ISSN: 2454-1907
DOI: 10.5281/zenodo.1006814

IJETMR
International Journal of Engineering Technologies and Management Research
A Knowledge Repository

UNDERSTANDING WEB TRAFFIC ACTIVITIES USING WEB MINING TECHNIQUES

Ng Qi Yau[1], Wan Mohd Nazmee Wan Zainon[*2]
[1, *2] School of Computer Sciences, UniversitiSains Malaysia, 11800, Penang, Malaysia

Abstract:
Web Usage Mining is a computational process of discovering patterns in large data sets involving methods using the artificial intelligence, machine learning, statistical analysis and database systems with the goal to extract valuable information from accessing server logs of World Wide Web data repositories and transform it into an understandable structure for further understanding and use. Main focus of this paper will be centered on exploring methods that expedites the log mining process and present the result of log mining process through data visualization and compare data-mining algorithms. For the comparison between classification techniques, precision, recall and ROC area are the correct measures that are used to compare algorithms. Based on this study it shows that Naïve Bayes and Bayes Network are proven to be the best algorithms for that.

Keywords: *Web Usage Mining; Data Mining Algorithms; Mining Techniques and Pattern Discovery.*

Cite This Article: Ng Qi Yau, and Wan Mohd Nazmee Wan Zainon. (2017). "UNDERSTANDING WEB TRAFFIC ACTIVITIES USING WEB MINING TECHNIQUES." *International Journal of Engineering Technologies and Management Research*, 4(9), 18-26. DOI: 10.5281/zenodo.1006814.

1. Introduction

The process of extracting useful patterns in log files is called log mining. Log mining itself is just like any web mining processes. Web mining can be divided into three main categories, Content Mining, Structure Mining, and Usage Mining. In this work, we mainly focus on Web Usage Mining (WUM) that has been defined as "application applying data mining techniques to discover usage patterns from WWW data" [1]. Web usage mining can allow organizations or enterprises to unwrap users' patterns in website browsing or present users' potential preferred products or pages. The success of data mining applications just like as other applications is depended on development of standard.

CRISP-DM (Standard Cross-Industry Process for Data Mining) is a conglomeration of companies that has defined and validated a data mining process that can be used into different data mining projects [2]. There are six stages of CRISP-DM: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. Business Understanding focuses on understanding the project objectives and requirements from

ISSN: 2454-1907
DOI: 10.5281/zenodo.1006814

perspective view point of business then converting knowledge into a data mining problem definition and preliminary plan is designed to achieve objectives. Data Understanding starts with data collection and proceeds to activities to identify data quality problems and also detect interesting subsets from hidden information. Data Preparation which covers all activities to construct final dataset from initial raw data usually will be performed several times. In modelling stage, multiple modelling techniques are applied. For same data mining problem, several techniques can be used. Evaluation stage is to determine if there are some importance business issues that have not been considered. Deployment stage involves activity which presents organized knowledge to customers or users in a way they can understand. Log mining is basically similar with web usage mining.

