Indian Journal of Fundamental and Applied Life Sciences ISSN: 2231– 6345 (Online)
An Open Access, Online International Journal Available at www.cibtech.org/sp.ed/jls/2014/03/jls.htm
2014 Vol. 4 (S3), pp. 1511-1518/Kashani et al.
Research Article

CONVERSION LOG DATASET INTO UNDERSTANDABLE FORMAT FOR DATA MINING

***Ali Azimi Kashani, Alimohamad Monjezi Noori and Hadi Mahriyar**
Department of Computer Engineering, Shoushtar Branch, Islamic Azad University, Shoushtar, Iran
**Author for Correspondence*

ABSTRACT
In recent years, the importance of log analysis has grown. Log data constitute a relevant aspect in understanding user behaviour. The volume of log is increasing rapidly and the format is differ to each organization. Having knowledge in analysing log would be an important part in log analysis. An efficient data pre-processing technique can help to improve the quality of data thus help to improve the accuracy and the efficiency of the subsequent data mining analysis. In this research, systematic analysis is done on existing techniques to transform raw dataset into readable format. Hence, transforming the webserver logs into a database for data mining analysis with the aims to help researchers gather knowledge about log data.

Keywords: *Log Analysis, Data pre-processing, Data Transformation, Data Mining, Knowledge Database Discovery, Systematic Analysis*

INTRODUCTION
Nowadays, the importance of log analysis has grown. Log data constitute a relevant aspect in understanding user behaviour. When user accesses to the website, log files are created. Log file recorded all related information regarding to the user such as host name, IP address, date and time, URL and request type. Information are usually stored in log file by using different log format such as NCSA (National Center for Supercomputing Application) common log, W3C Extended Log File and IIS log file format. The log file can be found on different location, for example, in web server, web proxy server and client browser (Patil, 2012). In order to understand the results of search engines or user behaviour, applications and computer system, it is necessary to do analysis on log file related to the website. However, the volumes of log nowadays have growing rapidly in respond to the tremendous uses of web. Therefore, there is a crucial need for a new generation of theories and tools to assist human in extracting useful information.

Knowledge database discovery (KDD) is considerable importance and necessity because of the accessibility and abundance of data today (Maimon and Rokach, 2005). KDD processes comprise of nine steps which are domain understanding and KDD goals, selection and addition, pre-processing, data transformation, choosing the appropriate data mining task, choosing the data mining algorithm, employing the data mining algorithm, evaluation and utilize the discovered knowledge (Fayyad *et al.*, 1996). For this research, the scope of dataset will be the USIM web server log and KDD technique was used to do the data transformation.

Data transformation is the process where the data are transformed or consolidated into forms appropriate for subsequent analysis. The data are presented in certain representation in order to be transformed into nominal data. This research aims to transform the raw dataset into understandable format for subsequent data mining analysis. The domain categorisation were conducted and obtained prior to the data transformation process.

Apart from that, the motivations for this research are:

i. The Importance of how to Analyse Data and Transform Data into Readable Format

There are many information display in the raw log data. Analysing web server log helps to understand who visited the website, on what application, how often and also provide with information on errors, problems and performance problems of the web application. However, analysing the data without having knowledge to analyse the log would lead to wrong interpretation.

ii. To Analyse log would be a Daunting Task

Nowadays, the size of log is increasing. The level of difficulty to analyse log become higher proportional to the number of user's accessess.

iii. The Difficulty for the Researcher to Transform Raw Dataset into Readable and Understandable Format for Subsequent Analysis

Transforming raw dataset into readable and understandable format need to be carried out prior to data mining analysis in order to ensure the outputs produced in the subsequent analysis will have a better accuracy and low false positive rates.

This research paper aims to transform the raw dataset based on the motivation above. The log dataset was obtained from the Pusat Teknologi Maklumat (PTM), Universiti Sains Islam Malaysia (USIM).

This paper is organised as follows. Section II presents the related works with big data and log. Section III explains the methodology used in this paper which consists of general overview on process to transform log dataset into understandable format and lab environment. Section IV presents the research finding which consists of log dataset classification, data transformation and machine learning algorithm results and section V is a closing remarks and summaries the future work of this research paper.

