УДК 004.021

WEB-SCRAPING КАК ИНСТРУМЕНТ ИЗВЛЕЧЕНИЯ ПОЛЕЗНЫХ ДАННЫХ ИЗ СЕТИ ИНТЕРНЕТ

Н.О. ШЕРШНЁВ
(Представлено: Ю.Н. КРАВЧЕНКО)

Рассматриваются преимущества и недостатки получения данных из сети Интернет способом Web-scraping. Показаны основные способы анализа и обработки данных, их назначение и применение в повседневной жизни человека.

Информация играет чрезвычайно важную роль в жизни человека. Известно, что «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.

Веб-скрапингом называется такой подход к извлечению полезных данных веб-скрапингом, который подразумевает под собой создание специального программного обеспечения, позволяюшего получать пользователю всю необходимую информацию с одного или нескольких интернет-ресурсов.

Существующие веб-скраперы имеют узконаправленную специализацию и зачастую создаются для конкретного веб-ресурса, что предполагает 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. Также 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, программу, 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].

Заключение

В ходе проведенного исследования рассмотрены основные способы анализа и извлечения данных, использующие технологию веб-скрапинга:

- Web-scraping как инструмент извлечения полезных данных из сети Интернет;
- основные способы анализа и обработки данных;
- назначение и применение их в повседневной жизни человека.

ЛИТЕРАТУРА

1. Википедия [Электронный ресурс] Web-scraping. – Режим доступа: https://en.wikipedia.org/wiki/Web_ scraping. – Дата доступа: 26.09.2017.
2. Habrahabr [Электронный ресурс] Web-scraping с помощью Pyton. – Режим доступа: https://habrahabr.ru/ post/280238/. – Дата доступа: 26.09.2017.
3. Habrahabr [Электронный ресурс] Web-scraping при помощи Node.js. – Режим доступа: https: //habrahabr. ru/post/301426/. – Дата доступа: 26.09.2017.