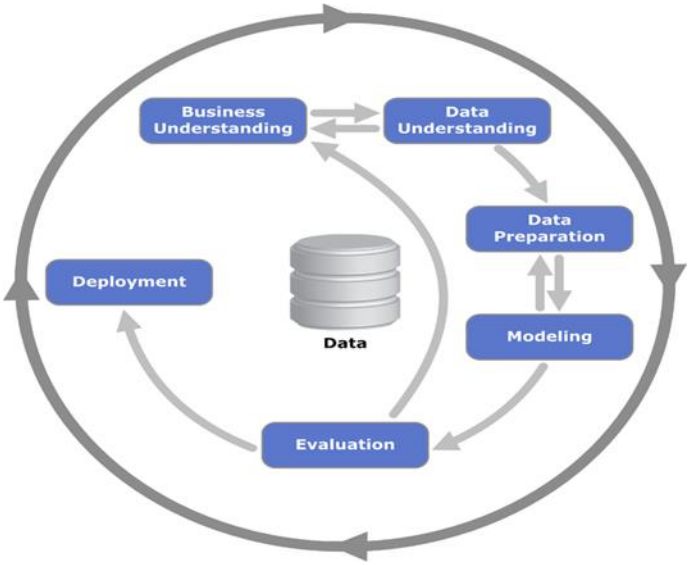

Figure 1: CRISP-DM model

In this paper, WEKA will be the tool to implement data mining algorithms for exploration purpose. WEKA (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. WEKA is free software available under the GNU General Public License. WEKA provides several very useful functions from pre-processing to visualize the patterns. WEKA's classification panel provides algorithms that will be applied to dataset and to be compared. In the Introduction section, present clearly and briefly the problem investigated, with relevant references.

Data Mining and Web Usage Mining

Application of data mining techniques into World Wide Web (WWW) is also known as web mining. Web mining can be divided into three main categories, Content Mining, Structure Mining, and Usage Mining. In this work, we mainly focus on Web Usage Mining (WUM) that has been defined as "application applying data mining techniques to discover usage patterns from WWW data" [1]. Content mining is referred as application of data mining technique to the contents of websites. A conceptual schema is created to define semantics of large volume unstructured datasets. Unstructured data is also known as no pre-defined data model, usually text but also can be images or video files. Structure mining identifies the relationship between Web

ISSN: 2454-1907
DOI: 10.5281/zenodo.1006814

pages linked by information or direct link connection. This could be done by using spider crawler to pull out data from the links.

Usage mining can be described as the discovery and analysis of user access patterns, through the mining of log files and associated data from a particular Web site. Through web usage mining, we can uncover these underlying patterns such as "what pages are the most popular and the least popular", "What section did users browse the most?" and "Which geographic area did most of the users come from?"

Data Preprocessing and Log Mining

As shown in Fig. 1, the data mining model starts from business understanding, which corresponding with objectives in this thesis. Only fully understand the objectives of carrying out data mining tasks we can perform the next step: data preprocessing. To carry out data preprocessing, firstly we need to understand what our data is. In this paper, our data will be log files extracted from storage of servers. Several log file formats can be analyzed by Web Site Analyzer which are:

a) NCSA (Common or Access, Combined, and Separate or 3-Log)
b) W3C Extended (used by Microsoft IIS 4.0 and 5.0)
c) SunTM ONE Web Server (iPlanet)
d) IBM Tivoli Access Manager WebSEAL
e) WebSphere Application Server Logs
f) FTP Logs
g) Custom Log File Format (field information defined by user)

The web access log in CLF format has information of the IP address of a visitor's machine, the user's ID of visitor if available and date/time of the page request. The method is a means of page request. It can be GET, PUT, POST or HEAD. The URL is the page that is requested. The protocol is the means of communication used, HTTP/1.0 for example. The status is the completion code. For example, 200 is the code for success. The size of field shows the bytes transferred as a result of a page request. The Extended Log Format, in addition to information, stores referrer, which is the page this request has come from and agent is the web browser used. During the data preprocessing process, Lelani [3] has concluded that the complexity of web log preprocessing has made it the most difficult and the most time consuming process compared with other data mining processes. It can be divided into four major steps:

1) **The Removal Of Web Robot Accesses**
   The web robot accesses included indexes, web crawlers and so on which is usually "robots.txt" created by Web administrator for access purpose. The web server also records accesses to file /proxy.pac" which is an auto-configuration file to configure all web browser clients. Typically, these system-generated entries are needed to be filtered out to avoid their effect on discovered patterns.
2) **Filtering Of Images And Noisy Data**
   Usually website contains images or even video in the page, but the real purpose is not at images or videos that embedded in the page. They're just for enhancement purposes. So filtering will be done. As we can see from image 2.1, when a webpage is loaded, all the

 ISSN: 2454-1907
DOI: 10.5281/zenodo.1006814

contents with images, etc will be loaded at the same time. During data preprocessing, they must be removed for real users' purposes.

