доказательства совершения преступления, предыдущая судимость обвиняемого. Эти задачи можно решать при помощи технологии Data Mining. Для решения поставленных задач была использована аналитическая служба Microsoft SQL Server Analysis Service 2012. Этот сервер предназначен для создания OLAP-кубов на основе реляционных хранилищ данных. Построенный OLAP-куб содержит все данные из таблиц, а также агрегатные значения для групп записей из таблиц [1].

Текущие результаты и перспективы. Был реализован парсер для извлечения содержимого HTML-страниц, анализатор текста, основанный на расстоянии Левенштейна, алгоритме шинглов и регулярных выражениях, а также была построена информационная модель данных судопроизводства. В будущем планируется решение задач интеллектуального анализа с помощью технологий OLAP и Data Mining.

Список литературы

1. Барсегян А.А., Куприянов М.С. Методы и модели анализа данных: OLAP и Data Mining. – СПб.: БХВ-Петербург, 2004. – 336 с.