ВЕБ-СКРЕЙПИНГ, МЕТОДЫ ПРОТИВОДЕЙСТВИЯ И ИХ ЭФФЕКТИВНОСТЬ

Т.Е. Белогривая, Р.А. Фортель

Учреждение образования «Гродненский государственный университет имени Янки Купалы», Гродно, Беларусь

Веб-скрейпинг является технологией получения данных путем извлечения их со страниц веб-ресурсов. Веб-скрейпинг может быть сделан пользователем компьютера вручную, но обычно данный термин относится к автоматизированным процессам, которые реализованы с помощью кода, выполняющего GET-запросы на сайт.

Основной задачей веб-скрейпинга является сбор данных из интернет-источников. С помощью данной технологии можно не только искать и копировать информацию, но и мониторить обновление информации на веб-сайтах.

Средств противодействия веб-скрейпингу довольно много, но нет ни одного на 100 % эффективного. Пока пользователи сайта могут получить к нему доступ, веб-скрейпер также может это делать, так как он имитирует действия реального пользователя. К тому же, методы противодействия веб-скрейпингу могут мешать реальным пользователям.

Рассмотрим основные контрмеры противодействия автоматизированному сбору данных.

1. Блокировка IP-адресов. Это один из способов противодействия веб-скрейпингу, когда данные крадут постоянно и, как правило, в большом объеме. Здесь речь идет уже не только о текстах, но и других сведениях, представляющих стратегический интерес для конкурентов. Разрешить или запретить доступ к сайту некоторым IP-адресам можно через менеджера IP-адресов на используемом хостинге.

Недостаток: бот может использовать прокси (поддельные IP-адреса).

2. Использование капчи. Этот метод заключается в добавлении капчи на сайт. Триггером вывода капчи может является переход на определенную страница сайта, определенное действие пользователя и многое другое. Однако при использовании данного метода следует понимать, что капча будет мешать реальным пользователям. Самый популярный сервис на данный момент это reCAPTCHA от компании Google.

Недостаток: капчи можно обойти с помощью сторонних сервисов, где реальные пользователи их решают.

3. Электронная почта и номер мобильного телефона. При выполнении определенных действий на сайте проводится проверка по мобильному телефону или электронной почте.

Недостаток: бот может использовать одноразовые электронные ящики или временные мобильные номера виртуальных сотовых операторов.

4. Изменение структуры сайта. Интернет-ресурс может регулярно обновлять структуру сайта. Из-за этого злоумышленнику придется переписывать код сбора данных.

Недостаток: современный веб-скрейпинг не завязывается на структуре сайта, а использует более точные идентификаторы.

Таким образом, хорошо написанный код имитирует активность реального пользователя и вычислить бота очень трудно. Также следует понимать, что, защищаясь от веб-скрейпинга, сайт может оказаться заблокированным для краулеров Google и Яндекса, которые являются такими же ботами. Последствия этого очевидны: сайт частично или полностью потеряет лидирующие позиции в поисковиках.