3) **Extracting Transactions**
   According to Margaret Rouse, a professional journalist in WhatIs.com, in computer programming, a transaction usually means a sequence of information exchange and related work (such as database updating) that is treated as a unit for the purposes of satisfying a request and for ensuring database integrity. The problems discussed by Pitkow [4] are the maximum time limit session given by proxy servers, the hyperlink structure and multiple IP address used by a single user to access the same website. Besides that, a time session that can fix a certain time limit, such as 30 minutes can be used to determine a single transaction. In order to focus at users' navigational purposes, reference's length can be used.
4) **Features Selection And Formatting**
   Features selection is a process where only desired attributes and remove the irrelevant attributes are chosen and reformatting those attributes into a file format where data mining software can understand. Feature selection can improve model interpretability and reduce overfitting. Feature selection is to be distinguished from feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Usually these types of file formats are acceptable by data mining tools: csv, arff, matlab file and so on.

Data Mining Algorithms

In order to dig out hidden patterns of every structured, semi-structured and structured data, various data mining algorithms are developed. Data mining algorithms are not optimized but heuristics because every algorithm has its limitations. For example, to describe how the relation of each dataset is, clustering can be used but how to predict outcomes with different criteria, a decision tree algorithm can be used.

Choosing the right algorithm can be the toughest yet most critical decision to make as for same business task, each algorithm produce a different result, some even produce more than one result. However there is no need to limit one algorithm for each problem as in order to produce different views on datasets. Regression can be used to predict financial forecasts while neural network algorithm can be used to perform analysis of factors affecting economics. Teressa [5] has concluded several possible factors to be considered when choosing suitable data mining algorithms.

a) The limitation of tools available for data mining and data mining algorithms which can be performed by these tools.
b) Main goals of the problems and structure of datasets
c) Adapt of more than one algorithms in order to get desired results
d) The data mining algorithms must be fast and clean-documented
e) Data miners must understand the algorithm and if possible perform test with other input files before implementing it on the actual data set.

In choosing methods to analyze data, Berson et.al. [6] Has stated statistics and clustering could be very useful in data mining. By strict, statistics is not considered as a branch of data mining but

ISSN: 2454-1907
DOI: 10.5281/zenodo.1006814

it has been proved very useful in real world environment. Neural networks and nearest neighbours technique tends to be more robust and not user-friendly for those who are not in expert in data mining, the industry jargons might suffocate the users and spend a lot of time to perform mining.

2. Research Methodology

Fig. 2 shows the basic framework of this work. It is modified from CRISP-DM model which has been shown in Fig. 1. A big difference apart with CRISP-DM model is in this paper, comparison of classification and clustering techniques will be added as one of the objectives in this research is to determine the efficiency of data mining algorithms. Meanwhile, business understanding is very importance too. The whole process will be constantly revised to make sure that the progress is on the track which are discovering users' pattern and collect measurable results to enhance website's features.

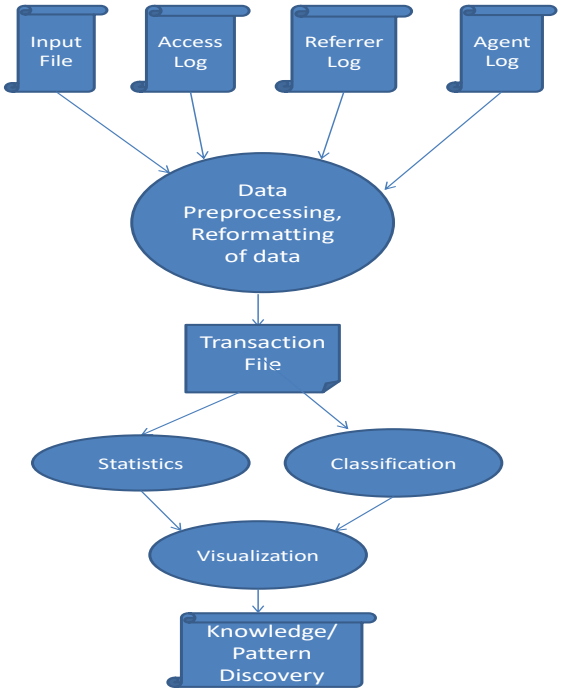

Figure 2: The framework for the proposed log mining process.

