УДК 004.056
DOI: 10.15827/0236-235X.128.607-612
Дата подачи статьи: 08.04.19
2019. Т. 32. № 4. С. 607–612

# *Метод обнаружения веб-роботов на основе анализа графа пользовательского поведения*

*А.А. Менщиков* [1], *аспирант, menshikov@itmo.ru*
*Ю.А. Гатчин* [1], *д.т.н., профессор, gatchin@mail.ifmo.ru*

[1] *Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Университет ИТМО), г. Санкт-Петербург, 197101, Россия*

Согласно отчетам компаний, занимающихся защитой веб-ресурсов, каждый пятый запрос к типовому сайту в сети Интернет направлен вредоносными автоматизированными системами – веб-роботами. Веб-роботы по объему трафика уже преобладают над рядовыми пользователями веб-ресурсов. Своими действиями они угрожают приватности данных, авторскому праву, несут угрозы несанкционированного сбора информации, влияют на статистики и приводят к ухудшению производительности веб-ресурса. Возникает необходимость обнаружения и блокирования источников таких средств.

Существующие методы предполагают использование синтаксической и аналитической обработки логов веб-сервера для обнаружения веб-роботов. Этого недостаточно, чтобы надежно выявлять веб-роботов, скрывающих свое присутствие и имитирующих поведение легитимных пользователей.

В статье предлагается новый метод, отличительной особенностью которого является использование характеристик графа связности страниц защищаемого веб-ресурса для формирования признаков роботизированных пользовательских сессий. Характеристики анализируемых сессий включают не только особенности графа перемещений самого пользователя, но и признаки каждого из посещенных им узлов веб-ресурса (степени входа и исхода, меры центральности и другие). Для расчета таких характеристик строится граф связности страниц веб-ресурса.

Данный метод заключается в анализе характеристик перемещений для каждой пользовательской сессии с целью классификации ее на роботизированную или принадлежащую легитимному пользователю.

В статье проводится анализ шаблонов поведения пользователей веб-ресурса, описываются основные принципы извлечения необходимых данных из логов веб-сервера, способ построения графа связности страниц веб-ресурса, а также наиболее значимые характеристики сессий. Обсуждаются процедура обнаружения и выбор подходящей классификационной модели. Для каждой из исследуемых моделей производятся отбор гиперпараметров и перекрестная проверка результатов. Анализ точности и полноты обнаружения показывает, что при использовании библиотеки XGboost можно получить $F_1$-меру порядка 0.96.

***Ключевые слова:*** *веб-роботы, информационная безопасность, защита веб-ресурсов, парсеры, обнаружение веб-роботов, граф веб-ресурса, теория графов, защита информации.*

Посетителей сегодняшних веб-ресурсов можно условно разделить на две категории: легитимные пользователи, совершающие действия при помощи веб-браузеров и мобильных приложений, и веб-роботы, выполняющие на сайте автоматизированные действия [1]. Веб-роботы могут выступать в роли индексаторов ресурса, проверять ссылки и работоспособность функционала, но могут и нести различные автоматизированные угрозы – от кражи информации до совершения мошеннических действий и манипуляций с целью получения преимущества над обычными пользователями [2].

Отчеты компаний, которые занимаются мониторингом Интернета, показывают, что до 50 % трафика на сайте приходит от веб-роботов [3]. Различие статистических параметров поведения пользователей веб-ресурсов и веб-роботов можно использовать также для улучшения системы кэширования и настройки систем управления статистикой для исключения веб-роботов из различных маркетинговых отчетов [4].

Типичный веб-ресурс – это ориентированный граф, узлами которого являются веб-страницы с информацией (HTML-страницы, документы, файлы, изображения, скрипты), а ребро проводится из узла, где есть гиперссылка в узел, на который она ведет. Ссылкой также могут являться вложение ресурса (например, изображения) и переход, выполняемый из JavaScript-сценария [5].

Знание о структуре веб-ресурса и данные о поведении легитимных пользователей на нем можно использовать для формирования модели поведения. Отличие поведения пользователя от данной модели позволяет сделать вывод об автоматизации его посещений и о других целях перемещения по сайту. Данные факты могут быть использованы для обнаружения веб-роботов и уточнения классических синтаксических и аналитических методов классификации пользователей.

***Обнаружение веб-роботов.*** Обнаружение происходит на основе анализа данных о пользователе. Часто такими данными являются обычные логи веб-сервера, но это могут быть дампы трафика или данные уровня приложений [6].