Related Work

Big Data

The variety, velocity and volume of data coming into an organization continues to reach unprecedented levels. This phenomenal growth of data requires not only to understand the big data but also what do with it. Table 1 shows the summarisation of the challenges of handling big data.

Table 1: Summarisation of the challenges of handling big data

Author and Title	Challenges
Michael and Miller (2013) Big Data: New Opportunities and New Challenges	New challenges with respect to how much data to store, whether the data will be secure, how long it must be maintained and how much it will cost.
Hemerly (2013) Public Policy Considerations for Data-Driven Innovation	Policy makers must take into account the possibility that regulation could preclude economic and social benefits in order to achieve the maximum benefits from data-driven innovation.
Tallon (2013) Corporate Governance of Big Data: Perspectives on Value, Risk and Cost	Developing governance mechanisms, policies and structures become the challenges to the organisations. It strike the balance between reward and risk in the face of growing quantities of data and innovation that deliver cheaper storage technology, faster and better.
Pitt *et al*., (2013) Transforming Big Data into Collective Awareness	Numerous challenges revolving around the notion of justice including distributive justice, procedural justice, interactional justice, natural justice and retributive justice. Another challenges include to resist the spread of misinformation.
Wagon and Clarke (2013) Big Data's Big Unintended Consequences	Data quality, semantic coherence and legality appear to be of little concern to those responsible for national security applications.

Log

Log file is an auxiliary text files that software application often produce (Valdman, 2011). Logging is a process of recording events or statistics to provide information about performance or system use (Bishop *et al.*, 1996). Log is used for record data on what, when, who, where and why (W5 questions) an event occurred for security professional on a particular application or device. Log is composed of log entries where information related to a specific event that has occurred within a network or system are contain in each entry (Allen, 2001). Different web servers maintain different types of information in the log files (Ratnesh *et al.*, 2009). The basic information in the log files are;

[x,y,z,a,b,c,d,e,f]

- **User Name (x):** Identifies who had visited the website. The identification of user is based on the IP address that is assigned by the Internet Service Provider (ISP).

- **Visiting Path (y):** The path taken by the user while visiting the website by using URL directly or by clicking on the link or through the search engine.

- **Path Traversed (z):** Identifies the path taken by the user within the website using several of links.

- **Time Stamp (a):** The time or session spent by the user in web page while browsing.

- **Page Last Visited (b):** The last page visited by the user before he or she leaves the website.

- **Success Rate (c):** The success rate of the website which determined by the number of downloads made and the number of copying activity done by the user.

- **User Agent (d):** User's browser information.

- **URL (e):** The resource that had been accessed by the user.

- **Request Type (f):** The information transfer method such as GET and POST.

The activity is recorded in web log file when user submit request to a web server.

Nowadays logs serve many functions within an organizations such as optimizing network and system compared to previous era which used primarily for troubleshooting the problems (Stout and Kent, 2002). Log is useful when involving with audit and forensics analysis, establishing baselines, supporting internal investigations and identifying operational trends and long-term problems (Ahmed *et al.*, 2009). Log file is used to study a user's query behaviour while user navigates a search site (Saxena *et al.*, 2012). Understanding user's navigational preferences is the way to improve query behaviour. The user access patterns information help service provides to modify and adapt their sites interface for individual users as well as to improve the site's static structure within the wider hypertext system.

Log files can be complex and often very large. Analysing log files could be a difficult task although the process of generating log files is straightforward and simple.

MATERIALS AND METHODS

Methodology

The followings are the procedure to transform log dataset into understandable format.

