ISSN 2356-4407

www.STIKI.ac.id

Bersama memajukan ICT Indonesia

PROCEEDING

IC - ITECHS 2014

The 1st International Conference on Information Technology and Security

Malang, November 27, 2014

Published by:

Lembaga Penelitian dan Pengabdian pada Masyarakat

Sekolah Tinggi Informatika dan Komputer Indonesia

PROCEEDING

The 1st International Conference on Information Technology and Security (IC-ITechs)

November 27, 2014

Editors & Reviewers:

Tri Y. Evelina, SE, MM Daniel
Rudiaman, S.T, M.Kom Jozua
F. Palandi, M.Kom

Layout Editor:

Eka Widya Sari

LEMBAGA PENELITIAN & PENGABDIAN KEPADA MASYARAKAT
Sekolah Tinggi Informatika & Komputer Indonesia (STIKI) – Malang
Website: itechs.stiki.ac.id E-mail: itechs@stiki.ac.id

PROCEEDING

The 1st International Conference on Information Technology and Security (IC-ITechs) November 27, 2014

ISSN 2356 - 4407

viii + 276 hlm; 21 X 29,7 cm

Reviewers & Editors:

Tri Y. Evelina, SE, MM
Daniel Rudiaman, S.T, M.Kom
Jozua F. Palandi, M.Kom

Layout Editor:

Eka Widya Sari

Published by:
LEMBAGA PENELITIAN & PENGABDIAN KEPADA MASYARAKAT
Sekolah Tinggi Informatika & Komputer Indonesia (STIKI) – Malang
Jl. Raya Tidar 100 Malang 65146, Tel. +62-341 560823, Fax. +62-341 562525
Website: itechs.stiki.ac.id E-mail: itechs@stiki.ac.id

GREETINGS

Head of Committee IC-Itechs

For all delegation participants and invited guest, welcome to International Conference on Information Technology and Security (IC-Itechs) 2014 in Malang, Indonesia.

This conference is part of the framework of ICT development and security system that became one of the activities in STIKI and STTAR. this forum resulted in some references on the application of ICT. This activity is related to the movement of ICT development for Indonesia.

IC-Itechs aims to be a forum for communication between researchers, activists, system developers, industrial players and all communications ICT Indonesia and abroad.

The forum is expected to continue to be held continuously and periodically, so we hope this conference give real contribution and direct impact for ICT development.

Finally, we would like to say thanks for all participant and event organizer who involved in the held of the IC-Itechs 2014. We hope all participant and keynote speakers got benefit from this conference.

ISSN 2356 - 4407
The 1st International Conference
on Information Technology and Security

LIST OF CONTENT

BAUM-WELCH ALGORITHM IMPLEMENTATION FOR KNOWING DATA CHARACTERISTICS RELATED ATTACKS ON WEB SERVER LOG

Triawan Adi Cahyanto
Department of Informatics Engineering, University of MuhammadiyahJember
triawanac@unmuhjember.ac.id

Abstract

A web server log on the web server which contains all the activities of a web-based application. Activity stored in the log contains request and response from a web server to a web-based application that user accessed. This activity is an important data and can be used to search for occurrence of attack if there are problems on a web-based application. This paper uses a dataset of web server logs to analyze data using dataset which is taken from the event forensics contest because the web server log dataset has a record of activity form attack on a web-based application. Baum-Welch algorithm is used to search a web server log data so that the characteristics of the data are considered as attacks can be classified according to the level of success based on the records stored in the log. The results of the analysis based on dataset used calculated that the amount of data related to the attack were identified as many as 1.120 incidents of a total 202.145 records of data in web server log file.

Keywords: *Baum-Welch Algorithm, Data Characteristics, Log File*

1. INTRODUCTION

Currently, server and web-based applications are popular targets for attacker. In addition to ease of access, resource availability application content through the network (internet) make attacker have plenty of time to do the analysis and resource attacks against the target. It is a negative impact derived from the development of technology and computers. In this regard, there is a field of science and computer technology is relatively well developed at this point is digital forensics. The scientific field of digital forensics is used to conduct investigations related to high-tech crime or computer crime, so it can be used to search for digital evidence so as to entrap criminals [1].Logging mechanism in a web-based application is done by storing the data of each visitor who sends a request to the web server into a file, called a web server log. Visitor data contained in web server logs will be very useful if there is a problem that occurs on the web server, such as a web application attack (deface), DOS (denial of service), etc.

The data of the perpetrators will be known by checking one by one every record stored in the log. Data perpetrator is known one way is to look at the IP address used to access the web server. Records stored in the log must contain a record of all visitors who access the web server via a web-based application, it will be inefficient if the search data logs record by examining one by one of the many data stored in the log. Selection of data has its own mechanism, especially when you want to find data related to the attacks.Web server log data selection can be done by classifying objects into categories or classes based on a specific pattern (pattern recognition). The method can be used to perform data selection based on a pattern of an object, such as Neural Network (NN), Hidden Markov Model (HMM), K-Nearest Neighbor (KNN).