Логи веб-сервера – это наборы строк, содержащих следующие данные о каждом запросе к веб-ресурсу: дата, путь до запрашиваемого узла, код ответа, страница, с которой совершен переход, браузер пользователя, IP-адрес источника, уникальный идентификатор сессии (если настроен).

Для каждой пользовательской сессии рассчитываются уникальные характеристики, описывающие поведение данного пользователя на веб-ресурсе. На их основе каждая сессия классифицируется на легитимную или роботизированную.

Классические методы обнаружения сегодня используют данные пользовательских запросов и логов без привязки к реальной структуре и контенту, расположенному на веб-ресурсе [7]. В основном исследователи применяют различные методы классификации или кластеризации на основе информации, полученной из веб-логов. Такие подходы позволяют добиваться точности обнаружения вплоть до 0.9 [8, 9], однако результаты очень зависят от набора данных и наличия в нем сложных веб-роботов, скрывающих свое присутствие [10].

***Построение графа веб-ресурса.*** Отдельной задачей является получение графа веб-ресурса (рис. 1). Для этого могут использоваться как внешние системы (краулер, совершающий обход всех страниц веб-ресурса), так и внутренний подход (генерация графа связности страниц через расширенные возможности фреймворка, на котором основан веб-ресурс). Связность также можно генерировать на основе пользовательских сессий, но такой подход приводит к ошибкам и ложным срабатываниям за счет устаревания данных в логах, нерелевантных запросов от веб-роботов, скрывающих

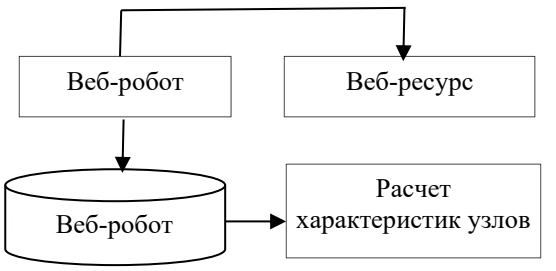



*Рис. 1. Процесс сбора графовых характеристик узлов веб-ресурса*

*Fig. 1. The process of collecting graph characteristics of web resource nodes*

свое присутствие, и переходов из закладок браузера.

Веб-ресурс можно представить в виде ориентированного графа $G = (V, E)$, который для удобства опишем матрицей смежности $A$ (рис. 2).

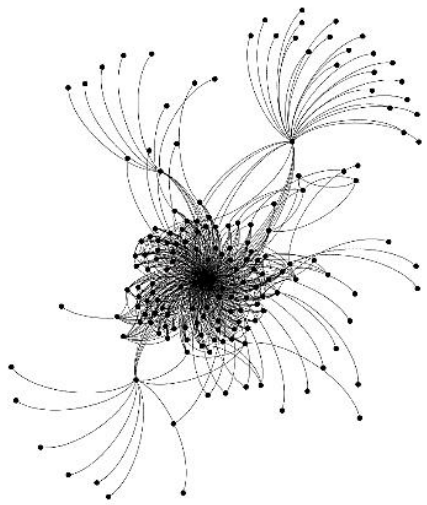


*Рис. 2. Граф страниц веб-ресурса*

*Fig. 2. A web resource page graph*

Для каждого узла рассчитываются следующие характеристики, которые затем будут использоваться для формирования признаков пользовательских сессий:

– степени входа и исхода каждой вершины:

$$k_i^{in} = \sum_{j=1}^{|V|} A_{ji}; \quad (1)$$

$$k_i^{out} = \sum_{j=1}^{|V|} A_{ij}; \quad (2)$$

– эксцентриситет вершины;

– меры центральности (Closeness centrality, Betweenness centrality, Harmonic centrality, eigencentrality);

– значения алгоритма HITS (ранги авторов и посредников) [11];

– PageRank [12]:

$$R'(u) = c\sum_{v\in B_u}\frac{R'(v))}{N_v} + cE(u). \quad (3)$$

***Характеристики сессий.*** Каждая сессия представляет собой набор запросов к веб-ресурсу от одного источника за определенный временной интервал. Каждый запрос направлен к определенному узлу графа, что позволяет оценивать изменение характеристик каждой из вершин графа, а также вычислять комбинированные показатели.

В результате каждую из сессий можно характеризовать следующими типами признаков, на основе которых использовать классификацию:

– среднее значение каждой из характеристик каждого узла;

– среднеквадратические отклонения характеристик узлов;

– распределение значений по каждой характеристике;

– дополнительные характеристики переходов между узлами.

Изучение распределений значений для различных характеристик позволяет утверждать, что шаблоны поведения легитимных пользователей и веб-роботов отличаются (рис. 3).