The beginning starts with data understanding process. Web log files consist many attributes therefore at least 10 of data mining thesis are reviewed. Experiment stars after understanding data mining models, data mining algorithms and related concepts. 10 days of School of Computer Science servers' log files which begins from 1st December to 10h December 2014 will be collected. The preprocessing is similar with data preparation in CRISP-DM.

In this paper, the log file format is NCSA Combined log file format with little customization of removal of cookie. The typical format of NCSA combined log file is looked like [7]:

Host, rfc931, username, date: time, request, status code, bytes, referrer, user agent, cookie

ISSN: 2454-1907
DOI: 10.5281/zenodo.1006814

For example: *125.125.125.125 - dsmith [10/Oct/1999:21:15:05 +0500] "GET /index.html HTTP/1.0" 200 1043 "http://www.ibm.com/" "Mozilla/4.05 [en] (WinNT; I)" "USERID=CustomerA; IMPID=01234"*.For business understanding purpose, only request and host (IP address) will not be filtered.

A transaction is determined by IP address. Pitkow [4] assumed that if IP address is same then same user is on the transaction, Cooley et. al. [8] has suggested that 30 minutes as a time limit to differentiate different users however Spilipoulouet. al. [9] has suggested that 24 hours because people at age of moderns usually perform multitasking, so they usually open multiple tabs and continue their tasks perhaps hours later. In this paper, suggestion of Spilipoulouet. al. will be taken. Multiple duplicated transactions caused by downloads will be removed as time interval in between two transactions is less than 1 second, only opening and ending transactions are kept.

This experiment will be using WEKA as data mining tool. WEKA's main user interface is the Explorer, but essentially the same functionality can be accessed through the component-based Knowledge Flow interface and from the command line. There is also the Experimenter, which allows the systematic comparison of the predictive performance of WEKA's machine learning algorithms on a collection of datasets [10]. In this paper, the knowledge discovery process will be focused at the Explorer panel.

In Explorer interface, several panels are created for knowledge discovery process. However in this paper, only preprocessing and classification panel will be used as knowledge discovery process due to time limitation.

Preprocessing

Raw data from databases needed to be preprocessed before raw data is imported into WEKA. Cleaning and reformatting of raw data refers to transform of unstructured raw data into a structured dataset which is understandable by WEKA. The common formats that WEKA can understand are arff, csv, json, matlab and so on. Filtering is a powerful function that can be used to perform supervised or unsupervised filtering of attribute or instance. In this paper, preprocessing is done by using Microsoft Excel 2007 before file is being loaded into WEKA. Microsoft Excel is able to load files from text and convert them into csv or arff file. The original log files from School of Computer Science is opened in text file form and loaded into Excel for cleaning and reformatting. Preprocessing panel is only involved in loading csv files into WEKA in this paper as Microsoft Excel can perform preprocessing in a faster way.

Classification

In this panel, instances can be classified 6 main categories of data mining techniques. In this thesis, only classification will be used due to time constraint and well documentation of classification techniques in data mining or more specifically text mining so that comparison result is based on previous works and is useful. Classification also performs well under supervised learning which means we predefined rules for instances to be grouped. Clustering techniques mentioned in this thesis have been tested separately on different datasets and yet recorded good result on accuracy so they are used to be compared with each other's in order to get useful comparison results. Naïve Bayes has been one of the popular machine learning methods for many years. Its simplicity makes the framework attractive in various tasks and

ISSN: 2454-1907
DOI: 10.5281/zenodo.1006814

reasonable performances are obtained in the tasks although this learning is based on an unrealistic independence assumption.

