Poster: A Web Server Identified Model based on Mean Shift

Junwei Su[1,2], Qixu Liu[1,2], Zhi Wang[1,2], Xiaoyun Li[3]
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences
Beijing University of Posts and Telecommunications
Beijing, China
liuqixu@iie.ac.cn

Abstract— **In view of the existing Web server identification method, there is a problem that the Web server that shields the Banner information has a low recognition rate and is strongly dependent on the fingerprint library. According to the difference of the server processing mechanism for malformed HTTP requests, a new identification method is proposed. This paper introduces the principle of the Web server identified model based on HTTP, Banner information and Mean Shift. We propose a method that how to send special HTTP requests to Web server, extract characteristics from HTTP response to identify Web server and use Mean Shift algorithm for Web component identification.**

I. Introduction

The Web server can serve contents to the World Wide Web, and store and provide documents to browser. The Web server identified model can analyze the of Web servers and judge Web servers' type and version. The accurate identification of Web servers is of great significance on assessing security of Web system, researching market share and forecasting the scope of influence on a network security incident.

HTTP is an application protocol for distributed, collaborative, and hypermedia information systems. The communication between browser and server takes place using HTTP [1]. The providers of the Web server must comply with RFC. However, the functions and the implementation methods of different Web servers are different, even though they comply with RFC standards. We can send specific requests and analysis response information to identify the type and the version of Web servers.

Mean Shift is a non-parametric feature-space analysis technique for locating the maxima of a density function, a so-called mode-seeking algorithm. Mean Shift is a procedure for locating the maxima—the modes—of a density function given discrete data sampled from that function [2].

In previous studies, Web servers that can be identified were only Apache, Nginx, IIS, and their common versions. However, Web server types in cyberspace are complex and numerous, and developers can block the traditional identified model that based on Banner by blocker or modified source code. And the researchers used supervised learning on the Web server identified model, which can identify the marked Web server [3]. But it can't identify the entire network space all types of Web server. There are so many types of Web server that we can't mark them artificially.

II. System Design and Implementation

Therefore, a Web server identified model based on HTTP technology, Banner technology and Mean Shift is proposed. This method solves the problem traditional Web server identification method can not identify Web server which blocking Banner information, and the problem of unlabeled Web server. The structure of our identified model showed in Figure 1.

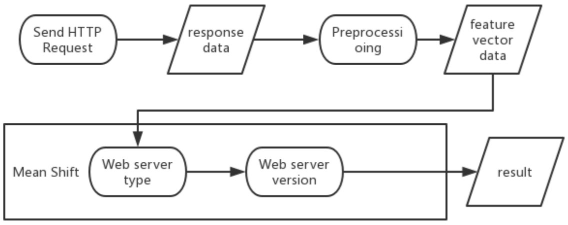

Fig. 1. The architecture of the identified system

The Web server identification model that based on HTTP technology, Banner technology and Mean Shift includes the following steps:

A. *specific HTTP request design:*

HTTP servers differ in the implement of the HTTP protocol. In the case where the HTTP request is well formed and legitimate, the response returned by all HTTP servers is more or less compliant with the specifications laid out in the RFCs for HTTP [4]. However, when confronted with malformed HTTP requests, these servers differ in their responses. Table1 shows the comparison of different responses.

TABLE1: WEB SERVER RESPONSE

Server	Field Ordering	DELETE Method	Improper protocol
Microsoft-IIS/7.5	Server, Date	405	405
Nginx	Server, Date	405	400
Apache/2.2.22	Date, Server	405	501

We design plenty of specific HTTP requests, which include normal GET, POST, OPTION, HEAD, DETELE request and malformed HTTP request. We construct deformed requests by modifying keywords, protocols, and versions. Table2 shows part of HTTP requests.

TABLE2: HTTP REQUESTS

HTTP Request	Description
GET/HTTP/1.0	Normal HTTP1.0 GET request
GET/HTTP/1.1	Normal HTTP1.1 GET request
PUT	Incomplete request
DELETE /HTTP/66.6	The wrong version number

B. Information collection and Preprocessiong:

The response packets returned by different Web Server have certain differences that reflected in the header composition of the response packet, response status code, response packet length, response packet content. We use the following methods to preprocess and vectorize data:

a) The response status codes are characterized by one-hot Encoding. If the Web server is not response, the NO_RESPONSE lable is 1.

b) We count the total number of header fields in the HTTP response packets and select the top 20 most frequently occurring header fields as feature 1 to feature 20, which shows in Table 3. If there is a header field in Table3, the corresponding position value is 1 and if it does not exist, the value is 0.