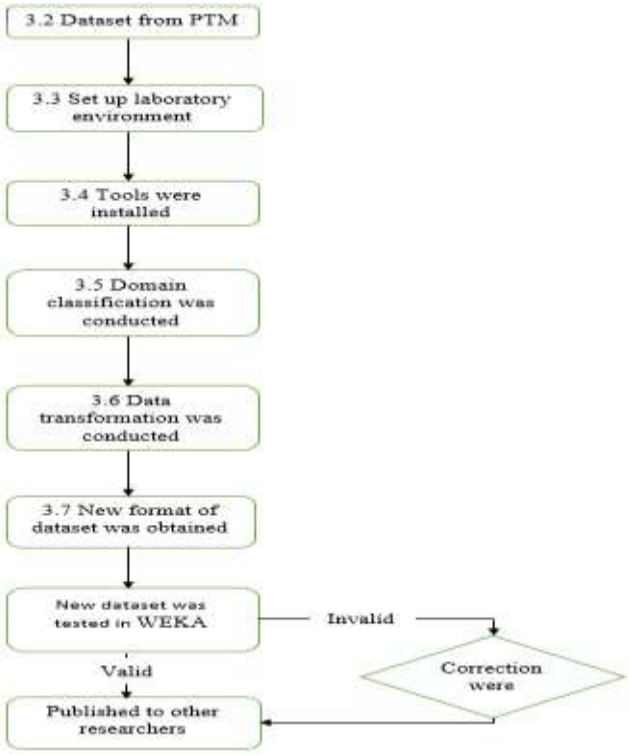

Figure 1: General overview on process to transform log dataset into understandable format

Dataset

The first step was cleaned up the dataset. The dataset is cleaned in excel format. Figure 2 shows the log format.

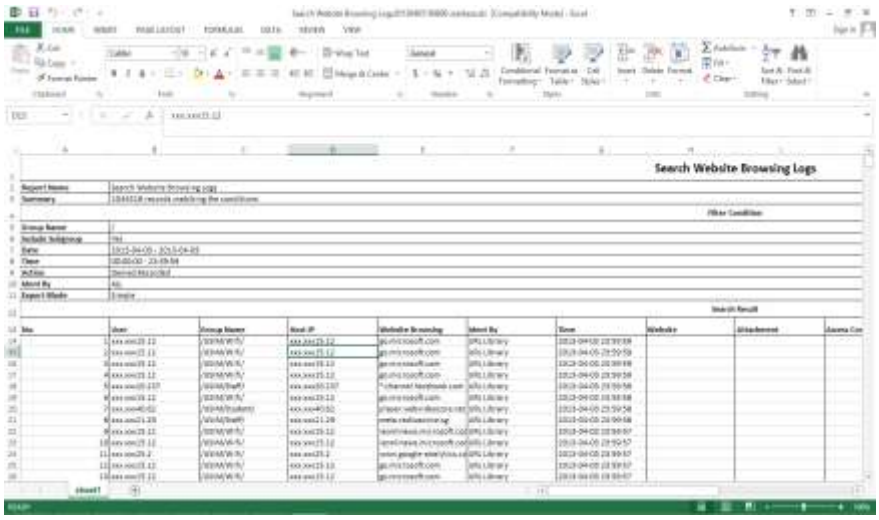

Figure 2: Log file in xlsx format

Based on the log, the data are separated into 15 tabs which are user, group names, host IP, website browsing, identify by, time, website, attachment, access control policy, website link, destination IP, service port, application type, URL type and action. However, in this research only 3 tabs will be extracted along with its data which are website link, service port and URL type. The other information is not important as it may have same data, dynamic host IP, or incomplete data. The dataset will then be classified into three categories which are website link, service port and URL type.

Research Article

Research Environment

The laboratory is set up with one pc only. The lab use virtual private network for domain classification. Figure 3 shows the architecture of the laboratory.

Figure 3: PC 1 for storing and log files analysis

The following in Table 2 is the lists of the tools used in this lab. Almost all tools used are an open source software or free basis except Microsoft Excel.

Table 2: Tools list

Function	Software/Hardware	Description
Display log file	Excel	Display log dataset in xlsx format.
Data mining tool	WEKA	Cluster and classify the new format dataset.
Storing and testing tool	PC 1	For storing log dataset and testing the dataset in WEKA.
Domain classification	Check URL category (http://www.commtouch.com/url-miscat/)	Identifying the web URL type for *Others* category. It is an open source tool.