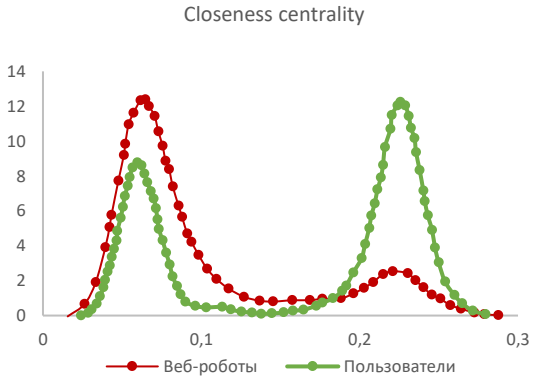
К дополнительным характеристикам переходов относятся технические особенности перемещения по графу:

– количество переходов между страницами, не связанными ссылкой;

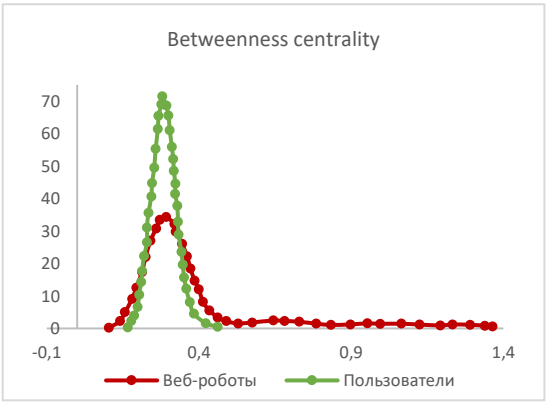
– количество возвратов на предыдущую страницу.

В данном исследовании не рассматривается временной контекст каждого из запросов, однако стоит отметить, что учет временных интервалов между разными типами запросов может принести дополнительные знания о том, как быстро пользователи принимают те или иные решения о возврате на предыдущую страницу или о переходе на главную страницу.

***Сравнительный анализ.*** В исследовании использовался архив трафика к веб-ресурсу за один месяц. Архив содержит HTTP-запросы к сайту от шестидесяти тысяч источников за рассматриваемый период. Веб-ресурс использует специальное ПО для идентификации сессий, что позволяет однозначно идентифицировать




*а)*




*б)*

*Рис. 3. Разница значений характеристик а) Closeness centrality и б) Betweenness centrality для людей и веб-роботов*

*Fig. 3. The difference in the values of the characteristics: a) Closeness centrality and б) Betweenness centrality for people and web robots*

связанные сессии легитимных пользователей без использования нечетких алгоритмов идентификации.

С помощью разработанного ПО в полуавтоматическом режиме производилась предварительная классификация сессий с использованием как однозначных признаков веб-робота (запросы к файлам-ловушкам, известные адреса источников, известные значения User-Agent), так и дополнительных параметров, оцениваемых человеком (повторяемость запросов, аномалии поведения, исполнение JavaScript, географическая привязка источника и другие).

При помощи ExtraTreesClassifier оценивались значимости всех непосредственных и усредненных графовых признаков для уменьшения признакового пространства (см. таб-

лицу). Дополнительно производилась оценка корреляции признаков.

**Наиболее значимые признаки**

**The most significant features**

| Признак | Важность |
|---|---|
| clustering_std | 0.060268 |
| harmonicclosnesscentrality_std | 0.054645 |
| eigencentrality_std | 0.054076 |
| degree_std | 0.053189 |
| hub_std | 0.051802 |
| outdegree_std | 0.051132 |
| closnesscentrality_std | 0.050760 |
| authority_std | 0.045223 |
| indegree_std | 0.043336 |
| clustering_avg | 0.043311 |

Для классификации использовались несколько разных моделей: Gradient Boosting, XGboost, Multilayer perceptron. Набор данных был разделен на тренировочный и тестовый. Тренировочный использовался для сравнения моделей классификации и подбора гиперпараметров моделей с использованием кроссвалидации с разделением на 10 блоков. Итоговая оценка производилась на тестовом наборе данных.

Оптимизация гиперпараметров производилась при помощи Grid Search с минимизацией значения площади под кривой ошибок (рис. 4).

Все рассматриваемые модели после оптимизации гиперпараметров показали приемлемые результаты обнаружения. Модель XGboost была наиболее точна с $F_1$-мерой, равной 0.96.

