Expert Systems with Applications 39 (2012) 8707–8717

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

# Feature evaluation for web crawler detection with data mining techniques

Dusan Stevanovic *, Aijun An, Natalija Vlajic

*Department of Computer Science and Engineering, York University, 4700 Keele St., Toronto, Ontario, Canada M3J 1P3*

ARTICLE INFO

*Keywords:*
Web crawler detection
Web server access logs
Data mining
Classification
DDoS
WEKA

ABSTRACT

Distributed Denial of Service (DDoS) is one of the most damaging attacks on the Internet security today. Recently, malicious web crawlers have been used to execute automated DDoS attacks on web sites across the WWW. In this study we examine the effect of applying seven well-established data mining classification algorithms on static web server access logs in order to: (1) classify user sessions as belonging to either automated web crawlers or human visitors and (2) identify which of the automated web crawlers sessions exhibit 'malicious' behavior and are potentially participants in a DDoS attack. The classification performance is evaluated in terms of classification accuracy, recall, precision and $F_1$ score. Seven out of nine vector (i.e. web-session) features employed in our work are borrowed from earlier studies on classification of user sessions as belonging to web crawlers. However, we also introduce two novel web-session features: the consecutive sequential request ratio and standard deviation of page request depth. The effectiveness of the new features is evaluated in terms of the information gain and gain ratio metrics. The experimental results demonstrate the potential of the new features to improve the accuracy of data mining classifiers in identifying malicious and well-behaved web crawler sessions.



## 1. Introduction

Today, the world is highly dependent on the Internet, the main infrastructure of the global information society. Therefore, the availability of Internet is very critical for the economic growth of the society. For instance, the phenomenal growth and success of Internet has transformed the way traditional essential services such as banking, transportation, medicine, education and defence are operated. These services are now being actively replaced by cheaper and more efficient Internet-based applications. However, the inherent vulnerabilities of the Internet architecture provide opportunities for various attacks on the security of Internet-based applications. For example, distributed denial-of-service (DoS) is a type of security attack that poses an immense threat to the availability of any Internet-based service and application. The DoS effect is achieved by sending a flood of messages to the target (e.g., a machine hosting a web site) with the aim to interfere with the target's operation, and make it hang, crash, reboot, or do useless work (see Fig. 1). In general, single-source DoS attacks can be easily prevented by locating and disabling the source of the malicious traffic. However, distributed DoS (DDoS) attacks launched from hundreds to tens of thousands of compromised zombies can present a much more complex challenge. Namely, unlike in the single-source DoS attack scenarios, the problem of locating the malicious hosts responsible for a DDoS attack becomes extremely difficult due to the sheer number of hosts participating in the attack. Furthermore, the larger collection of malicious hosts can generate enormous amount of traffic towards the victim. The result is a substantial loss of service and revenue for businesses under attack. According to the United States' Department of Defence report from 2008 presented in Wilson et al. (2008), cyber attacks from individuals and countries targeting economic, political, and military organizations may increase in the future and cost billions of dollars.

Now, attackers launching the traditional DDoS attacks by employing illegal Network Layer packets can be easily detected (but not easily stopped) by the signature detections systems such as intrusion detection systems. However, an emerging (and increasingly more prevalent) set of DDoS attacks known as Application Layer or Layer-7 attacks are shown to be particularly challenging to detect. The traditional network measurement systems often fail to identify the presence of Layer-7 DDoS attacks. The reason for this is that in an Application Layer attack, the attacker utilizes a legitimate network session. More specifically, the attacker utilizes a web crawler[1] program that performs a clever semi-random walk of the web site links, intended to resemble the web site traversal of an

* Corresponding author. Tel.: +1 416 736 2100x70143.
*E-mail address:* dusan@cse.yorku.ca (D. Stevanovic).

[1] Crawlers are programs that traverse the Internet autonomously, starting from a *seed* list of web pages and recursively visit documents accessible from that list. Crawlers are also referred to as robots, wanderers, spiders, or harvesters. Their primary purpose is to discover and retrieve content and knowledge from the Web on behalf of various Web-based systems and services. For instance, search-engine crawlers seek to harvest as much Web content as possible on a regular basis, in order to build and maintain large search indexes. On the other hand, shopping bots crawl the Web to compare prices and products sold by different e-commerce sites.

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.01.210

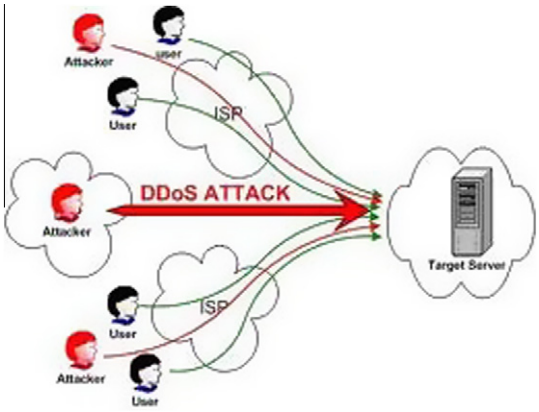


**Fig. 1.** Application layer denial of service attack.

actual human user. Since such attack signatures look very much like legitimate traffic, it is difficult to construct an effective metric to detect and defend against the Layer-7 attacks.

