Presented at the DLSU Research Congress 2017
De La Salle University, Manila, Philippines
June 20 to 22, 2017

Extracting Features from Web Logs for Web Usage Mining

Andrea Dominique Cortez[1], Paolo Gabriel Gamab [1,*], Joseph Emmanuelle Julian[1], Benjamin Gabriel Tan[1], and Jocelynn Cu[1]

[1] *Center for Networking and Information Security, De La Salle University*
**Paolo Gabriel Gamab: paolo_gamab@dlsu.edu.ph*

Abstract: Web usage mining (WUM) is a direct application of data mining techniques to discover patterns from web log data. These log files are vital in generating patterns since it contains copious amount of data gathered from the activity of the user in the web. Techniques such as data collection, pre-processing and determining a methodology to extract useful features for online user are used to generate these patterns. The formulated patterns that are obtained through the user's web activity can be beneficial to certain fields such as e-commerce wherein companies may use such information to distinguish the needs of its customers and improve their marketing strategies and forensics wherein possible crime suspects can be monitored which could help anticipate and prevent crime-related issues. The main focus of this paper is to show the results of the first half of WUM which involves data collection and feature extraction. The data collected was from the browsing activity of four users with the use of an extension installed in a Chrome browser on their personal computers. The collected data are web logs that were then turned into four datasets (one per person). Each dataset consists of 13 features. The data collected from the four users showed a total of seven hours worth of browsing activity, generating 31,469 instances.

Key Words: web usage mining; web log; data collection; feature extraction; browsing activity

1. INTRODUCTION

Web browsing is any type of online activity done in a Chrome browser. Whether it be going to a web page, reading a PDF online, or streaming videos, as long as it was done in the Chrome browser, it can be detected by the Log Catcher. Web Usage Mining is the application of data mining algorithms in order to discover user patterns from web data. The purpose of WUM is "to capture, model, and analyze the behavioral patterns and the profiles of users interacting with web sites" [6]. The patterns that are obtained are usually represented as objects, resources or pages which are frequently accessed by users with common interests. The important stages of Web Usage Mining (WUM) are the gathering the

data of user sessions using heuristics techniques and discovering patterns by using pattern discovery techniques. The whole procedure of using WUM is summarized into three steps, namely: (1) data collection and pre-processing; (2) pattern mining; and, (3) knowledge application [5]. The first step is further divided into three more steps: Collection, sorting and processing, and feature extraction. Data collection is the first step of WUM [7]. Web servers are the optimal place to collect web usage data since copious amounts of information can be found in their log files. Log files are simple text files in which information is jotted down each time a user accesses a resource from a website, these can be found in web servers, web proxy servers, and client browsers. Data Preprocessing is the technique where the information contained by the web log data file is cleaned and processed to obtain a quality data that can be used in pattern mining. It has been stated in numerous studies that raw web log data files must be preprocessed first before directly using it for pattern mining. That is to say, raw web log data files are inconsistent due to many undesirable information, information that are diverse and noisy [4, 7]. Feature extraction involves identifying ideal features from a dataset that may be used for activities like modelling. This step includes the following processes: data cleaning/reduction, which aims to remove unwanted data inside raw log files [7]. After this, user and session identification can be done in order to extract more information about the user and his browsing activity.

The main focus of this paper is the data collection and the feature extraction processes of WUM. The data that was used were web logs collected from the participants of the experiment. The researchers' motivation for doing this study is to assert the feasibility of web usage mining and that it is useful in various fields. Possible models that can be generated may be used to improve e-commerce by monitoring and tailoring the advertisements to the customer's needs and wants. It can also be used in crime prevention wherein authorities may investigate the web activity of possible crime suspects, gather web evidence, and anticipate different types of web attacks. It can also monitor conversations that may seem suspicious. Section II focuses on the methodology that was used to fulfil the focus of the paper. It discusses the main setup and the program that was used. Section III shows the results gathered from the two processes. A brief discussion of the results is also done. Lastly, Section V discusses the summary of the entire paper, as well as on-going work being done by the group in order to further the exploration on WUM. For the main text (contents) and references, two column formatting is required.

