УДК 004.056.3 Вестник УрФО № 2(32) / 2019, с. 25–29

Менщиков А. А., Гатчин Ю. А., Коробейников А. Г. DOI: 10.14529/secur190204

КОМБИНИРОВАННЫЙ МЕТОД ОБНАРУЖЕНИЯ И ПРОТИВОДЕЙСТВИЯ АВТОМАТИЗИРОВАННОМУ СБОРУ ИНФОРМАЦИИ С ВЕБ-РЕСУРСОВ

Статья посвящена вопросу разработки комбинированного метода обнаружения и противодействия автоматизированному сбору информации с веб-ресурсов. Проблема противодействия веб-роботам является важной, согласно отчётам аналитических компаний. Веб-роботы угрожают приватности данных, авторскому праву и несут угрозы работоспособности веб-ресурсов. В данной статье предлагается метод противодействия веб-роботам, использующий комбинированный подход к обнаружению на основе семантических и графовых поведенческих методов. Приводится исследование характеристик обнаружения и алгоритм выбора стратегии реагирования. Результаты применения данного подхода показывают точность обнаружения и противодействия выше 95%.

Ключевые слова: *веб-роботы; обнаружение веб-роботов; противодействие веб-роботам; безопасность веб-ресурсов.*

Menshchikov A. A., Gatchin U. A., Korobeynikov A. G.

COMBINED DETECTION AND COUNTERACTION METHOD OF AUTOMATED INFORMATION DATA GATHERING FROM WEB RESOURCES

The article is devoted to the development of a combined web-robot detection and counteraction method. The problem of web-robot prevention is important, according to reports from analytical companies. Web robots threaten data privacy, copyright and affect the performance of web resource. This article proposes a new method of web-robot counteraction using a combination of detection approaches based on semantic and graph behavioral methods. A study of the characteristics of the detection method and the algorithm for selecting a response strategy is provided. The results of applying this approach show the accuracy of detection and counteraction above 95%.

Keywords: *web-robot; web-robot detection; web-robot counteraction; website security.*

Введение. Веб-роботы – это специализированные средства сбора информации с веб-ресурсов [1]. Значительную долю пользователей веб-ресурсов составляют автоматизированные средства, ведущие несанкционированную деятельность от кражи информации с целью размещения на другом ресурсе до выполнения действий с целью получения выгоды и преимущества над рядовыми пользователями ресурса. В 2018 году OWASP выпустил документ об автоматизированных угрозах, где привёл классификацию 21 различного вектора атак на веб-ресурс со стороны автоматизированных средств [2].

Методы обнаружения и противодействия веб-роботам заключаются в поиске характерных признаков роботизированного поведения и сравнения профилей поведения в рамках пользовательских сессий. Характеристиками могут выступать различные параметры, получаемые на уровне клиента, веб-сервера и веб-приложения [3]. На уровне клиента данные собираются посредством JavaScript кода и иного активного содержимого. На уровне веб-сервера собирается статистика по структуре и содержанию HTTP и WebSocket трафика, а также информация об источнике запросов (база GeoIP). На уровне веб-приложения анализируется логика и структура поведения пользователя [4].

Помимо данных характеристик предлагается использовать информацию о структуре и содержании защищаемого веб-ресурса, что позволит связать поведение пользователя с той средой, с которой он взаимодействует.

Методы противодействия. Для применения сценариев противодействия веб-роботам необходимо осуществить процедуру обнаружения. Запросы от пользователей группируются в сессии, что позволяет строить поведенческий профиль на основе связанных последовательных запросов от каждого пользователя.

На первом этапе происходит сбор данных от веб-сервера. Рассчитываются семантические характеристики каждого из узлов веб-ресурса, а также графовые характеристики на основе построенного графа связности страниц сайта. Основные характеристики приведены в таблице 1.

