UDC 004

RESEARCH AND ANALYSIS OF WEB- RESOURCE PROTECTION AGAINST ATTACKS

Zhamisheva Nuray Mazhitkyzy
zhamishevan@gmail.com
L.N.Gumilyov Eurasian National University 2nd year student in the field of Information technology, Nur-Sultan, Kazakhstan
Scientific leader - Tashenova Zhuldyz Mussagulovna

Abstract: The article discusses the most important problems that determine the pace and future development of the Internet, and information security is becoming. Web vulnerabilities today outnumber any other information security issues in terms of number and associated risks. Explored several web application vulnerability detection technologies.

Keywords: web-resource, security, analysis, vulnerability, database, Internet, web application protection

Introduction

These days a problem of unauthorized massive and automated information crawling on the Internet becomes more and more serious. Website protection systems tend to be essential. Important resources and services migrate to the Internet, where they experience a wide variety of risks such as automated information gathering by web-automatons and competitive intelligence. In Russia e-commerce market has a strong tendency to grow. According to different research, its growth is about 15% a year. In 2016 e-commerce revenue accounted for 850 billion rubles. Because of these facts, it is important to provide higher data integrity, confidentiality and availability of websites.

There are special tools for information gathering on the Internet. These programs called web-robots, parsers and crawlers. We can divide them into two groups according to their objectives: robots, used for legal purposes (content analysis, indexing for search systems, site mirroring etc.) and robots, used by criminals.

Web robots can not only collect information, but also actively act on web resources: purchase goods, write advertising messages and comments, send spam, and exploit vulnerabilities. In addition, web robots can respond to intense actions that cause a high load on web servers and therefore slow down the operation of a website, causing access problems for ordinary users. Scrapers require significant bandwidth and usually run on multiple threads over a long period of time. Poorly written searchers can endlessly load dynamic pages or send distorted requests to the web server.

Website owners use robots.the txt file is called the Robot Exclusion Protocol, which provides instructions on how to split your site into robots. About a third of the internet uses this standard for regulation. Not every web robot interacts with the standard, namely email collectors, spambots, malware, and site scanning robots for security vulnerabilities, while other robots ignore these recommendations. Figs. 1. average site traffic analysis, (%)

The volume of scraper traffic detected in 2014 increased by 17% compared to 2015 and by 59% compared to 2010. To disable new sources of scraping, you need to develop new methods for detecting bypassing web pages.

Scrapers have become more aggressive and complex, and they have started using more and more IP addresses to conduct their actions and avoid detection. They also use many infected computers and other devices for their own purposes.

Today, we need comprehensive methods that combine round-the-clock monitoring and international best practices in the field of web robot detection. The business is also interested in protecting information from automated data collection, as this directly affects its profits.

Classification of web robots

One of the most interesting features of web robots is purposefulness. Robots are usually designed for specific purposes related to obtaining specific information by reducing costs and increasing the speed of collection by eliminating incorrect behavior and excessive requests. This behavior is inherent in legal and unethical robots and allows you to track the relationship between behavioral patterns of information processing on a web resource. In other words, it allows you to distinguish robot traffic from human traffic by their behavior. Robots are usually divided into three main categories:

1) amateur web robots that use Direct browsing of web pages and only simple queries. These scrapers have a small amount of loyalty and resources. They are usually used by inexperienced users who do not have a large budget for collecting information.

2) Advanced Web robots that try to act as legitimate users. They use multiple IP addresses and periodically change the paths and viewing methods of the user agent.

3) professional web robots that use complex behavioral algorithms and are often manually configured for each web resource.