2. METHODOLOGY

The methodology for this study is focused on data collection and feature extraction. The physical setup for data collection starts with the user browsing the Internet with the Log Catcher Chrome extension. The user can browse various web sites for a minimum session duration of one hour. The user can browse the Internet at any time of the day. The Log Catcher, which is a Javascript tool that collects HTTP headers, is installed on the computers of the users. The program is enabled to run in the background while the user is browsing the Internet. The web logs of the users are collected and the features are extracted. The figure of this setup is shown in Figure 1.

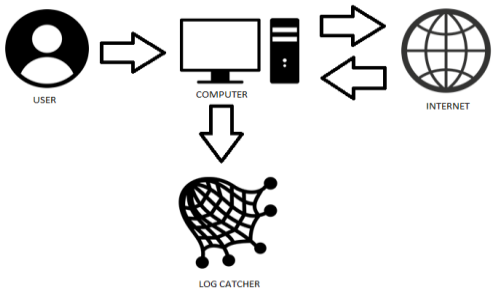

Fig. 1. Physical Setup for Data Collection

The Log Catcher is a Chrome extension we developed and is created in JavaScript programming language that catches HTTP headers when it is enabled. The Log Catcher is the tool that aids the researchers in acquiring the features that the study

needs. The general overview of how the Log Catcher works is that it gets instances of the HTTP headers, specifically request headers, when a user browses the Internet. The program logs these headers in the Chrome browser console. The program iterates through the headers until the specific features are found. Once the feature has been found, it logs the name of the feature and its value. This process is done in the background and its flow can be seen in Figure 2.

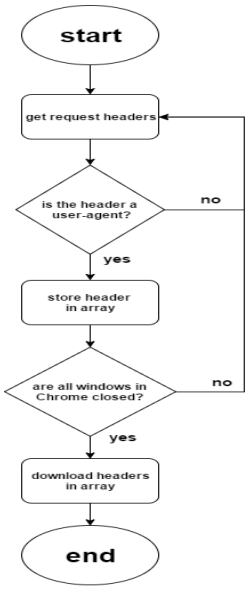

Fig. 2. Log Catcher Flowchart

The process continues until the user decides to close the Chrome browser. While it is true that these features are readily available in most forms of web logs, another method is to consider a fixed set of features. This may be done so that the processing time in the experiment becomes more efficient. Should this method be done, it has to be made sure that certain parameters are made to allow proper identification of different users or different sessions that may come from similar users or similar devices [2]. That being said, some of the features this paper focused on can be seen in Table 1 below. These features were indicated since these are the set of features that the log catcher program is able to collect.

Table 1. List of Features Considered for Feature Extraction

Feature	Description	Relevance
Request Method	Message communicated to a web server once a client has initiated a URL Request. Usually either a GET or a POST	May indicate if user is either downloading or uploading data
Accept Headers	Advertises which content is understood by the user. Determines the content that may be received from the server	May indicate what the user can actually see on the visited webpage
Tab ID	Identifier for specific tabs opened in a browser	May indicate behavior of a user by means of opening or changing tabs
URL Origin	Identifier where the URL Request may begin. It does not appear all the time.	May indicate the webpage where user came from when the intent to visit another website was made
URL Requested	Message sent by user/client to a server	May indicate the webpage the user would like to visit
URL Referred	Identifier where URL request came from	May indicate the webpage where user came from when the intent to visit another website was made
Content Type	Refers to the type of media found in a webpage	May indicate the type of content the user likes or needs from the webpage
Day-of - Week	Refers to the day the user browses a webpage	May indicate when the user decides to browse a particular webpage
Time-of-Day	Refers to the time when the user visits a webpage on a particular day	May indicate when the user decides to browse a particular webpage
User Agent	Identifier for the operating system and	May indicate the source of the

	the browser used	requests made by the user
Request ID	Identifier for specific requests in a web log	Helps sort out different requests in the dataset
Frame ID	Identifier for specific frames contained in a web log	Helps sort out different web log entries in the dataset

Once the data has been collected, it will undergo data cleaning which means that features not needed in the study are not included in the dataset of a user. The collected logs are saved in an Excel workbook. In cleaning the data, the researchers use the filtering feature of Excel in order to sort through the features and retrieve only the specific values. This allows the researchers to build the dataset of a certain user with the features as the headers and its corresponding values. The dataset is manually observed and analyzed by the researchers to see patterns like the amount of time spent browsing, number of sites visited and progression of sites visited in the online activity of a user. These patterns are recorded and thoroughly studied in order to come up with a generalization for how a certain user browses the Internet. The conclusions made from the different users are analyzed in order to note possible similarities and differences that can be taken from the users.

