Research on the application of crawler technology in machine learning

Lei Zhang*, Zening Chen and Shufeng Yang
Shenyang Fire Science and Technology Research Institute of MEM, Shenyang, China

*Corresponding author: leizhang@syfri.cn

Abstract. With the rapid development of the Internet, the world wide web has become the carrier of a large number of information, and the search engine has entered the public life. In order to help search engines capture related web resources, web crawlers emerge as the times require. Web crawler is a program that automatically extracts web pages. Some malicious web crawlers will not only steal the information of the website, but also cause irreparable damage to the website. Along with it, anti-crawler, the enemy of web crawler, comes into our life. This paper studies the use of machine learning for crawler recognition, so as to make a better anti-crawler strategy, use machine recognition instead of manual recognition, reduce the workload of manual recognition of crawler, reduce the loss of human and material resources, and improve the recognition rate of crawler, so as to protect the website information and website equipment.

1. Introduction
At this stage, with the continuous development of science and technology, people's search engines have become diverse. In order to quickly obtain a large amount of target data from the Internet, we need to write some script programs to obtain data in batches according to certain rules, which is the so-called crawler [1]. The emergence of web crawlers not only benefits the society, but also pollutes the network environment. Malicious crawlers will seriously increase the burden of the website server, and may also cause the emergence of botnets. Not only that, the emergence of malicious crawler also provides a convenient way to steal other people's important website data. Illegal crawling of important website data will cause the disclosure of important website data and user information, leading to significant business losses. For web crawlers, anti-crawler mechanism is born. Nowadays, most web crawlers still rely on human identification, which is time-consuming and laborious. In recent years, machine learning appears in people's life because it can greatly reduce the human workload and reduce the loss of human and material resources. This paper will study the application strategy of anti-crawler based on machine learning. According to the characteristics of web crawlers, machine learning can automatically identify malicious web crawlers, so as to realize the anti-crawler strategy, which can better protect the website and its important data.

2. Realization of machine learning crawler recognition
At present, artificial intelligence has created a very considerable economic benefits for human beings. Artificial intelligence can replace human beings to do a lot of work that human beings don't want to do and can't do. Moreover, the probability of machine making mistakes is lower than that of human beings,
and it can work continuously, greatly improving work efficiency and saving a lot of costs. Machine learning (ML) is a multidisciplinary interdisciplinary, involving probability theory, statistics, approximation theory, convex analysis, algorithm complexity theory and other disciplines, is a branch of artificial intelligence.

In order to better identify the web crawler and implement the anti-crawler strategy, the characteristics of the web crawler are collected. At present, web crawlers on the network have the following characteristics: ① the request frequency of the same IP is high; ② the time interval of each visit of the same IP is small; ③ the IP location is unstable; ④ the user agent is not a common identifier; ⑤ the number of requests for verification code is high; ⑥ the trap of activating crawler; ⑦ the percentage of map access is high; ⑧ the error response is high; and ⑨ the error response is not correct robot.txt Visit.

Decision tree is one of the classical algorithms of machine learning. Decision tree is composed of a decision graph and possible results (including resource cost and risk), which is used to create a plan to reach the goal. Decision tree is a special tree structure, which is established and used to assist decision-making. According to the characteristics of the web crawler, the decision tree algorithm is used to train the training set to generate a decision model of probability operation, and then the decision model is used to identify the web crawler.

By connecting to the website database that needs to detect the web crawler, we traverse all the access request characteristics of the website that needs to be detected, and then use the decision number to analyze the collected access request characteristics through the identification module trained by the training set, so as to realize the identification of the web crawler in the access, and finally output the detection results and judgment basis of the corresponding IP. Through such feedback, the website administrator can restrict the access to the website, manually correct errors and other related operations. Using machine learning to identify crawlers can not only reduce the labor cost and improve the efficiency of the website, but also avoid the misjudgment of crawler detection of special access nodes.

3. Implementation of anti-crawler strategy
There is a spear, there is a shield. The counter mechanism of web crawler is a kind of strategy against malicious crawler and to prevent crawler from infringing the website. Anti-crawler mechanism is a series of strategies to prevent web crawlers according to their characteristics. Anti-crawler mechanism greatly reduces the harm of crawler to the website, protects the data security of the website to a certain extent, and reduces the burden of the website server. But at this stage, most of the anti-crawler work still relies on human operation, artificial judgment, identification of crawlers, so as to achieve the ban of crawlers and other operations, wasting a lot of human, material and financial resources, and the efficiency of identification of crawlers is also in a low state.

Through machine learning to identify web crawlers, not only can website managers limit suspicious IP and other related operations, but also website designers can optimize the anti-crawler mechanism in their own system according to the output information after identifying crawlers, so as to better protect the security of the website. If the IP access frequency is too high, you can increase the limit of IP access frequency on the server. If the IP access frequency exceeds a certain frequency, it will be considered as a web crawler, so as to defend. Website managers can also directly analyze the system's crawler detection data according to the crawler detection, and design a reasonable theme anti-crawler scheme, so as to more effectively protect the website, reduce the infringement of website hardware resources and important data theft by web crawlers, and make the website maintain a competitive advantage in the trend of the Internet era.

4. Conclusions
At present, more than 60% of the access requests on the network come from crawler robots, and many of them are malicious crawlers. This kind of crawler has some commonalities, such as stealing site data, stealing sensitive information, attacking the site and so on. [3] Because the strategy of web crawler is to "crawl" as much valuable information as possible, it will visit as many pages as possible according to
the specific strategy, occupy the network bandwidth and increase the processing cost of the web server. Many small site owners find that when the web crawler patronizes, the access traffic will increase significantly. Malicious users can use crawlers to launch DoS attacks on Web sites, which makes web services run out of resources and cannot provide normal services under the violent access of a large number of crawlers. Malicious users may also use web crawlers to capture all kinds of sensitive information for improper purposes. Web crawler and its corresponding technology bring considerable visits to the website, but also bring direct and indirect security threats. More and more websites begin to pay attention to the restriction of web crawler. In the network world, as a site administrator, it is very important to ensure the security of the website and its data. This paper studies the combination of machine learning and web crawler recognition, which makes the recognition of web crawler more effective and intelligent, reduces the infringement of web crawler on the website, and greatly reduces the error of human identification of web crawler. To a certain extent, it prevents the infringement of web crawler on the website, and achieves the purpose of data protection, system stability guarantee and competitive advantage maintenance.

Acknowledgments
This work was financially supported by Fire Rescue Bureau of emergency management department Science and technology projects, 2020XFZD11.

References
[1] E. Walker. Benchmarking amazon ec2 for high-performance scientific computing[J]. USENIX, 2018, 72(22): 87-89 Sci. Commun. 163 (2000) 51-59.
[2] O. Papapetrou, G. Samaras. IPMicra: An IP-address based location aware distributed Webcrawler[C]. Proceedings of the 5th International Conference on Internet Computing, Las Vegas, 2014, 694-699
[3] A. Singh, M. Srivatsa, L. Liu, et al. Apoidea: A decentralized peer-to-peer architecture for crawling the World Wide Web[C]. Proceedings of the SIGIR 2003 Workshop on Distributed Information Retrieval, Toronto, 2013, 126-142
[4] J. Zawodny. Redis: Lightweight key/value store that goes the extra mile[J]. Linux Magazine, August, 2019(31):77-79
[5] J. Dean, S. Ghemawat. MapReduce: simplified data processing on large clusters[C]. Proceedings of the 6th conference on Symposium on Operating Systems Design & Implementation, San Francisco, 2004, 382-387