УДК 004.056

ВЛИЯНИЕ ЗАЩИТЫ ОТ ПАРСИНГА ДАННЫХ НА ПОЛОЖЕНИЕ В ПОИСКОВОЙ ВЫДАЧЕ

Бабарицкий П.А. (Университет ИТМО)
Научный руководитель – к.п.н., доцент Государев И.Б.
(Университет ИТМО)

Настоящая статья охватывает проблему влияния наличия защиты от парсинга данных на выдачу веб-ресурсов в поисковой выдаче. Рассматриваются различные автоматизированные способы сбора данных на основе различных языков программирования (python и php). Кроме того, рассмотрен набор защит, с которыми данные автоматизированные приложения могут столкнуться. Включает результаты для выборки интернет-магазинов такие как посещаемость, релевантность и наличие защиты.

Постановка научной проблемы, описание существующего положения, анализ отечественного и зарубежного опыта в решении данной проблемы и т.д.
При рассмотрении научных публикаций, раскрывающих понятия парсинга данных и такого явления как web-scrapping становится понятно насколько актуальна и важна проблема анализа методов извлечения и обработки данных веб ресурсов для области веб-безопасности и их коммерческого применения в интернете. Актуальность определена тем, что ранее не проводилось подобных исследований, нацеленных на поиск связи между использованием методов защиты от парсинга данных и их влиянием на веб-ресурсы с применением теоремы Байса. В данной работе объектом исследования выступают торговые площадки. Предметом исследования стали средства защиты информации от веб-извлечения. Выборкой стали релевантные сайты, то есть те, что смогли попасть в выдачу поисковой системы Google или Yandex. Для анализа была применена теорема Байеса, которая позволяет дополнять уже произведенный анализ, и учитывать получение ранее результаты.

В рамках аналитического обзора рассматриваются существующие методы и средства по противодействию несанкционированному извлечению и обработке данных с веб-ресурса. По результатам аналитического обзора был выявлен ряд моделей, средств и методов противостояния воздействию нежелательных веб-роботов, который был разделен на две глобальные группы: базовые и высокоэффективные и полечена условная схема иерархии методов противодействия нежелательным веб-роботам, нацеленным на несанкционированную обработку и извлечение данных веб-ресурсов. Каждая из приведенных методологий кратко рассматривалась для раскрытия теоретических аспектов противодействия веб-роботам, нацеленным на несанкционированную обработку данных веб-ресурсов. Данный аналитический разбор позволил провести тестирование гипотез на предмет соответствия действительности методом Байеса. Целью данного эмпирического исследования был поиск наличия зависимостей между защитой и популярностью торговых интернет-ресурсов, релевантных для поисковых систем. Для достижения поставленной цели, были решены следующие задачи: аналитический обзор средств защиты; разработка специализированного ПО для автоматического сбора целевой информации; расчет полученных данных на базе теоремы Байеса. Сбор информации осуществлялся с использованием Python 3. На данном языке было разработано приложение, которое осуществляло сбор данных о торговых сайтах. Основной выборкой выступили те площадки что вернула поисковая система. В данном исследовании использовались две из существующих поисковых систем Google и Yandex. Для автоматического сбора и эмуляции работы пользователя использовалась библиотека Selenium и через парсинг были проверены сайты на наличие на них соответствующих защит. На основе выборки определенной выдачей поисковой системы, сформировался ряд гипотез:

1) P(H|E) - гипотеза о том, что популярные торговые веб-сайты используют защиту от веб-ботов;

2) P(H|E) - гипотеза о том, что не популярные торговые веб-сайты используют защиту от веб-ботов;
3) P(H|E) - гипотеза о том, что релевантные торговые веб-сайты используют защиту от веб ботов;
4) P(H|E) - гипотеза о том, что менее релевантные торговые веб-сайты используют защиту от посещения их страниц веб ботами.

Где H – случай, когда сайт обладает защитой, а E –ограничивающие условие (популярные, не популярные, релевантные, не релевантные сайты). P(E) – вероятность наступления такого условия. Релевантными признаны первые сто веб-страниц, а популярными – те сайты, которые смогли превысить планку в 9500000 человек в месяц.

На основе вышеприведенных данных был сформирован прогноз о вероятности наличия защиты в зависимости от популярности сайта или положения его в поисковой системе и получен ряд гипотез, из которых верной можно считать одну из предложенных: о том, что популярные торговые веб-сайты используют защиту от веб-ботов. Остальные предположения, которые имели низкие показатели соответствия действительности, не подтвердились и были отклонены. Также полученные результаты позволяют сделать предположение о внутренней работе поисков систем и о том, что могло повлиять на итоговые результаты исследования. В итоге, для получения более обширного ответа и выведения более точных результатов, следует использовать большее количество входных данных и предполагается рассмотрение влияния определенных видов защит по отдельности. Кроме того, разработанный способ сбора данных позволяющий получить результаты анализа защиты смежных по тематике сайтов, может быть усовершенствован и внедрен в сервисы SEO-анализа сайтов в качестве дополнительных данных.