3. RESULTS AND DISCUSSION

The researchers have collected data from four users ages 20-25 during three specific time periods: Morning (12:00MN to 11:59AM), Afternoon (12:00NN to 5:59PM), and Evening (6:00 to 11:59PM). Three of whom are male and one female. Some of the results found were indicators of behavior from certain features collected from the dataset. Figure 3 shows a sample web log. Each dataset was cleaned and sorted using Microsoft Excel in order for the features to be put in separate columns.

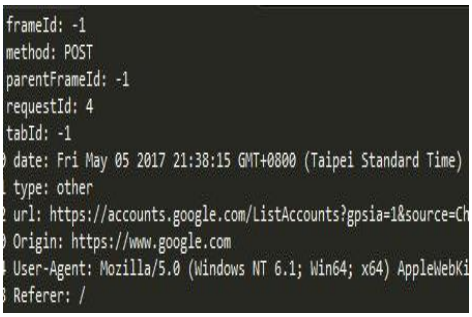

Fig. 3. Sample Web Log

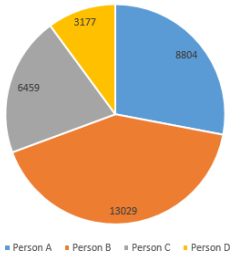

Fig. 4. Instances of Web Browsing per User

Figure 4 shows the number of instances or activity each user has accumulated with a minimum duration of an hour. The duration, shown in minutes, have also been collected. As seen in Figure 5, Person A spent the most time on the internet with over 261 minutes and 8804 instances. Person B however, who only spent 193 minutes on the internet produced much more instances than Person A with over 13, 029 instances. The same goes for Person C and D who only spent approximately 60-70 minutes each but one produced double instances than the other. This may indicate that participants who browse the Internet for a longer duration may not necessarily yield more instances which could imply that each user may have a unique browsing behavior.

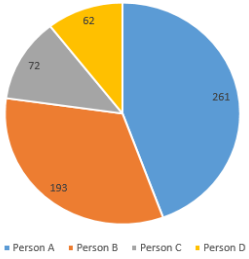

Fig. 5. Duration of Browsing Activity per User in Minutes

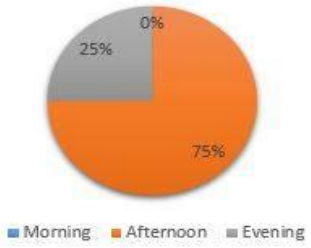

Fig. 6. Percentage of Activity During a Specific Time Period

Manual observations of the data led the researchers to list what the users usually do on the internet which include browsing through articles, chatting, watching a video or checking one's email. Most of the users tend to start browsing in the afternoon based on the timestamps seen in the datasets. This is shown in figure 6 where 75% or 23,602 instances were recorded in the afternoon while the remaining 25% or 7,867 instances were recorded in the evening. Another observation was the top sites that were visited by the users which include Facebook, Twitter, YouTube, 9gag and Google Mail. Upon further analysis of the data, patterns were discovered with regards to the browsing behavior of the users. Some of these patterns include all users starting their browsing session with opening a social media website, Facebook being the best example for this. Another pattern discovered is that while Facebook is one of the sites that showed the most instances, the users would often open another tab on the browser in order to navigate to another social media site like Twitter or YouTube. Another observation was that users tend to stay idle on a web page for up to 2 to 3 hours. This can be inferred as the user leaving the computer while doing other activities outside of browsing the Internet. Some productivity in the users was also found as some opened multiple tabs to view articles but immediately closed them after a few minutes.