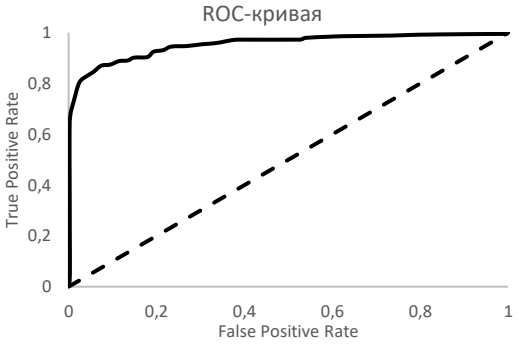



*Рис. 4. ROC-кривая*

*Fig. 4. ROC curve*

**Выводы**

В статье предложен метод обнаружения веб-роботов на основе анализа графа пользовательского поведения. За счет анализа связности страниц веб-ресурса и расчета характеристик графа веб-ресурса удалось добиться улучшения точности и полноты обнаружения веб-роботов. Данные характеристики могут быть скомбинированы с классическими методами и приводить к улучшению показателей обнаружения веб-роботов. Были использованы несколько методов классификации, произведены подбор гиперпараметров, а также перекрестная проверка результатов обнаружения. В итоге достигнута $F_1$-мера обнаружения веб-роботов, равная 0.96, что превышает существующие показатели методов, основанных на синтаксическом и аналитическом обнаружении.

*Работа выполнена в рамках гранта РФФИ № 17-07-00700-а «Методы формальной и функциональной верификации вычислительных процессов, основанные на знаниях и графоаналитических моделях».*

***Литература***

1. Stassopoulou A., Dikaiakos M.D. Web robot detection: A probabilistic reasoning approach. Computer Networks, 2009, vol. 53, no. 3, pp. 265–278.

2. Zabihimayvan M., Rude H.N., Sadeghi R., Doran D. A soft computing approach for benign and malicious web robot detection. Expert Syst. with Appl. 2017, vol. 87, pp. 129–140. DOI: 10.1016/j.eswa.2017.06.004.

3. Bad bot report. URL: https://resources.distilnetworks.com/travel/2018-bad-bot-report (дата обращения: 10.03.2019).

4. Xie N., Rudeet N., Brown K., Doran D. A soft computing prefetcher to mitigate cache degradation by web robots. Proc. Intern. Sympos. Neural Networks. Springer, Cham, 2017, pp. 536–546. DOI: 10.1007/978-3-319-59072-1_63.

5. Wang W., Lei Y., Kuhn D.R., Sampath S., Kacker R., Lawrence J.F. Using combinatorial testing to build navigation graphs for dynamic web applications. Software Testing, Verification and Reliability, 2016, vol. 26, no. 4, pp. 318–346.

6. Jacob G., Kirda E., Kruegel C., Vigna G. PUBCRAWL: Protecting Users and Businesses from CRAWLers, Proc. 21st USENIX Conf. on Security Sympos. 2012, pp. 507–522.

7. Zabihi M., Jahan M.V., Hamidzadeh J. A density based clustering approach to distinguish between web robot and human requests to a web server. The ISC Int. J. Inf. Secur., 2014. vol. 6, no. 1, pp. 77–89.

8. Tan P.N., Kumar V. Discovery of web robot sessions based on their navigational patterns. Intelligent Technologies for Information Analysis, 2004, pp. 193–222.

9. Suchacka G., Sobkow M. Detection of Internet robots using a Bayesian approach. Proc. 2nd CYBCONF. IEEE, 2015, pp. 365–370. DOI: 10.23919/FRUCT.2017.8071322.

10. Menshchikov A., Gatchin Yu., Tishukova N., Komarova A., Korobeynikov A. A study of different web-crawler behavior. Proc. 20th Conf. FRUCT, 2017, pp. 268–274. DOI: 10.23919/FRUCT.2017.8071322.

11. Xing W., Ghorbani A. Weighted pagerank algorithm. Proc. Sec. Conf. Communication Networks and Services Research, IEEE, 2004, pp. 305–314.

12. Page L., Brin S., Motwani R., Winograd T. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab, 1999. URL: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf (дата обращения: 10.03.2019).

Software & Systems
DOI: 10.15827/0236-235X.128.607-612

Received 08.04.19
2019, vol. 32, no. 4, pp. 607–612

## Web-robot detection method based on user's navigation graph

***A.A. Menshchikov*** [1], *Postgraduate Student, menshikov@itmo.ru*
***Yu.A. Gatchin*** [1], *Dr.Sc. (Engineering), Professor, gatchin@mail.ifmo.ru*

[1] *The National Research University of Information Technologies, Mechanics and Optics, St. Petersburg, 197101, Russian Federation*

**Abstract.** According to reports of web security companies, every fifth request to a typical website is from malicious automated system (web robots). Web robots already prevail over ordinary users of web resources in terms of traffic volume. They threaten data privacy and copyright, provide unauthorized information gathering, lead to statistics spoiling, and performance degradation. There is a need to detect and block the source of robots.

