Методы и инструменты автоматического сбора информации с Веб-ресурсов

Н.В. Сахаровский

При поиске в Интернете определённых товаров пользователю приходится сравнивать предложения различных электронных магазинов и сайтов объявлений. При этом возможности фильтрации и поиска информации на исходных сайтах часто оказываются слишком ограниченными, поэтому возникает желание собрать информацию из различных источников в единую БД, для которой могут быть реализованы более мощные механизмы поиска. Задача автоматизации сбора информации с Веб-ресурсов называется Веб-скрейпингом.

Для отображения в браузере на сервере по содержимому БД HTML страницы, также для передачи информации в браузер может использоваться обычный текст, XML или JSON. Для массового извлечения информации необходимо сделать запросы к серверу для получения списка страниц, которые он может выдать, далее запросить у него каждую из страниц. При этом требуется извлекать из получаемых с страниц интересующую информацию. Программа Веб-скрейпинга извлекает из Веб-страниц определённую часть данных, например с различных сайтов может собираться информация об именах, электронных адресах, номерах телефонов и т.д для формирования базы контактов.

Веб-скрейпинг в наше время является серьёзным инструментом автоматизированного поиска информации в глобальной сети Интернет. Все информационные поисковики тесно связаны с Веб-скрейпингом. В основном Веб-скрейпинг используется для поиска информации, копирования различных данных из Интернета, мониторинга объявлений на различных сайтах по продаже товаров и услуг. С помощью Веб-скрейпинга можно анализировать и сравнивать цены на различные товары для корректировки цен, наблюдения за конкурентами и оценки их возможностей, поэтому такие инструменты очень востребованы практически во всех отраслях. Веб-скрейпинг и Интернет появились примерно в одно время, сделав всемирную паутину доступной для поиска информации.

В декабре 2019 года 17 летний школьник из США, Ави Шиффманн создал очень популярный сайт ncov2019.live про коронавирус. Одним из основных инструментов наполнения сайта является Веб-скрейпинг, с использованием которого собирается информация со всего мира о количестве заразившихся вирусом, погибших и вылечившихся и демонстрирует эти показатели в реальном времени. В какой-то момент трафик сайта достиг 350 миллионов посещений, а в день сайт посещают примерно 30 миллионов человек.

Область Веб-скрейпинга активно развивается во взаимодействии компьютера и человека, также используя искусственный интеллект для структурирования информации в определённом формате. Программы Веб-скрейпинга не рассчитаны на обычных пользователей, поскольку для своей работы они требуют написания кода под конкретную задачу.

Во всемирной паутине можно найти много инструментов для Веб-скрейпинга. Одним из популярных инструментов скрейпинга является «Scrapy» - программная платформа с бесплатным, открытым кодом. Среди платных платформ можно выделить Import.IO. Для написания алгоритмов разбора HTML и XML кода на Python удобно использовать библиотеку BS4.

В ответ на создание инструментов Веб-скрейпинга разрабатываются программы, которые обнаруживают и блокируют таких ботов. Это делается для того, чтобы информация не была использована конкурентами и для снижения нагрузки на сайт. Есть несколько признаков, по которым можно обнаружить Веб-скрейпинг. Например, обнаруживает себя нестандартное поведение пользователя в виде большого количества переходов на страницу сайта, когда пользователь циклично выполняет одни и те де действия. В таком случае блокируется IP-адрес, с которого бот многократно обращался к страницам сайта. Поэтому работа Веб-скрейпинга должна быть максимально похожа на действия обычного пользователя, чтобы избежать блокировки. Периодически следует ротировать IP-адреса, изменять скорость обращения к серверу, добавлять случайные действия на сайте, чтобы не вызвать подозрений.