HMM cansolve the problem of NN and KNN whichassumes a system modeled as a Markov process withconditionsthatare not observed. Therefore, the possibilityof the transitionbetween the conditionofbeing the only parameter wasobserved, resulting in HMM state is not seendirectly, but the output is dependent on the visiblestate. HMM appropriatewhenapplied in the selectionof log data toclassify the data relatedto the attack, by looking at the order of the data contained in log records and detect the typeof attack attempts. HMM has the disadvantageofrequiring a long timeto test the stateof the sequence. That'sbecause HMM is verydependent on the statesequencemodelsarebuilt from a case. Increasingnumberof log data, there is a record in the log is not relatedto the attack. The record is called the false alarm. Alternatives areusedtoreduce the numberoffalse alarms caused by logs thatfiltering and attack patternrecognitionbased on the urlcontained in the log using the HMM. Thisstudyuses the Baum-Welch algorithmtodetermine the statisticalcharacteristicsof the data based on the log is alreadysaved. Baum-Welch algorithm is oneofseveralalgorithmsin hidden Markov models.

Thisstudy is expectedtobuild and develop the software easytouse by system administrators in analyzing the log data relatedto the attacks thatarespecificto web-basedapplications. The software is built, is expectedtoseek the data associatedwithtrials and attacks canvisually display the statisticsof all log data associatedwith the attempted attack.

2. RESEARCH METHOD

Research methodology is divided into two:

2.1 Analysis and Results

Part of the analysis and results will be illustrated in the following diagram:

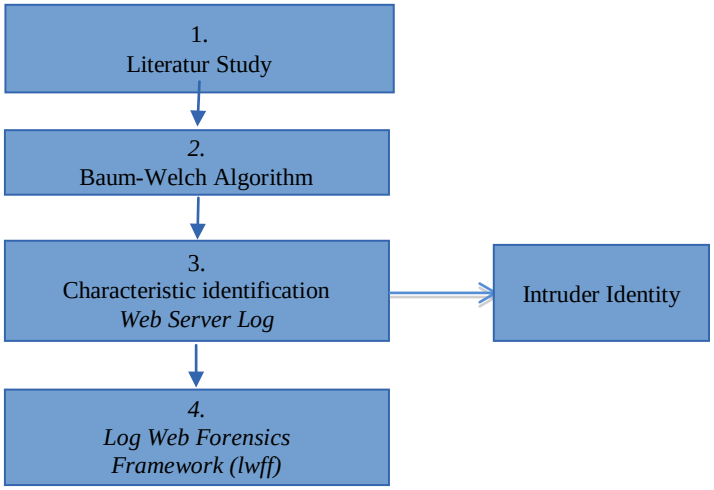

Figure 1: Analysis and Results Diagram

Log files have a tendency to change significantly. Therefore, the log data is stored on the server is the data that is important to do an analysis, especially when the server system is impaired. All activities contained on the system will be recorded on the server in the form of log files, and therefore very necessary to do further investigation about the log.

a. Damage Detection Data
A mechanism to check the web server logs the data to be analyzed. Web server log data were examined in order to know the "potential" modification of data by looking at the timestamp data contained in the metadata of the file log. Potential modifications to the data obtained from the difference between the current timestamp data with a timestamp data contained in log files metadata.

b. Identification of User Identity
The identity of users accessing the resource on the web server takes the data stored hostname parameters of web server logs.

2.2 Implementation and Testing

Part of the implementation and testing will be illustrated in the following diagram:

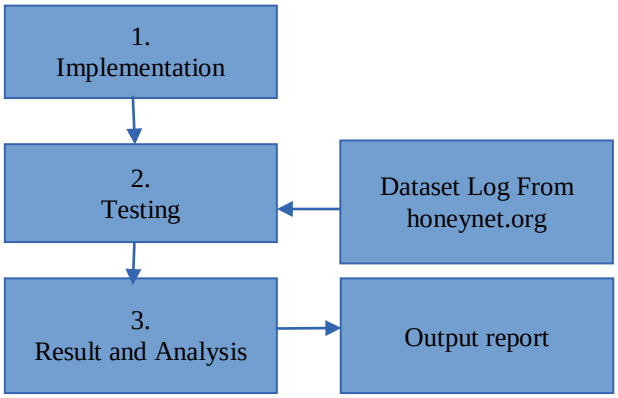

Figure 2 : Implementation and Testing

a. Implementation

Applications are made to be run on the terminal (in this study uses kali linux) and using the Apache web server log data, will then conduct an analysis based on the data log as desired based on certain parameters eg timestamp, hostname, referrals, status code, and others.

b. Testing Software

Software testing is done by simulation to determine how the software is running, how the software works, whether it is in accordance with the original purpose or not. Lwff software will be tested and the results of the analysis will be a comparison with the results of the analysis of the quality of other software. Testing lwff with other software is expected to know the effectiveness of the performance of the software.

c. Analysis of Results

Analysis of the results of the analysis results of the investigation of the web server log data. The output of the analysis of these results is in the form of a simple report with html format files.

