УДК 004.021

СПОСОБЫ ИЗВЛЕЧЕНИЯ ДАННЫХ, ИСПОЛЬЗУЮЩИЕ ТЕХНОЛОГИЮ ВЕБ-СКРАПИНГА. DIFFBOT КАК ГОТОВОЕ РЕШЕНИЕ ДЛЯ ИЗВЛЕЧЕНИЯ ДАННЫХ

Н.О. ШЕРШНЕВ
(***Представлено: канд. техн. наук, доц. А.Ф.ОСЬКИН***)

В данной статье рассматривается подход к получению данных из сети Интернет Web-scraping. Приводится пример готового решения (Diffbot).

Введение. Информация играет чрезвычайно важную роль в жизни человека. Ведь, как известно: «Кто владеет информацией, тот владеет миром». В процессе развития современного общества, все большие объемы несортированных данных хранятся в сети Интернет в открытом доступе. Поэтому при создании нового веб-ресурса перед разработчиком стоит выбор: использовать готовые данные из сети Интернет, или прикладывать усилия для создания собственного контента. Это обусловило создание и разработку технологий, которые бы позволяли собирать полезную информацию в интернете, анализировать ее и предоставлять в структурированном виде конечному пользователю.

Основной раздел. Такой подход к извлечению полезных данных называется веб-скрапингом, который подразумевает под собой создание такого программного обеспечения, которое позволило бы получать пользователю всю необходимую информацию с одного или нескольких интернет-ресурсов.

Существующие веб-скраперы имеют узконаправленную специализацию и зачастую создаются для конкретно веб-ресурса, что предполагает под собой большие человеческие усилия для автоматизации процессов получения и дальнейшего преобразования информации к структурированному виду.

Веб-скраперы имеют широкое применение в разных сферах жизни человека. При помощи веб-скраперов можно решить следующие задачи: мониторинг данных о погоде, сбор личных данных пользователей, поиск вакансий, работа с частными объявлениями по поиску жилья, отслеживание цен на товары в различных магазинах и т.д.

Среди готовых решений для скрапинга веб-сайтов стоит отметить следующие:

– веб-сервисы, которые работают через API (DiffBot, Embedly и др.).
– проекты с открытым кодом (Goose, Goutte, Morph, Scrapy и др.) [1].

Далее рассмотрим основные способы анализа и извлечения данных, использующие технологию веб-скрапинга:

1) **Copy-and-paste.** В некоторых случаях, лучшим решением для получения данных из сети Интернет является простое копирование необходимой информации пользователем. Так же этот способ полезен в том случае, когда веб-скрапер не может преодолеть защиту веб-сайта от машинной автоматизации;

2) **Text pattern matching.** Довольно простой, однако мощный подход к извлечению информации с веб-сервиса может быть основан на использовании регулярных выражений;

3) **HTTP programming.** Данный способ основывается на отправке http-запросов к удаленному веб-серверу, используя программирование сокетов;

4) **DOM parsing.** Программа, созданная на основе данного подхода, встраивается в полноценный веб-браузер, например, Internet Explorer или Mozilla Firefox. Система управления браузером анализирует веб-страницы в DOM-дереве, на основе которых программа получает необходимую информацию;

5) **Computer vision web-page analysis.** Существуют попытки создания такого программного обеспечения, которое использует машинное обучение и компьютерное зрение, идентифицирует и извлекает информацию с веб-страниц;

6) **HTML parsing.** Многие веб-сайты содержат коллекции страниц, которые были сгенерированы автоматически из таких источников, как базы данных. Данные обычно кодируются в похожие страницы с помощью общих скриптов и шаблонов. В процессе интеллектуального анализа данных, программа, которая определяет такие шаблоны, извлекает их содержимое и переводит в понятную форму называют оболочкой или wrapper; [2]

7) **Web-scraping.** Существует программное обеспечение, которое может быть использовано для настройки веб-скрапинг решений. Такое программное обеспечение может автоматически распознать структуру веб-страницы, что освобождает от написания веб-скрапинг кода; [2]

Среди готовых решений для веб-скрапинга стоит отметить DiffBot. DiffBot – это эффективное и революционное решение, которое предоставляет возможность автоматического извлечения структури-

рованных, нормализованных и точных данных из любой точки мира. DiffBot использует передовую технологию искусственного интеллекта и набор API для анализа веб-страниц и извлечения данных. Данное программное обеспечение снабжено всеобъемлющей диаграммой знаний, которая содержит точную и подробную информацию о различных объектах, найденных в Интернете, таких как люди, места, компании, организации, предприятия, продукты, статьи и обсуждения. Пользователи могут легко и точно запросить этот граф знаний, чтобы отобразить любые необходимые им данные. Кроме того, Diffbot определяет, как информация и объекты связаны друг с другом, облегчая пользователям понимание данных, предоставляемых программным обеспечением, и их использование для достижения своих конкретных целей. [4]

Рисунок. – Diffbot лучшее готовое решение для веб-скрапинга

Важно отметить следующие плюсы использования Diffbot:

– автоматическое определение страниц – при помощи API Analyze осуществляется автоматический поиск и извлечение всех продуктов, статей, обсуждений, видео или изображений при сканировании любого сайта;

– подробные данные о продуктах – Product API автоматически возвращает полную информацию о продукте, включая все данные о ценах, идентификаторы продуктов, марки и полные таблицы спецификаци;

– чистый текст и HTML – статьи, обсуждения, описания продуктов и подписи к изображениям возвращаются в виде чистого текста и HTML;

– структурированный поиск – поиск структурированного контента на лету из любого сканирования.

Заключение.В данной статье рассмотрен Web-scraping как инструмент извлечения полезных данных из сети Интернет, основные способы анализа и обработки данных, назначение и применение в повседневной жизни человека. Приведен пример готового решения для веб-скрапинга Diffbot.

ЛИТЕРАТУРА

1. Википедия [Электронный ресурс] Web-scraping Режим доступа: https://en.wikipedia.org/wiki/Web_scraping. Дата доступа: 25.09.19.
2. Habrahabr [Электронный ресурс] Web-scraping с помощью Pyton. Режим доступа: https://habrahabr.ru/post/280238/. Дата доступа: 25.09.19.
3. Habrahabr [Электронный ресурс] Web-scraping при помощи Node.js. Режим доступа: https://habrahabr.ru/post/301426/. Дата доступа: 25.09.19.
4. Finances online [Электронный ресурс] Diffbot Review. Режим доступа: https://reviews.financesonline.com/p/diffbot/#overview-benefits. Дата доступа: 25.09.19.