The existing methods and algorithms involve syntactic and analytical processing of web server logs to detect web robots. Such approaches cannot reliably identify web robots that hide their presence and imitate the behavior of legitimate users.

This article proposes a method of web-robot detection based on the characteristics of the page web-graph. The characteristics of the analyzed sessions include not only the features of a user web graph, but also parameters of each node visited by him (in and out degrees, centrality measures, and others). To calculate such characteristics, a connectivity graph of pages was constructed.

Based on the analysis of these parameters, as well as the characteristics of the web robot's behavioral graph, the authors make a decision to classify the session.

The authors provide an analysis of different behavioral patterns, describe the basic principles of extracting the necessary data from web server logs, and the method of the connectivity graph construction as well as the most significant features. The paper conciders a detection procedure and selection of an appropriate classification model. For each studied model, the authors select optimal hyperparameters and perform cross-validation of the results. The analysis of the accuracy and precision of such detection shows that the usage of XGboost library allows obtaining $F_1$ measure equals 0.96.

**Keywords:** web-robots, information security, website protection, parsers, web-robot detection, website graph, graph theory, information protection.

***Acknowledgements.*** *The work has been done within the framework of the RFBR grant no. 17-07-00700-a "Methods of formal and functional verification of computational processes based on knowledge and graph-analytical models".*

### *References*

1. Stassopoulou A., Dikaiakos M.D. Web robot detection: A probabilistic reasoning approach. *Computer Networks*. 2009, vol. 53, no. 3, pp. 265–278.

2. Zabihimayvan M., Rude H.N., Sadeghi R., Doran D. A soft computing approach for benign and malicious web robot detection. *Expert Systems with Applications*. 2017, vol. 87, pp. 129–140. DOI: 10.1016/j.eswa.2017.06.004.

3. *Bad Bot Report*. Available at: https://resources.distilnetworks.com/travel/2018-bad-bot-report (accessed March 10, 2019).

4. Xie N., Rudeet N., Brown K., Doran D. A soft computing prefetcher to mitigate cache degradation by web robots. *Intern. Symp. on Neural Networks*. Springer, Cham, 2017, pp. 536–546. DOI: 10.1007/978-3-319-59072-1_63.

5. Wang W., Lei Y., Kuhn D.R., Sampath S., Kacker R., Lawrence J.F. Using combinatorial testing to build navigation graphs for dynamic web applications. *Software Testing, Verification and Reliability*. 2016, vol. 26, no. 4, pp. 318–346.

6. Jacob G., Kirda E., Kruegel C., Vigna G. PUBCRAWL: Protecting Users and Businesses from CRAWLers. *Proc. 21st USENIX Security Symp*. 2012, pp. 507–522.

7. Zabihi M., Vafaei Jahan M., Hamidzadeh J. A density based clustering approach to distinguish between web robot and human requests to a web server. *The ISC Intern. J. of Information Security*. 2014, vol. 6, no. 1, pp. 77–89.

8. Tan P.N., Kumar V. Discovery of web robot sessions based on their navigational patterns. *Intelligent Technologies for Information Analysis*. Springer, Berlin, Heidelberg Publ., 2004, pp. 193–222.

9. Suchacka G., Sobkow M. Detection of Internet robots using a Bayesian approach. *2015 IEEE 2nd Intern. Conf. on Cybernetics (CYBCONF)*. 2015, pp. 365–370. DOI: 10.23919/FRUCT.2017.8071322.

10. Menshchikov A. A study of different web-crawler behavior. *2017 IEEE 20th Conf. of Open Innovations Association (FRUCT)*. 2017, pp. 268–274.

11. Xing W., Ghorbani A. Weighted pagerank algorithm. *IEEE Proc. 2nd Annual Conf. on Communication Networks and Services Research*. 2004, pp. 305–314.

12. Page L., Brin S., Motwani R., Winograd T. *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab, 1999. Available at: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf (accessed March 10, 2019).

**Для цитирования**

Менщиков А.А., Гатчин Ю.А. Метод обнаружения веб-роботов на основе анализа графа пользовательского поведения // Программные продукты и системы. 2019. Т. 32. № 4. С. 607–612. DOI: 10.15827/0236-235X.128.607-612.

**For citation**

Menshchikov A.A., Gatchin Yu.A. Web-robot detection method based on user's navigation graph. *Software & Systems*. 2019, vol. 32, no. 4, pp. 607–612 (in Russ.). DOI: 10.15827/0236-235X.128.607-612.