Applied Soft Computing 13 (2013) 698–708

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

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

Detection of malicious and non-malicious website visitors using unsupervised neural network learning

Dusan Stevanovic*, Natalija Vlajic, Aijun An

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

ARTICLE INFO

Article history:
Received 8 August 2011
Received in revised form 24 April 2012
Accepted 6 August 2012
Available online 23 August 2012

Keywords:
Web crawler detection
Neural networks
Web server access logs
Machine learning
Clustering
Denial of service

ABSTRACT

Distributed denials of service (DDoS) attacks are recognized as 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 use of two unsupervised neural network (NN) learning algorithms for the purpose web-log analysis: the Self-Organizing Map (SOM) and Modified Adaptive Resonance Theory 2 (Modified ART2). In particular, through the use of SOM and modified ART2, our work aims to obtain a better insight into the types and distribution of visitors to a public web-site based on their browsing behavior, as well as to investigate the relative differences and/or similarities between malicious web crawlers and other non-malicious visitor groups. The results of our study show that, even though there is a pretty clear separation between malicious web-crawlers and other visitor groups, 52% of malicious crawlers exhibit very 'human-like' browsing behavior and as such pose a particular challenge for future web-site security systems. Also, we show that some of the feature values of malicious crawlers that exhibit very 'human-like' browsing behavior are not significantly different than the features values of human visitors. Additionally, we show that Google, MSN and Yahoo crawlers exhibit distinct crawling behavior.

1. Introduction

The today's business world is critically dependent on the availability of Internet. 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. In ever increasing numbers, these services are being offered by means of web-based applications. However, the inherent vulnerabilities of the Internet architecture provide opportunities for various attacks on the security of web-based applications. Distributed denial-of-service (DDoS) is an especially potent type of security attack, capable of severely degrading the response-rate and quality at which web-based services are offered. According to the United States' Department of Defence report from 2008, presented in Ref. [1], the number of cyber attacks (including the DDoS attacks) from individuals and countries, targeting economic, political and military organizations, are expected to increase in the future and cost billions of dollars.

The most common way of conducting a denial of service (DoS) attack is 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. In the past, most DoS attacks were single-sourced, which means they were reasonably easy to prevent by locating and disabling the source of the malicious traffic. Nowadays, however, almost all DoS attacks involve a complex, distributed network of attacking machines – comprising thousands to millions of hijacked zombies. These, the so-called DDoS attacks, are extremely difficult to detect due to the sheer number of hosts participating in the attack. At the same time, they can generate enormous amount of traffic toward the victim and result in substantial loss of service and revenue for businesses under the attack.

An emerging (and increasingly more prevalent) set of DDoS attacks, known as *application-layer* or *layer-7* attacks [2], are shown to be particularly challenging to detect. The reasons for this are: (1) in an application-layer attack, the attacker utilizes a legitimate-looking layer-7 network session, and (2) HTML requests sent to a web server are often conducted by a cleverly programmed crawler,[1]

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

1568-4946/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2012.08.028

[1] *Web-crawlers* are programs that traverse the Internet autonomously, starting from a *seed* list of web pages and then recursively visiting documents accessible from that list. Crawlers are also referred to as *robots* (*bots*), *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

in a way that mimics a semi-random walk through the web site links, and thus appears as a web site traversal conducted by a legitimate human user. Given the fact that application-layer DDoS attacks resemble the legitimate traffic, it is quite challenging not only to defend against these attacks but also to construct an effective metric for their detection.

So far, a number of studies on the topic of application-layer DDoS attacks have been reported. Thematically, these studies can be grouped into two main categories: (1) detection of application-layer DDoS attacks during a *flash crowd* event based on aggregate-traffic analysis [3,4] and (2) differentiation between well-behaved and malicious web crawlers based on web-log analysis [5–7].

The study presented in this paper falls in the latter of the above mentioned categories, as we examine the use of two unsupervised neural network (NN) learning algorithms for the purpose web-log analysis: the Self-Organizing Map (SOM) [8] and Modified Adaptive Resonance Theory 2 (Modified ART2)[2] [9]. In particular, through the use of SOM and ART2, our work aims to obtain a better insight into the types and distribution of visitors to a public web-site based on their browsing behavior, as well as to investigate the relative differences and/or similarities between malicious web crawlers and other non-malicious visitor groups.