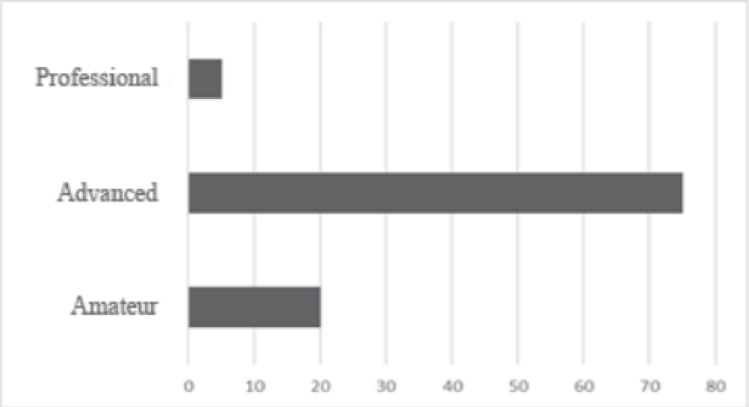

Fig. 2. Scraper types usage, (%)

Fig. 2 shows the percentage of categories of web robots. The most popular are advanced scanners, and the least used are professional analysts.

Attacks using web robots are aimed at obtaining the necessary information on the website. An attacker can use software versions, knowledge of the web server, and information about appropriate updates for the appropriate types of attacks. In addition, the information itself may contain trade secrets and personal data.

Web resources often provide more information than users, and criminals can use this fact for their own purposes. Even the slightest information about the system can lead to full reputation [10]. Today, automated information collection tools allow attackers to massively collect data from various resources.

Sliding systems are used for competitive exploration. Business competitors collect additional information to create their own effective system with stolen content.

Web robots are very dangerous for web resources related to e-commerce. Such resources display unique content that has commercial value, such as:

* Travel
* Online Notifications
* Online Catalogs
* Ticket sales
* Blogs and sites with unique content
* Information resources and libraries
* Social networks
* Other resources containing personal data

Tourism industry companies remain the main targets of scrapers, followed by online catalogs and online advertising. All the worst industries have the same problem. They have a lot of public data and rely on them for their success in business. If competitors or other operators steal data and use it for their own purposes, it negatively affects them and threatens their business model in the future.

Web-robot detection

Web-robot detection methods can be classified by their operational principles, launching strategies and using techniques. According to the first criterion robots are divided into four categories (Fig. 3). The second criterion divides them into active ones, which work during robot query, and delayed ones, which run afterwards. These techniques include filtration, machine learning methods etc.

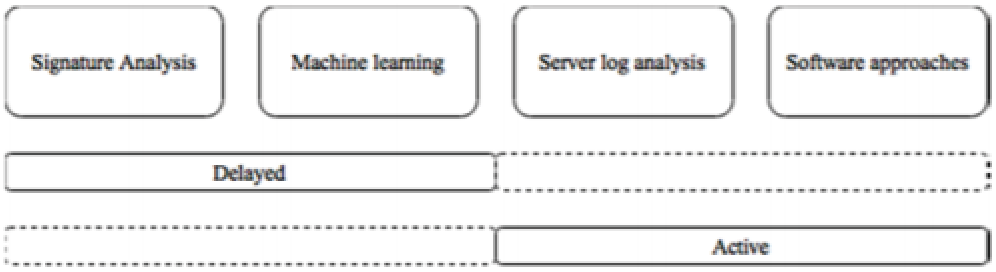

Fig. 3. Detection method classification

Real-time log analysis is a simple processing of web server logs. It includes metrics such as detecting suspicious HTTP headers, analyzing User-Agent and Referer fields, and filtering IP addresses by country or organization. The main advantage of these methods is the ease of implementation and speed of data processing. However, it can only detect known web robots. Log analysis is usually used to identify amateur analysts.

Signature traffic analysis is based on the identification of certain characteristics inherent in robotic systems, as opposed to the human user. For example, using navigation with the same nesting level, the request frequency is too high, and loading only HTML pages (without scripts and CSS files). This method uses deviation of metric values based on the usual user behavior, unlike the previous method, which searches for specific patterns in logs. The advantage of this method is high availability, but it requires sensitivity adjustment for each metric.

