УДК 004

ПРОГРАММА СБОРА ДАННЫХ О СТРУКТУРЕ ВЕБ–САЙТОВ

***А.С. Потехин**, 2 курс*
Белорусский государственный университет информатики и радиоэлектроники

В настоящее время всеобщие глобальные тенденции приближаются к тому, что все операции и торговые сделки будут проходить с использованием веб–ресурсов. Для того, чтобы успешно вести бизнес очень важно получать актуальные данные о движения рынка (динамика цен и товаров) и локальные новости, которые порой всецело влияют на формирование спроса, своевременно. Но необходимые данные не всегда легко доступны пользователю и чаще всего они неструктурированны. Рассматривается приложение, которое будет обладать необходимым функционалом для сбора и структурирования данных с различных веб–ресурсов.

Целью исследования, для которого необходим сбор данных из Сети Интернет, является сентимент–анализ данных с различных новостных сайтов. Данные должны содержать полную информацию о новости, включая заголовок, текст, дату и автора новости. Для того, чтобы обеспечить сбор указанной информации, необходимо реализовать инструмент – web–scraper.

В широком понимании web scraping — это сбор данных с различных интернет–ресурсов. Общий принцип его работы можно объяснить следующим образом: автоматизированный код выполняет запросы на целевой сайт и получая ответ, парсит HTML–документ, ищет данные и преобразует их в заданный формат. Т.е. инструменты веб–скрапинга позволяют вручную или автоматически извлекать новые или обновленные данные и сохранять их для последующего использования.

Для того чтобы выполнять эту задачу, инструмент должен поддерживать работу со следующими данными:

1) HTML, JavaScript, так как большинство сайтов построены с использованием этих технологий;
2) Plain text, PDF и другие форматы представления текстовых данных;
3) URLs, с возможностью построения ни их основе графа веб–ресурсов.

Также инструмент должен обладать требованиями [1],[2],[3]:

1) Надежность – Веб содержит ресурсы, которые могут вводить скрапер в бесконечный цикл или недоступные сервисы, ожидать выполнения которых, он не должен. Скрапер должен быть устойчивым к таким ловушкам;
2) Вежливость – интернет–ресурсы имеют явные и неявные политики, регулирующие частоту, с которой скрапер может посетить их. Они описаны в файле robots.txt и эти политики должны соблюдаться;
3) Распределенность – скрапер должен иметь возможность выполняться в распределенном режиме на нескольких машинах;
4) Масштабируемость – скрапер должен поддерживать возможность увеличения производительности за счет добавления дополнительных вычислительных узлов, на которых он исполняется;
5) Производительность и эффективность – скрапер должен обеспечивать эффективное использование системных ресурсов, включая процессор, память и полосу пропускания сети;
6) Качество – скрапер должен уметь отделять спам–страницы от полезных и извлекать последние;

7) Актуальность – скрапер должен поддерживать обновление собранных данных;

8) Расширяемость – скрапер должен быть модульным, т.е. позволять добавлять новую функциональность, для анализа новых форматов данных, протоколов и т.д.

Помимо описанных общих требований для скраперов, можно обозначить основные требования, для конкретной задачи исследования:

1) Скрапер должен быть кросплатформенным, чтобы его можно было одинаково настраивать и конфигурировать на вы– числительных узлах с разными операционными системами;

2) Скрапер должен обеспечивать производительность обработки порядка 100 стр/сек, чтобы время сбора описанного выше объема данных составляло часы, а не дни. В том случае если окажется, что данных для сбора и анализа больше предполагаемого, скрапер должен предоставлять возможность легко увеличить его производительность путем выделения ему для работы большего числа потоков или добавления дополнительных вычислительных узлов;

3) Скрапер должен быть интегрирован с базой данных для хранения собранной информации и полнотекстовым индексом, позволяющим быстро извлекать данные для последующего анализа, отвечающие указанным условиям;

4) Требуется скрапер для сбора данных в ширину и вертикального поиска, так как в указанной задаче необходимо извлечь информацию о конкретной предметной области, а не узкое множество фактов;

В настоящее время существует множество готовых решений веб–скраперов, но готового решения для данной задачи исследования нет, поэтому, для реализации поставленной задачи был разработан собственный веб–скрапер.

Созданный скрапер является эффективным инструметом для поиска в Вебе, ядро написано на C++ с которым взаиможействует Ruby–оболочка., поддерживает граф связей узлов, различные парсеры, фильтры и нормализаторы URL. Он позволяет использовать различные хранилища данных, такие как Cassandra, Hbase и др. Скрапер также является масштабируемым (до 100 узлов в кластере и легко настраивается и расширяется, в полной мере является “вежливым”.

Список используемых источников

1. PAPAVASSILIOU V., PROKOPIDIS P., THURMAIR G. A modular open–source focused crawler for mining monolingual and bilingual corpora from the web // Proceedings of the 6th Workshop on Building and Using Comparable Corpora. — 2013.

2. AHUJA M.S., BAL J.S., VARNICA Web Crawler: Extracting the Web Data // International Journal of Computer Trends and Technology. — 2014.

3. YADAV M., GOYAL N. Comparison of Open Source Crawlers–A Review// International Journal of Scientific & Engineering Research. — 2015.