Таблица 1

Основные используемые категории характеристик

Категория	Описание основных признаков
Структурные	Поля HTTP запроса; данные о браузере; количество запросов в сессии; типы файлов; номера ошибок.
Временные	Частота и длительность запросов; распределение запросов в сессии; время происхождения запроса.
Графовые	Степени входов и исходов посещённых вершин; эксцентриситеты; значение мер центральности; значения алгоритма HITS; PageRank пройденных вершин; переходы между несвязанными страницами.
Семантические	Число тематик; число уникальных тематик; подобие тематик в сессии; вариативность тематик; распределение переходов между тематиками.
Поведенческие	Информация об источнике запросов; JavaScript метрики; запросы файлов-детекторов.

Для расчёта данных характеристик в рамках сессии строятся комбинации из средних и медианных значений параметров каждого запроса, а также изучается их распределение и среднеквадратическое отклонение значений.

На втором этапе происходит формирование профиля для легитимных пользователей и веб-роботов. Рассчитываются сессионные характеристики, учитывающие распределение значений параметров каждого из запросов в рамках сессии. На основе данных характеристик, а также достоверной информации о происхождении сессии формируется классификационная модель. На третьем этапе происходит процедура идентификации сессий [5]. Для каждой сессии вычисляется результат комбинации решений о принадлежности пользователя к роботизированным сессиям по приведённой формуле, а также происходит выбор подходящего сценария реагирования.

$$\sum_{i=1}^{n} p_i \times P(y = 1 \mid using\ i^{th}\ \text{method}) \quad (1)$$

Весовой параметр p_i подбирается экспертным образом, где $\sum_{i=1}^{n} p_i = 1$. n – количество используемых методов обнаружения

и противодействия, *y* – результат классификации (равен 1, если сессия отнесена к роботизированной). *P* – вероятность отнесения сессии к роботизированной.

На четвёртом этапе происходит реагирование (рисунок 1) и формируется сообщение об инциденте.

Рис. 1. Схема процесса противодействия

Выбор процедуры реагирования состоит из следующих шагов:

1. Определение коэффициента ущерба от пропуска веб-робота (ошибка первого рода);

2. Определение коэффициента ущерба от неверной классификации легитимного пользователя (ошибка второго рода);

3. Выбор порогов срабатывания трёх сценариев реагирования экспертным методом (блокировка, ограничение лимитов действий, проверка на основе CAPTCHA [6]).

Величина ущерба напрямую связана с используемыми стратегии противодействия. Например, в случае использования противодействия в виде показа CAPTCHA, ошибки второго рода несут небольшой репутационный ущерб веб-ресурсу по сравнению с использованием стратегии блокировки по IP адресу. Ущерб от пропуска веб-робота зависит от данных, расположенных на веб-ресурсе, их стоимости и объема недополученной прибыли в связи с кражей информации и возможным появлением ресурсов-агрегаторов. Структура всех подсистем, задействованных в реализации процессов обнаружения и противодействия приведена на рисунке 2.

Мониторинг является отдельным этапом и осуществляется непрерывно. Каждый запрос пользователя к веб-ресурсу отражается в логах веб-сервера. При обнаружении роботизированной сессии создаётся отчёт, содержащий логи запросов данной сессии, вероятность обнаружения и средние статистические характеристики. Также, периодически происходит формирование отчёта о результатах обнаружения.

Результаты. Для проведения эксперимента использовалась система обнаружения и противодействия веб-роботам, использующая комбинацию метода обнаружения, основанного на анализе семантики страниц и запросов, а также метода, основанного на анализе поведения пользователей в сочетании с графом связности страниц веб-ресурса. В качестве входных данных использовались публичные датасеты [7], содержащие логи веб-сервера порталов MSNBC, NASA, а также данные нескольких веб-ресурсов в сети интернет. Общее количество анализируемых сессий составило 258431, точность обнаружения и противодействие предлагаемого метода на основе размеченных данных после применения проверки на тестовом и валидационном множествах, а также проведении 10-ти проходной перекрестной проверки результатов составила 95%. Для классификации использовались несколько различных моделей: Gradient Boosting, XGboost, Multilayer perceptron. В сравнении с результатами существующих методов, не учитывающих семантические и графовые характеристики, увеличение точности составило от 5 до 10%.