In our earlier work [10], we have investigated the use of supervised algorithms (as provided by WEKA data-mining software [11]) for the purpose of web-user classification, including: C4.5, RIPPER, k-Nearest Neighbours, Naïve Bayesian Learning, Support Vector Machine. The results of this study have shown that supervised classification of web-users into different visitor categories (malicious vs. well-behaved vs. unknown visitors) can be effective and ensure satisfactory levels of classification accuracy, but only if preceded by a reliable data-labelling process. Namely, the main known disadvantage of most supervised algorithm, including those studied in [10], is the fact that they are only as good as their respective data-labelling strategy. Put another way, a supervised algorithm can provide accurate classification only if it has been trained on correctly labelled data, which in turn requires that the human/labelling expert be very familiar with the type and nature of the data-set being studied. Unfortunately, in the emerging era of highly sophisticated and ever-evolving web crawlers and bots (i.e., crawlers and bots that aim to hide or fake their identity by mimicking the behaviour of regular human visitors), the use of 'expert knowledge' for the purpose of reliable data pre-classification/labelling will be increasingly more problematic. Clearly, from the perspective or web-user classification, the presence of crawlers and bots with dynamically changing human-like behaviour is likely to translate into highly irregular and noisy data, and as such present a great challenge for any supervised expert-based system. This, ultimately, explains our motivation to extend the work presented in [10] and look at use of unsupervised learning for the purpose of web-user classification.

The content of this paper is organized as follows: in Section 2, we discuss previous works on web crawler detection. In Section 3, we give an overview of our web-log analyzer that is used to generate a meaningful training dataset out of any given access log file. In Section 4, we briefly outline our experimentation setup. In Section 5, we present and discuss the obtained web-session clustering results. In Section 6, we conclude the paper with final remarks.

2. Related work

To date, in addition to our work [10], several other research studies have also looked at the use of supervised learning for the purposes of data-mining and/or clustering of web sessions. Note that supervised learning process clusters sessions based on previous a priori knowledge. In one of the first such studies [12], the authors attempt to discover distinct groups of web robot sessions by applying C.4.5 algorithm (i.e., a decision tree classifier) to 25-dimensional feature vector space. The 25 features, i.e., their respective values, are derived from the navigational properties of each identified robot session. In advance of clustering, and depending on the value of *user-agent fields*, each session is pre-labeled as known *robots*, *known browsers*, *possible robots*, and *possible browsers*. The results of the study show that, by applying the proposed feature set in combination with C.4.5 algorithm, robots can be detected with more than 90% accuracy after only four web-page requests. In [13], the authors utilize supervised Bayesian classifier to detect the presence of web crawlers from web server logs and, subsequently, they compare their results to the results obtained with the decision tree technique. The proposed methodology achieves very high recall and precision values in web robot detection. Another study utilizing logistic regression and decision trees has been reported in [6]. In this study, authors propose a robot detection tool that speeds up the tasks for pre-processing web server access logs and achieves very accurate web robot detection.

Several studies have looked at the use of unsupervised learning for the purpose of more general web log analysis. In [14], the authors employ the Self-Organizing Map (SOM) algorithm to achieve automatic demographic-based classification of human web-site visitors based on the number and sequence of their web-page visits. In [15], the authors also examine the application of the SOM algorithm on web-server access logs, with the aim to group human web-visitors thematically and, as a result of that, help them find relevant information in a shorter period of time. In a similar study [16], the authors propose employing the Adaptive Resonance Theory (ART) algorithms to cluster human web users according to their thematic interests.

In the view of the previous works, the novelty of our research is twofold:

1) Firstly, to the best of our knowledge, this is the first study that applies unsupervised learning to the problem of web-visitor categorization, ultimately aiming to promote effective differentiation between malicious web-crawlers and other (non-malicious) visitor groups to a web site. (Note, in [14–16], only human web-visitors have been considered, and little to no attention has been given to automated web-crawlers.) We have chosen to use the SOM and ART neural network algorithms in our study for the following reasons.
   - The goal of the SOM algorithm is to cause the underlying neural network to respond similarly to similar input patterns. In order to achieve this goal, the network undergoes multiple rounds of the so-called *competitive learning process*. (In competitive learning, a training sample is fed to the network, and its Euclidean distance to all weight vectors is computed. The neuron with weight vector closest to the training sample is pronounced 'winner'. Once identified, the weight vector of the winner, as well as a number of its nearest topological neighbors, are adjusted towards the given training sample. The

crawl the web to compare prices and products sold by different e-commerce sites. Malicious crawlers are type of web robots that, for instance, generate DDoS traffic that can overwhelm web server's resources and thus limit or unable legitimate users' access to the website. Another example of malicious activity attributed to malicious crawlers is collecting email addresses for spam mail.

[2] Modified ART2 is a variation of the original ART algorithm [24]. Its advantages over the original algorithm are: (1) stable learning that results in gradually increasing/merging clusters, and (2) learning/clustering that can be terminated either when the radius of the formed clusters reaches some predetermined size, or when the number of formed clusters reaches some predetermined number.