Data Mining Algorithms' Comparison

5 classification algorithms are implemented in WEKA to get result for comparison in term of performance. During the experiment of comparison, 10 folds of cross-validation is taken. In 10-fold cross-validation, the original sample is randomly partitioned into 10 equal sized subsamples. Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 9 subsamples are used as training data. The cross-validation process is then repeated 10 times (the folds), with each of the 10 subsamples used exactly once as the validation data. The 10 results from the folds can then be averaged to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once.

The data used to construct or discover a predictive relationship are called the training data set. Most approaches that search through training data for empirical relationships tend to overfit the data, meaning that they can identify apparent relationships in the training data that do not hold in general. A test set is a set of data that is independent of the training data, but that follows the same probability distribution as the training data. Again, top 20 most frequency terms will be used for classification purpose to test whether each of the 5 classification techniques can successfully classify them correctly or not. Among 5 techniques, 0R algorithm will be used as baseline benchmark of performance.

3. Results And Discussions

The performance of 0R trivial model is below par and it cannot be used for discovering "active" compounds. However, the accuracy of the model (Correctly Classifieds Instances) of this trivial model is quite high: 83.8733 %. This fact clearly indicates that the accuracy cannot be used for assessing the usefulness of classification models built using unbalanced datasets. For this purpose, precision, recall and ROC area will be used.

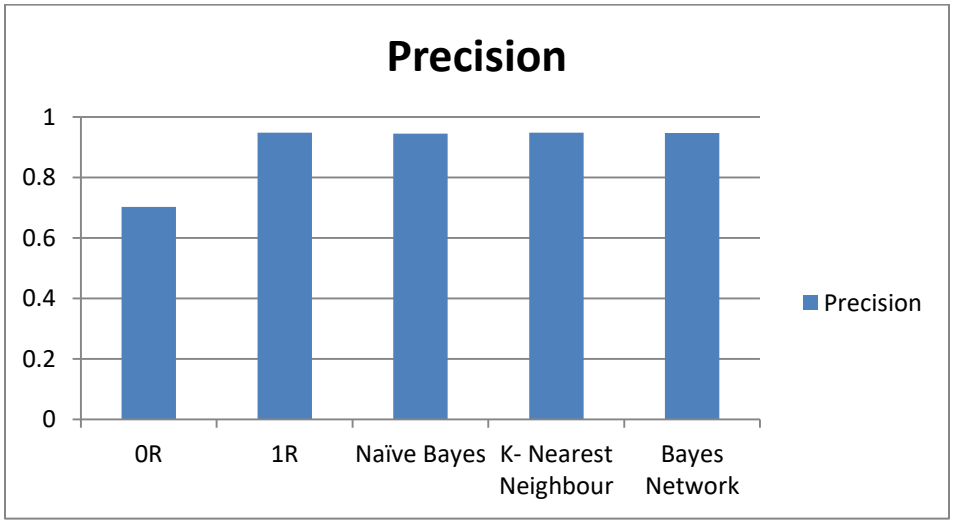

Figure 3: Comparison of precision between classification techniques

ISSN: 2454-1907
DOI: 10.5281/zenodo.1006814

As we can see, 0R as the baseline of benchmarking performance records lower precision, the remaining algorithms prove they are very precise in accuracy with more than 0.95 of precision. The improvement of algorithms from 0.70 to more than 0.95 indicates that from 10-folds cross validation of training and testing, these algorithms are learning the patterns.