Заключение. В данной статье предлагается комбинированный метод обнаружения и противодействия автоматизированному сбору информации с веб-ресурсов, основанный на изучении семантических характеристик страниц веб-ресурса, а также графа связности его страниц. Приводятся основные характеристики обнаружения, а также принцип комбинирования результатов обнаружения на основе нескольких методов. Описывается предлагаемая схема комплексной системы обнаружения и противодействия веб-

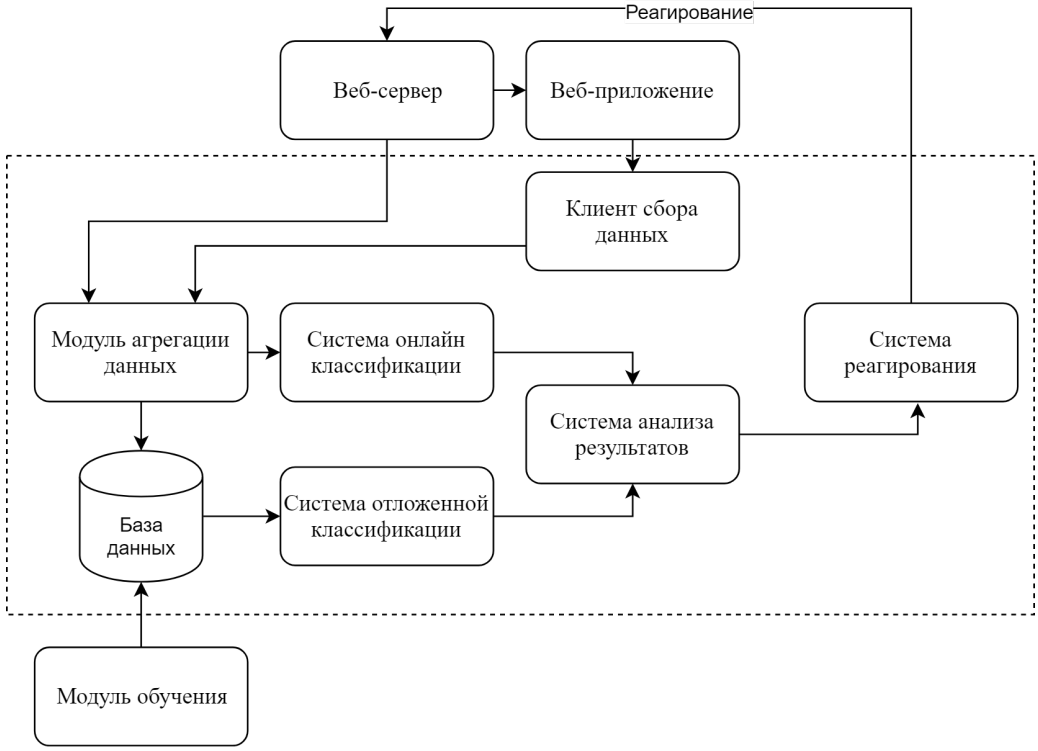

Рис. 2. Структура системы реагирования

роботам, алгоритм выбора сценария реагирования и экспериментальные результаты тестирования системы, использующей предлагаемый метод. Результаты позволяют говорить о теоретической и практической значимости данного подхода, а также применимости для задачи нейтрализации автоматизированных угроз.

Литература

1. Menshchikov A. et al. A study of different web-crawler behaviour //2017 20th Conference of Open Innovations Association (FRUCT). – IEEE, 2017. – Pp. 268-274.

2. OWASP Automated threat Handbook 2018. URL: https://www.owasp.org/images/3/33/Automated-threat-handbook.pdf (дата обращения: 03.06.2019).

