ОБНАРУЖЕНИЕ РОБОТИЗИРОВАННОГО ЗАПОЛНЕНИЯ ФОРМ НА ВЕБ-РЕСУРСЕ

А.А. Менщиков, А.В. Комарова, Ю.А. Гатчин
Университет ИТМО
Санкт-Петербург

Введение.

На сегодняшний день все большее количество информации переносится в сеть Интернет. Это и персональные данные людей, и личная переписка, и данные банковских карт, и многое другое. Такой порядок вещей требует обеспечения целостности, конфиденциальности и доступности информации, расположенной в Сети. Владельцам веб-сайтов необходимо уделять особенное внимание защите от краулеров, которые в значительной степени могут мешать легитимным пользователям комфортно использовать данный ресурс, а также являться источником финансовых потерь владельцев ресурса за счет понижения поискового рейтинга сайта, появления сайтов-клонов, манипуляций с функциональностью, махинаций с рекламой и данными, а также многих других угроз. Зачастую веб-ресурсы имеют закрытые разделы, доступные только для авторизованных пользователей. Это приводит к массовой регистрации поддельных аккаунтов с целью автоматизированного сбора информации из таких разделов. Одним из способов обнаружить поддельные аккаунты является анализ активности пользователя на этапе авторизации и заполнения регистрационных данных. В связи с актуальностью проблемы, в данной работе затронут вопрос обнаружения роботизированного заполнения форм на веб-ресурсе.

Описание метода

Как уже отмечалось выше, веб-роботы — это специальные средства, предназначенные для сбора информации с различных Интернет-ресурсов [1, 2]. Помимо легитимных роботов, которые могут анализировать контент веб-ресурса [3] или индексировать сайты для улучшения работы в поисковых системах, существуют также и роботы-злоумышленники [4, 5]. Такие роботы могут рассылать спам и рекламу, совершать покупки на сайтах, эксплуатировать различные уязвимости системы. Все эти злонамеренные действия приводят к уменьшению пропускной способности ресурса, увеличению траффика за единицу времени, к проблемам доступа у легитимных пользователей, что в свою очередь приводит к финансовым потерям владельца ресурса. По данным аналитических компаний, объемы парсинга с каждым годом увеличиваются [6]. По этой причине поиск методов различия обычных пользователей от вредоносных становится актуальной задачей.

Существующие системы противодействия данной угрозе являются закрытыми проектами, механизм действий которых ограничен коммерческой тайной [7]. Академические исследования носят более теоретический характер и имеют важный недостаток — сосредоточенность на одном методе обнаружения, либо жесткую привязку к конкретному защищаемому ресурсу [8]. Актуальной задачей становится поиск универсальных методов обнаружения веб-роботов. Мы предлагаем в качестве одной из ступеней такого обнаружения анализировать заполнение форм на веб-ресурсе как точку, где веб-робот вынужден выполнять активные действия на сайте. В отличие от существующих статических и динамических методов при заполнении форм робот вынужден выполнять большое количество активных действий, что увеличивает сложность эмуляции им поведения легитимных пользователей [9].

Целью данного исследования является разработка метода обнаружения веб-роботов при помощи анализа заполнения форм на веб-ресурсе.

Использовались техники пассивного и активного анализа поведения пользователей на веб-ресурсе. К активному анализу относится регистрация действий с использованием JavaScript. К пассивному — изучение структурных, временных и поведенческих характеристик запросов. Данные характеристики изучались при помощи шаблонного и статистического анализа. Были составлены списки правил для обнаружения некорректного заполнения форм, например, с использованием разного рода ловушек и невидимых для легитимных пользователей полей ввода. На этапе статистического анализа изучались аномалии заполнения форм с использованием методов машинного обучения. Комбинация данных подходов позволила авторам минимизировать ошибки и повысить точность обнаружения автоматизированных регистраций на веб-ресурсе.

Результаты

Были изучены несколько веб-ресурсов с месячной аудиторией более тысячи пользователей. На исследуемых сайтах были установлены специальные программные средства анализа заполнения форм [10]. Из базы активных и доверенных пользователей путём ручной и полуавтоматической верификации были выбраны характеристики, описывающие легитимного посетителя. Затем, эти характеристики сравнивались с данными недоверенных пользователей. Для обучения классификатора использовались также пользователи,

отнесённые к роботам при помощи ручного изучения их параметров, с использованием шаблонов, правил и ловушек на веб-ресурсе, позволяющих однозначно классифицировать посещение как роботизированное. Данные были разделены на несколько выборок для проведения перекрестной проверки. Полученные точность и полнота обнаружения позволяют сделать вывод о применимости метода анализа заполнения форм в рамках системы обнаружения автоматизированного сбора информации с веб-ресурсов.

Выводы

В связи со все большим ростом количества и активности веб-роботов, тема исследования на сегодняшний день является актуальной. Система обнаружения показала свою эффективность по скорости, точности и полноте, что позволяет использовать её для решения задачи обнаружения и противодействия автоматизированному сбору информации с веб-ресурса. Использование анализа заполнения форм позволяет повысить эффективность работы системы обнаружения веб-роботов, поскольку повышает сложность эмуляции активных действий на ресурсе.

В результате данного исследования было показано преимущество предлагаемого метода над существующими подходами к обнаружению на основе анализа запросов.

Работа в данном направлении может послужить хорошим базисом для будущих исследований, и будет способствовать дальнейшему развитию данной области.

ЛИТЕРАТУРА

1. Эркенов М.С. Основные принципы построения и функционирования информационно-поисковых систем в сети Интернет // Вестник Ростовского государственного университета путей сообщения. 2007. № 2 (26). С. 69-77.
2. Junsup L., Sungdeok C., Dongkun L., Hyungkyu L. Classification of web robots: An empirical study based on over one billion requests // Computers & Security. 2009. V. 28. No 8. Pp. 795-802.
3. Nesi Paolo, Gianni Pantaleo, Marco Tenti. Geographical localization of web domains and organization addresses recognition by employing natural language processing, Pattern Matching and clustering // Engineering Applications of Artificial Intelligence. 2016. 51. Pp. 202-211. URL: http://dl.acm.org/citation.cfm?id=2910172 (дата обращения 16.04.2018).
4. Menshchikov A., Komarova A., Gatchin Y.A., Korobeynikov A.G., Tishukova N. A Study of Different Web-Crawler Behaviour // Proceedings of the 20th Conference of Open Innovations Association FRUCT – 2017. Pp. 268-274.
5. G. Jacob, E. Kirda, C. Kruegel, G. Vigna. PUB CRAWL: Protecting Users and Businesses from CRAWLers // Proceeding Security'12. Proceedings of the 21st USENIX conference on Security symposium. 2012. Pp. 25–36.
6. Отчет компании distil networks. URL: https://resources.distilnetworks.com/travel/2018-bad-bot-report, свободный (дата обращения: 23.04.2018).
7. Сайт компании imperva. URL: https://www.imperva.com/products/threatradar/bot-protection/, свободный (дата обращения 23.04.2018).
8. Web Robot Detection in Academic Publishing. URL: https://arxiv.org/pdf/1711.05098.pdf (дата обращения 23.04.2018).
9. Кипаева Е.В., Кириченко М.И., Орлова Ю.А., Заболеева-Зотова А.В. Распознавание ботов в социальных сетях // Открытые семантические технологии проектирования интеллектуальных систем. 2014. № 4. С. 431-434.
10. Менщиков А.А., Комарова А.В. Система обнаружения автоматизированного сбора информации с веб-ресурсов // Свидетельство о государственной регистрации программы для ЭВМ № 2017661508, 21 августа 2017 г.