Machine learning methods. These methods include analyzing web server traffic and logs. It provides statistical analysis of traffic to identify searchers. It usually uses criteria specific to signature analysis. The advantage of this method is the ability to detect previously unknown analyzers, but manual work requires training and training to achieve the necessary accuracy to detect and eliminate false positives, which can be very difficult.

Traps are a purely technical way to separate a person from a robot. These include The Turing test, a special confusing JavaScript functionality, invisible links, flash applets, local browser storage, cookies,ETag, and real-time log analysis-all these are simple processing of web server logs. It includes metrics such as detecting suspicious HTTP headers, analyzing User-Agent and Referer fields, and filtering IP addresses by country or organization. The main advantage of these methods is the ease of implementation and speed of data processing. However, it can only detect known web robots. Log analysis is usually used to identify amateur analysts.

Crawling challenges

While developing any automated crawling system, it is essential to take into consideration various limiting factors and major challenges the developers of such programs may face with. They are:

1) The necessity of manual configuration and debugging of the system for parsing sites with a complex structure.

2) Information gathering systems have to be able to handle large amounts of data within a short period of time;

3) The design and layout of web-sites can change frequently. It affects to scraping systems and spoil parsing results. Operators have to check them regularly and fix parsing rules manually after every change.

It is important to understand the difficulties the parser developers face with and how they can be used to protect webresources against web-robots. When we increase the cost of web-robot development, the quantity of attackers can be reduced.

Conclusion

The web-robots detection problem requires a whole range of tools. Firstly, web-robots detection methods based on certain parameters and information about their activity. Secondly, a system that helps with the use of these methods, gathers all the necessary information to carry out its preprocessing, processing and decision-making. Third, the framework to adjust the detection system and monitoring of its operations.

The significance of the results is in new methodological approaches and developed tools. They can be used to protect web resources from automated information gathering. We studied a set of web server logs and found robotic sources by comparing the characteristics of the visitor’s behavior. The results allow automatic detection of web robots activity on website and disabling their sources.

This study will serve as a stepping-stone for the construction of an integrated approach to ensure the security of web-resources and for the generation of representative data sets that are necessary for machine learning methods applied to the problem of web-robots detection.

References

1. Report East-West Digital News [Electronic resource] – Mode of access: http://www.ewdn.com/files/ecom-rus-download.pdf/, free (date accessed: 27.10.2016).

2. Report of the company scrapesentry [Electronic resource] – Mode of access: https://www.scrapesentry.com/scrapesentry-scraping-threatreport-2015/, free (date accessed: 27. 10.2016).

3. Report of the Association of companies of Internet trade [Electronic resource] – Mode of access: http://www.akit.ru/wpcontent/uploads/2016/05/E-commerce_1Q2016-FINAL.pdf free (reference date: 01.11.2016).

4. Menshchikov A.A., Gatchin YU.A. Metody obnaruzheniya avtomatizirovannogo sbora informacii s veb-resursov // Kibernetika i programmirovanie [Cybernetics and programming]. – 2015. – ʋ 5. – S. 136-157

5. Junsup Lee, Sungdeok Cha, Dongkun Lee, Hyungkyu Lee, Classification of web robots: An empirical study based on over one billion requests // Computers & Security. – 2009. – V. 28. – ʋ 8. – P. 795-802.

6. Robots Exclusion Protocol Guide [Electronic resource]. – Mode of access: http://www.bruceclay.com/seo/robots-exclusion-guide.pdf free (reference date: 01.11.2016).

7. Sun, Yang, Isaac G. Councill, and C. Lee Giles. "The ethicality of web crawlers." Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. Vol. 1. IEEE, 2010.

8. D. Derek, S. Gokhale A Classification Framework for Web Robots // Journal of American Society of Information Science and Technology. – 2012. – V. 63. – P. 2549–2554.

9. 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. – P. 25–36.