3. Zabihimayvan M. et al. A soft computing approach for benign and malicious web robot detection // Expert Systems with Applications. – 2017. – Vol. 87. – Pp. 129-140.

4. Doran D., Morillo K., Gokhale S. S. A comparison of web robot and human requests //Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining. – ACM, 2013. – Pp. 1374-1380.

5. Hamidzadeh J., Zabihimayvan M., Sadeghi R. Detection of Web site visitors based on fuzzy rough sets //Soft Computing. – 2018. – Vol. 22. – №. 7. – Pp. 2175-2188.

6. Bursztein E., Martin M., Mitchell J. Text-based CAPTCHA strengths and weaknesses //Proceedings of the 18th ACM conference on Computer and communications security. – ACM, 2011. – Pp. 125-138.

7. UC Irvine Machine Learning Repository. URL: https://archive.ics.uci.edu/ml/index.php (дата обращения: 03.06.2019).

References

1. Menshchikov A. et al. A study of different web-crawler behaviour //2017 20th Conference of Open Innovations Association (FRUCT). – IEEE, 2017. – Pp. 268-274.

2. OWASP Automated threat Handbook 2018. Available at: https://www.owasp.org/images/3/33/Automated-threat-handbook.pdf (accessed: 03 June 2019).

3. Zabihimayvan M. et al. A soft computing approach for benign and malicious web robot detection // Expert Systems with Applications. – 2017. – Vol. 87. – Pp. 129-140.

4. Doran D., Morillo K., Gokhale S. S. A comparison of web robot and human requests //Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining. – ACM, 2013. – Pp. 1374-1380.

5. Hamidzadeh J., Zabihimayvan M., Sadeghi R. Detection of Web site visitors based on fuzzy rough sets //Soft Computing. – 2018. – Vol. 22. – №. 7. – Pp. 2175-2188.

6. Bursztein E., Martin M., Mitchell J. Text-based CAPTCHA strengths and weaknesses //Proceedings of the 18th ACM conference on Computer and communications security. – ACM, 2011. – Pp. 125-138.

7. UC Irvine Machine Learning Repository. Available at: https://archive.ics.uci.edu/ml/index.php (accessed: 03 June 2019).

МЕНЩИКОВ Александр Алексеевич, аспирант, Университет ИТМО. 197101, г. Санкт-Петербург, Кронверкский пр., 49. E-mail: menshikov@corp.ifmo.ru

ГАТЧИН Юрий Арменакович, доктор технических наук, профессор, Университет ИТМО. 197101, г. Санкт-Петербург, Кронверкский пр., 49. E-mail: gatchin@mail.ifmo.ru

КОРОБЕЙНИКОВ Анатолий Григорьевич, доктор технических наук, профессор, заместитель директора по науке, Санкт-Петербургский филиал Федерального государственного бюджетного учреждения науки Института земного магнетизма, ионосферы и распространения радиоволн им. Н.В.Пушкова Российской академии наук. 199034, г. Санкт-Петербург, Менделеевская линия, 3. E-mail: korobeynikov_a_g@mail.ru

MENSHCHIKOV Alexander, postgraduate student, St. Petersburg National Research University of Information Technologies, Mechanics and Optics. 197101, St. Petersburg, Russia. Kronverksky pr., 49. E-mail: menshikov@corp.ifmo.ru

GATCHIN Yurij, Dr.Sc., Professor, St. Petersburg National Research University of Information Technologies, Mechanics and Optics. 197101, St. Petersburg, Russia. Kronverksky pr., 49. E-mail: gatchin@mail.ifmo.ru

KOROBEYNIKOV Anatoly, Dr.Sc., Professor, Deputy Director for Science, Pushkov Institute of Terrestrial Magnetism, Ionosphere and Radio Wave Propagation of the Russian Academy of Sciences (IZMIRAN). 199034, St. Petersburg, Russia. Mendeleevskaya liniya, 3. E-mail: korobeynikov_a_g@mail.ru