TABLE3: TOP 20 MOST FREQUENTLY OCCURRING HEADER FIELDS

id	key	id	key
1	content-Length	11	server
2	content-type	12	expires
3	connection	13	cache-control
4	date	14	last-modified
5	set-cookie	15	accept-ranges
6	mime-verion	16	etag
7	x-powered-by	17	transfer-encoding
8	age	18	content-location
9	www-authenticate	19	location
10	x-frame-options	20	x-cache

c) The order of the header fields is different. The order of Microsoft-IIS/7.5 is “Server,Date”,but the order of Apache/2.2.22 is “Date,Server”. We record the order of the top 10 occurrence header fields id(Table3 header fileld id) in the returned packet. If the header field is not in Table3, enter 0. The sample feature vector is S={11,2,4,14,15,3,1,14,19,0}.

C. Model training and result output:

We design two types of Mean Shift model. The function of the first model is to divide Web server base on type. The second one is to divide Web server base on version. We import the entire feature vector data into first Mean Shift model, and the export is categorized Web server. Then we import feature vector data of each categorized Web server into second Mean Shift model, and the export is different versions of Web server.

Marked samples: Not all of Web developer will mask Banner, we use Banner technology to identify the Web servers that are exposed. We also set up the common Web servers. These known Web servers are labeled samples.

Sample recognition accuracy: According to the marked samples, we design the following accuracy calculation method.

In the model, the maximum number of labeled samples in a cluster is X, while the rest of the labeled samples are denoted as Y; therefore the recognition accuracy of the arithmetic Ra(X) is defined as Formulate (1).

$$\mathrm{Ra}(X) = \mathrm{sum}(X)/\mathrm{sum}(X+Y) \qquad (1)$$

We use Mean Shift in the model, which is based on kernel density estimation.

First, the feature vectors of the scanned data are passed into the cluster of the first Mean Shift model. In order to get the maximum average value of the sum of all cluster Ra(X), we adjust the bandwidth value of the Mean Shift algorithm to get the optimal bandwidth value. When the bandwidth value is optimal, the web server type of each cluster is the most sampled Web server type in the cluster.

Second, each cluster in the first model will pass into the cluster of the second Mean Shift model, and we use the same as the first step to get the optimal bandwidth value. After the optimal bandwidth value is obtained, the Web server version of each cluster is the most sampled Web server version in the cluster.

III. CONCLUSION AND FUTRUE WORK

In this poster, we propose a Web server identified model base on HTTP technology, Banner technology and Mean Shift. In the future, we will continue to complete this experiment and select more representative features based on the experimental results to improve the accuracy of arithmetic.

IV. ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their helpful comments for improving this paper. This work is supported by the Key Laboratory of Network Assessment Technology, Chinese Academy of Sciences and Beijing Key Laboratory of Network security and Protection Technology and The National Key R&D Program of China（No. 2016YFB0801604）.

REFERENCES

[1] Fielding, Roy T.; Gettys, James; Mogul, Jeffrey C.; Nielsen, Henrik Frystyk; Masinter, Larry; Leach, Paul J.; Berners-Lee, Tim (June 1999). Hypertext Transfer Protocol – HTTP/1.1. IETF. doi:10.17487/RFC2616. RFC 2616.

[2] Comaniciu, Dorin; Peter Meer (May 2002). "Mean Shift: A Robust Approach Toward Feature Space Analysis". IEEE Transactions on Pattern Analysis and Machine Intelligence. IEEE. 24 (5): 603–619. doi:10.1109/34.1000236.

[3] Wu Shaohua, SunDan, Hu Yong. Web Server Identification Based on Bayesian theory[J]. Computer Engineering, 2015, 41(7): 190-193,198.

[4] HTTP/1.1 RFC 2616 HTTP://www.ietf.org/rfc/rfc2616.txt.