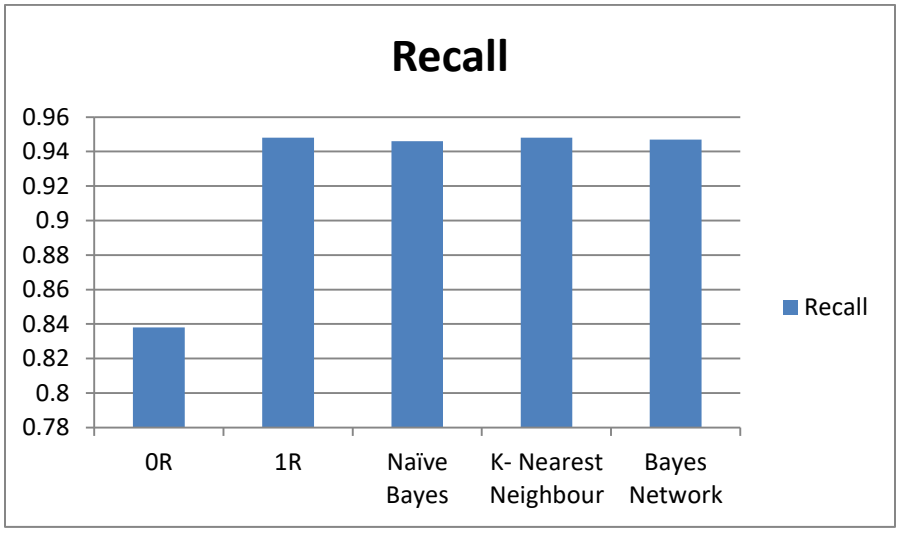

Figure 4: Comparison of recall between classification techniques

The results section should provide details of all of the experiments that are required to support the conclusions of the paper. The section may be divided into subsections, each with a concise subheading.

It is advised that this section be written in past tense. It is a good idea to rely on charts, graphs, and tables to present the information. This way, the author is not tempted to discuss any conclusions derived from the study. The charts, graphs, and table should be clearly labeled and should include captions that outline the results without drawing any conclusions. A description of statistical tests as it relates to the results should be included.

4. Conclusions and Recommendations

Naïve Bayes, Bayes Network and k-NN are algorithms with very high accuracy but as k-NN is lazy classifier; if the dataset is large the time consumed will be directly proportional to size of dataset so it is not effective. Thus, Naïve Bayes and Bayes Network are the best techniques to perform classification. Based on the experiment, we manage to determine the most suitable classification techniques that can be used in log mining, which is Bayes Network and Naive Bayes algorithm.

It is suggested different datasets and classification techniques to be used during classification comparison. Future works should also determine benchmark for clustering techniques in log mining in the future as unsupervised learning is more difficult. This section may also include also include discussion on theoretical and methodological implications of findings.

ISSN: 2454-1907
DOI: 10.5281/zenodo.1006814

Acknowledgements

This work was supported by UniversitiSains Malaysia under RUI Grant Scheme (1001/PKOMP/8012206).

References

[1] Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N. (2002): Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorations 1, pp. 12-23.
[2] Jose M. Domenech, Javier Lorenzo (2007). A Tool for Web Usage Mining. 8th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'07), Birmingham, UK.
[3] Anand S. Lalani (2003). Data Mining of Web Access Logs. A minor thesis, School of Computer Science and Information Technology, Faculty of Applied Science, Royal Melbourne Institute of Technology, Melbourne, Victoria, Australia.
[4] James Pitkow (1997). In search of reliable usage data on the www. In Proc. of the Sixth International WWW Conference, pp. 1-13.
[5] Teressa T. Chikohora (2014): A Study of the Factors Considered when Choosing an Appropriate Data Mining Algorithm. International Journal of Soft Computing and Engineering (IJSCE) , Volume-4, Issue-3
[6] Berson, A., Smith, S., Thearling, K. (ed.) (1999). Building Data Mining Applications for CRM. NewYork: McGraw-Hill.
[7] Aniket Dash (2010) Web Usage Mining: An Implementation. A minor thesis, National Institute of Technology, Rourkela, India.
[8] Robert Cooley, BamshadMobasher, and Jaideep Srivastava (1999). Data Preparation for Mining World Wide Web. Browsing Patterns Knowledge and Information Systems Vol. 1 Issue 1, pp 5-32.
[9] Spilipoulou M., Mobasher B, Berendt B. (2003) "A framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis," INFORMS Journal on Computing spring.
[10] ChitraNasa, Suman (2012). Evaluation of Different Classification Techniques for WEB Data. International Journal of Computer Applications (0975–8887), Volume 52.

*Corresponding author.
E-mail address: nazmee@ usm.my