3. RESULT AND DISCUSSION

Result was divided into two phases :

3.1 Preprocessing

Preprocessing stage is done to find the knowledge and gather data from log files. Before performing anomaly detection, make sure that the web server log data available, can be used to analyze input data.

After checking the data, then we can find two types of log data, namely:

- Basic log files
  Data is found directly on the web server logs.
  Examples of basic log file format:
  10.0.1.8 - - [12 / Dec / 2012: 11: 26: 24 +0200] "GET /my-webapp.php?id=1 HTTP / 1.1" 200 2769
- Supplemental log file data
  An additional data that can be generated from the data base. Told as supplemental data when added some more parameters of basic data is already stored log files.
  Example Supplemental log file data formats:
  10.0.1.8 - - [12 / Dec / 2012: 11: 26: 24 +0200] "GET /my-webapp.php?id=1 HTTP / 1.1" 200 2769 "http: //localhost/links.php" "Mozilla / 5.0 "

The log data will be reconstructed based on stored session and log data sets as clients who deserve to be identified. Reconstruction session contained in the log will result in the detection of activity, namely the automatic detection of the attack, the detection of non automated attacks.

- Automatic detection of attacks
  Detection is characterized by log records obtained from the software web application vulnerability scanner to test a website.
- Detection of attacks that are non-automatic
  Detection is characterized by log records obtained from the sequence of user activitywith the manual method.

This study uses the approach of time and the user agent string in the session identification process. Deadline set for the session analyzed were 60 minutes if the user agent string remains the same. If you want to change the time, quite a change in the variable $ max_session_duration contained in the program code. The algorithm used to identify the session are as follows [2] :

1. for all clients i do
2. for all requests j do
3. if delayj, j-1> 60min or (new agent and delayj, j-1> 60sec + μi + 3) then
4. the new session to true
5. end if
6. the end for
7. the end for

Pseudocode algorithm will perform repetitions of client data in this case is the log data found. Huang (2004) revealed that of the log data were then searched by the user id and the timestamp contained in the session so that the session can be analyzed from a client or user [3].

3.2 Training (Baum-Welch Algorithm)

Training phase is used as a first step to create and train a HMM Ensemble of each web application URL query parameter. Valid input sequence is made through the training data and then compare it with the data testing. The learning process and the number of inputs that may be available then all the data in the form of letters to be converted into a character and all the numbers will be converted to character N then all characters that exist on the web server logs will be stored.

Each client log data are only allowed to contribute once each generated training data to avoid changes in the dataset. The values were added before it was converted into a certain character is derived from the log data that has a response code 2xx / 3xx, to demonstrate that the log data is a normal operation.

For training HMM Ensemble, this study uses Baum-Welch algorithm [4]. Training data used take from the query parameters, argument names, url path, cookie, and the user agent of the preprocessing stage. Figure 4.1 below illustrates the training process so that the web server log data can be searched Ensemble HMM value of each parameter from the url query log data [5].

3.3 Software Visualization

Program code created just complement the features of the existing program. Programs created using PHP programming language based on the command line, so that the code can only be executed using the console terminal and the Apache web server. Software testing performed by accepting input log files from the log data honeynet.org since been identified containing data attacks. Testing using log data consisting of user history data that is not related to the attack so as to obtain the difference results from the analysis of the use of two different log data.