4. CONCLUSION

Generally speaking, the current focus of this paper tackled on the data collection and feature extraction methodologies involved in Web Usage Mining. The data was collected from four participants and it produced four unique datasets. Each dataset underwent data pre-processing that included cleaning and feature extraction. In the process of feature extraction, we were able to acquire 12 distinct features. At present time, we manually observed possible indicators or patterns of user behavior using these features. Patterns that were first considered included difference in the session duration per user, number of instances of browsing activity, number of unique tabs open, as well as top webpages visited, and the first webpage that a user often visits upon opening the Chrome browser. Although these observations seem apparent of user behavior, there is still a need to apply proper pattern discovery techniques and algorithms such as the Apriori algorithm for Association Rule Mining in order to automate and streamline the process of generating web usage patterns for various users.

In lieu of the study, the researchers' ongoing work includes engaging more participants in collecting more data from their browsing activity. The target number of users for the data collection is 10. Also, given the processed data that has been generated, the researchers look to model the information using the Apriori algorithm for Association Rule Mining in order to generate patterns of web behavior per user to further support the manual observations made in this part of the experiment. After which, validation of the collected data and the generated models will happen. Lastly, thorough and detailed analysis of the results will be concluded.

5. ACKNOWLEDGMENTS

First and foremost, we would like to praise and give thanks to the Almighty Father for blessing us in all of our research endeavors.

We would like to thank Mr. Arvin Corpuz for sharing his valuable knowledge in web usage mining

and for mentoring us during the development of the Log Catcher tool.

We would also like to acknowledge the assistance of Dr. Judith Azcarraga in teaching us the possible modeling techniques that can be done moving forward with this research.

We are also grateful for the confidence that Dr. Merlin Suarez and Dr. Ron Resurreccion have given us to pursue this research.

Last but not the least, we would like to give our deepest gratitude to our thesis adviser and co-author, Ms. Jocelynn Cu, for inspiring us to delve into the field of web usage mining and for devoting her time, effort, and guidance throughout the entire research. This would not be possible without you.

6. REFERENCES

Abramson, M., & Aha, D. W. (2013). User authentication from web browsing behavior. Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference, St. Pete Beach, Florida. Palo Alto, California: The AAAI Press. Available: www.aaai.org/ocs/index.php/FLAIRS/FLAIRS13/paper/viewFile/5865/6081

Chitraa, V., & Thanamani, A. S. (2012). An enhanced clustering technique for web usage mining. International Journal of Engineering Research & Technology (IJERT), 4. Retrieved from www.citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.680.4122&rep=rep1&type=pdf

Jafari et al. (2013). Extracting users' navigational behavior from web log data: a survey. Journal of Computer Sciences and Applications (JCSA), 1. Retrieved from www.pubs.sciepub.com/jcsa/1/3/3/

Lakshmi et al. (2013). An overview of preprocessing on web log data for web usage analysis. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2. Retrieved from www.jatit.org/volumes/Vol34No2/11Vol34No2.pdf

Losarwar, V., & Joshi, M. (2012). Data preprocessing in web usage mining. Paper presented at International Conference on Artificial Intelligence and Embedded Systems. Retrieved from psrcentre.org/images/extraimages/47%20712010.pdf

Mishra, R., & Choubey, A. (2012). Comparative analysis of apriori algorithm and frequent pattern algorithm for frequent pattern mining in web log data. (IJCSIT). Retrieved from ijcsit.com/docs/Volume%203/vol3Issue4/ijcsit2012030420.pdf

International Journal of Computer Science and Information Technologies, 3. Retrieved from www.ijcsit.com/docs/Volume%203/vol3Issue4/ijcsit2012030420.pdf

Singh A. P., & Jain R. C. (2014). A survey on different phases of web usage mining for anomaly user behavior investigation. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 3. Retrieved from www.ijettcs.org/Volume3Issue3/IJETTCS-2014-06-03-066.pdf

Sisodia, D. S., & Verma, S. (2012). Web usage pattern analysis through web logs: a review. Paper presented at 2012 Ninth International Joint Conference on Computer Science and Software Engineering (JCSSE), Bangkok, Thailand. Retrieved from http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6261924

Sundari et al. (2014). A review on pattern discovery techniques of web usage mining. International Journal of Engineering Research and Applications (IJERA), 4. Retrieved from www.ijera.com/papers/Vol4_issue9/Version%204/S4904131136.pdf