Numerous studies have been published on the topic of Layer-7 DDoS attacks. Given that the key challenge of Layer-7 DDoS attacks is their close similarity to the patterns of legitimate user traffic; researchers studying Layer-7 defence mechanism are mostly interested in devising effective techniques of attack detection. More specifically, the research works in this field are categorized into two main groups: (1) detection of application-layer DDoS attacks during a *flash crowd* event based on aggregate-traffic analysis (Oikonomou & Mirkovic, 2009; Xie & Yu, 2009) and (2) differentiation between wellbehaved and malicious web crawlers based on web-log analysis (Bomhardt, Gaul, & Schmidt-Thieme, 2005; Hayati, Potdar, Chai, & Talevski, 2010; Park, Pai, Lee, & Calo, 2006). (A more detailed overview of the works from the latter group is provided in Section 2.)

In this study, we pursue the line of research of the second group of works further, through two sets of experiments. In particular, the goal of the first set of experiments is to: (1) examine the effectiveness of seven selected classification algorithms in detecting the presence of (i.e. distinguish between) known well-behaved web crawlers and human visitors and (2) evaluate the potential of two newly proposed web-session features to improve the classification accuracy of the examined algorithms. The goal of the second experiment is to: (1) examine the effectiveness of seven classification algorithms in distinguishing between four visitor groups to a web site (malicious web crawlers, well-behaved web crawlers, human visitors and unknown visitors (either human or robot)) and (2) evaluate the potential of two newly proposed web-session features to improve the classification accuracy of the examined algorithms in this particular case. The datasets used in the experiments are generated by pre-processing web server access log files. The implementations of classification algorithms are provided by WEKA data mining software (WEKA, 2010).

The novelty of our research is twofold. Firstly, to the best of our knowledge, this is the first study that looks into the detection of the so-called malicious web crawlers, i.e. crawlers used to conduct Layer-7 attacks, and ways of distinguishing them from well-behaved web robots (such as Googlebot and MSNbot among others). Secondly, in addition to employing traditional web-session features in our classification, we also introduce two new features and prove that the utilization of these features can improve the classification accuracy of the examined algorithms.

The paper is organized as follows: In Section 2, we discuss previous works on web crawler detection. In Section 3, we discuss the advantages of utilizing supervised learning for the purpose of web visitor detection over using a simple rule-based web-log analyzer. In Section 4, we present an overview of our web-log analyzer and the process of dataset generation and labelling. In Section 5, we outline the design of the experiments and the performance metrics that were utilized. In Section 6, we present and discuss the results obtained from the classification study. In Section 7, we conclude the paper with our final remarks.

## 2. Related work

In over the last decade, there have been numerous studies that have tried to classify web robots from web server access logs. One of the first studies on classification of web robots using data mining classification techniques is presented in Tan and Kumar (2002). In this study, the authors attempt to discover web robot sessions by utilizing feature vectors derived from a number of different properties of recorded user sessions. In the first step, the authors propose a new approach to extract sessions from log data. They argue that the standard approach based on grouping web log entries according to their IP address and user-agent fields may not work well since an IP/user-agent pair may contain more than one session (for example, sessions created by web users that share the same proxy server). Next, authors derive 25 different properties of each session by breaking down the sessions into episodes, where an episode corresponds to a request for an HTML file. Among 25 different properties or features, authors identify three features that, in their belief, most distinctly represent sessions likely to be robots, and therefore can be used for the purposes of class labelling. These three features are: (1) checking if robots.txt (file that lists pages that may be accessed by the robots) was accessed, (2) the percentage of page requests made with the HTTP method of type HEAD and (3) percentage of requests made with an unassigned referrer field. These features most distinctly represent sessions likely to be robots since normally a human user would not request robots.txt, send a large number of HEAD requests, or send requests with unassigned referrer fields. As a result of the initial class labelling, the observed user sessions are partitioned into groups of known robots, known browsers, possible robots, and possible browsers. Finally, the technique adopts the C4.5 decision tree algorithm over the labelled human and robot sessions using all of the 25 derived navigational features. This classification model when applied to a dataset suggests that robots can be detected with more than 90% accuracy after only four requests.

In addition to C4.5, other data mining techniques have also been used for the purposes of log-session classification. In Bomhardt et al. (2005) and Stassopoulou and Dikaiakos (2009), for example, authors utilize Bayesian classification and neural networks respectively, to detect web robot presence in web server access log files. Many of the features used in Bomhardt et al. (2005) and Stassopoulou and Dikaiakos (2009) overlap with those from (Tan & Kumar, 2002), indicating an emerging consensus on what metrics should be used to characterize web robot traffic. Examples of works that utilize other (unrelated to data-mining) methods of robot identification are (Wei-Zhou & Shun-Zhenga, 2006) (use Markov Chain modelling), (Ahn, Blum, Langford, & Hopper, 2003) (use Turing tests), and (Lin, Quan, & Wu, 2008) (use aggregate traffic analysis). The malicious crawler detection has been addressed in studies in (Lin, 2009; Hou, Chang, Chen, Laih, & Chen, 2010).

## 3. Supervised learning versus log parser

Many of the early systems for classification of web-site visitors were based on simple rule-based logic. Namely, for any given web-log file, a rule-based classification system would first perform text pre-processing in order to identify (i.e., extract) individual user sessions. Subsequently, by focusing on one or a few particular features, the system would derive a numerical (i.e. vector) representation of