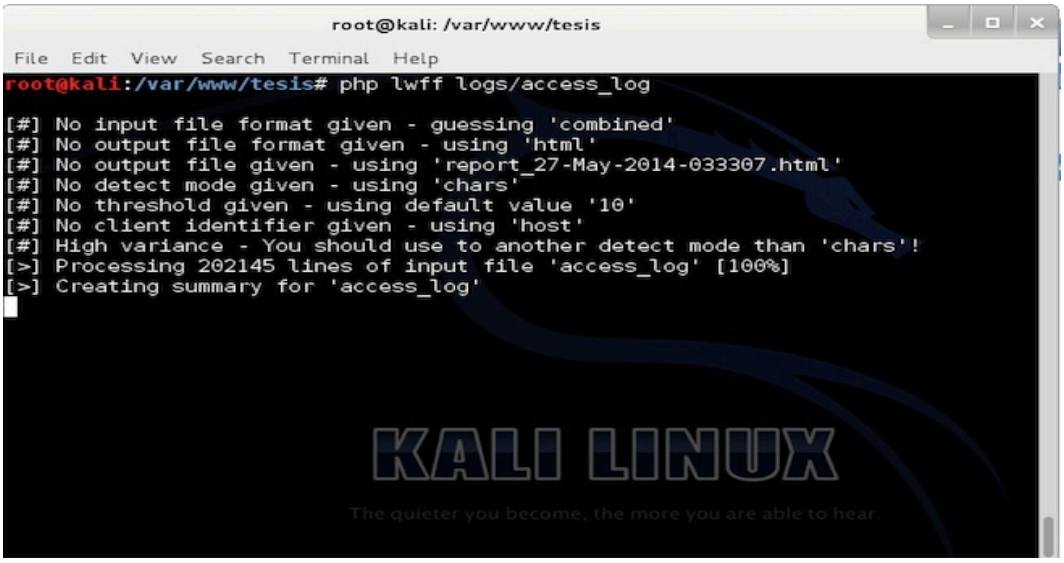

Figure 3 : Baum Welch Algorithm in Source Code Program

- To carry out the analysis of web server logs, simply type the location of the web server storage logs, ie "php lwff logs /access_log". For this case, the web server logs stored in the logs directory.
- The program will conduct an analysis of web server logs, if completed will generate its output in HTML format

The output will be generated after analysis by location or path where the web server storage logs, based on the data analyzed. There are drawbacks when generating the report analysis, ie when the web server log file size is quite large (about 10GB) then the data analysis process will take a very long time. The following is an example of the output in the form of the characteristics of the web server log data

Figure 4: Output Characteristics Web Server Log Data

Based on the characteristics of the image data can be seen with the data recorded detailed information such as the following picture

Date	Request	Final-Status
Sun, 22 Aug 2010 16:00:04 +0800	GET /elearning/index.php?cal_m=12&cal_y=1942 HTTP/1.1	200
Sun, 22 Aug 2010 17:33:12 +0800	GET /elearning/index.php?cal_m=11&cal_y=1942 HTTP/1.1	200
Sun, 22 Aug 2010 17:33:13 +0800	GET /elearning/calendar/view.php?view=month&course=1&cal_d=1&...	303
Sun, 22 Aug 2010 19:02:26 +0800	GET /elearning/index.php?cal_m=10&cal_y=1942 HTTP/1.1	200
Sun, 22 Aug 2010 19:02:28 +0800	GET /elearning/calendar/view.php?view=month&course=1&cal_d=1&...	303
Sun, 22 Aug 2010 20:30:56 +0800	GET /elearning/calendar/view.php?view=month&course=1&cal_d=1&...	303
Mon, 23 Aug 2010 11:00:06 +0800	GET /elearning/index.php?cal_m=12&cal_y=1941 HTTP/1.1	200

Figure 5 : Detail Data

4. CONCLUSION

Based on the explanationthat has beenwritten, it can be concluded as follows :

1. Characteristicsof web server log filescan be identifiedbased on the timedelayof log data in order todetermine the authenticityof log data is already in a state still in accordancewith the modified or original. To log datasetderived from honeynet.org, the resultsof the data containedtamperdetectionofpossiblechanges on March 14, 2004 16:50:20 is becausethere is a different timestamp in the log record is 3 hours 1 minute 44 seconds.
2. Hidden Markov ModelsMethod is a procedureto process log data relatedto the attack, in which the problem-solving process is done in threebasicstages, namely the evaluationof the problems with the forward-backwardalgorithm, read problems with the viterbialgorithm, and study the issueswithbaum-welchalgorithm.
3. Lwff software cananalyze log data quickly and produce a reportanalyzing the good and accompanied by statistical data in the form of a diagram chart.

BIOGRAPHY

The author is a lecturer at Departmentof Informatics Engineering, University of Muhammadiyah Jember. He graduated Master of Information Technology in the Islamic University of Indonesia to the field of digital forensic interest. Currently, he is trying to learn and write some papers, scientific journals related to information security, data mining and open source.

REFERENCE

[1] Al-Azhar, M.Nuh. 2012. Digital Forensik Panduan Praktis Investigasi Komputer. Penerbit Salemba : Jakarta.

[2] Muller, J. (2012). Web Application Forensics.*Ruhr-Universitat Bochum*.

[3] Huang, X., Peng, F., An, A., &Schuurmans, D. (2004). Dynamic Web Log Session Identification With Statistical Language Models, *55*(14), 1290–1303. doi:10.1002/asi.20084

[4] Frazzoli, E. (2010). Intro to Hidden Markov Models the Baum-Welch algorithm, 1–24

[5] Corona, I., Ariu, D., & Giacinto, G. (2009). HMM-Web □: a framework for the

[6] detection of attacks against Web applications, 1–6

[7] Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. IEEE.

Ross, S. M. (1983). Introduction to Stochastic Dynamic Programming (Probability and Mathematical Statistics)