RESULTS AND DISCUSSION

Findings

The following are the findings of this research. A new dataset classification is produced to ease the data transformation process. The output of the new clean dataset is transformed into nominal dataset which can be used for the subsequent analysis using data mining algorithm (Mertz and Murphy,1996). WEKA software is used to conduct the mining analysis.

Log Dataset Classification

The log dataset has been classified based on three classification which are website link, service port and URL type as displayed in Figure 4.

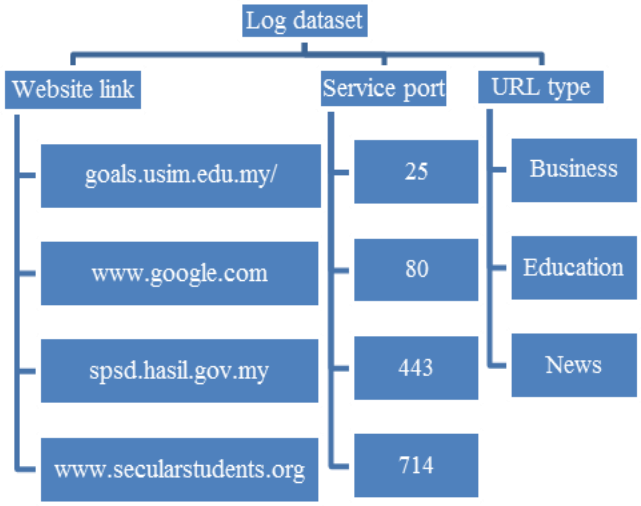

Figure 4: Log dataset classification format

Data Transformation

Dataset 1			New format for dataset 1
Website link	Service port	URL type	
mangastream.com/	80	Entertainment	P11, P22, P37

Log dataset attributes

Dataset transformed into nominal data with numeric

P11 represents website link – as mangastream.com/

P22 represents service port – as port 80

P37 represents URL type – as entertainment

Figure 5: Data Transformation (using certain number representation)

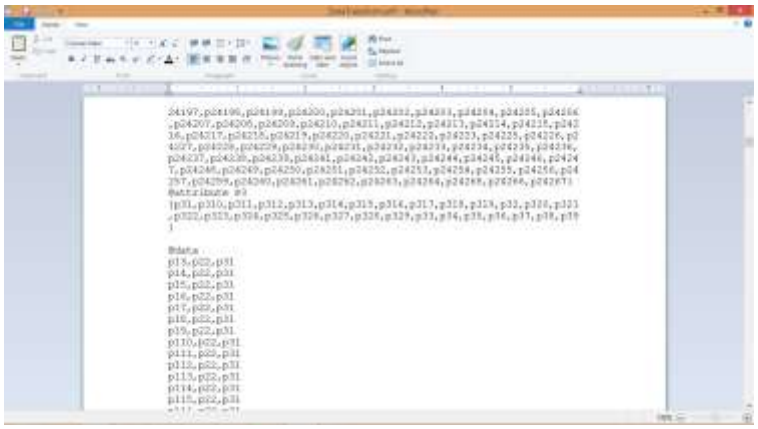

Figure 6: arff format of dataset

Figure 7: Arff file was uploaded in WEKA machine

The researcher identified the nominal data for each of dataset. Figure 5 shows the data transformation details. After conducting data transformation for each of dataset, the researcher came out with clean log dataset file that was compatible with WEKA machine. Figure 6 show the new format of dataset.

The new file is an arff format. The arff file will be used as input in WEKA machine to validate. Figure 7 show the arff file was uploaded into WEKA machine.

The purpose of this figure is to show how the data can be mined in WEKA software. Later, the data can be analysed using different machine learning algorithm inside WEKA.

Machine Learning Algorithm Results

This section presents the finding results of the dataset built. Naïve Bayes and k-nearest neighbour (IBk) algorithm is chose to classify the dataset and the result as displays in table 3. Both results from two

algorithms were compared. True positive rate (TPR) and false positive rate (FPR) are used during the experiment. The experiment was conducted using WEKA software.

Table 3: Machine learning algorithm result

	Naïve Bayes (%)	IBk (%)
TPR	100	99.6568
FPR	0	0.3432

TPR represents true positive rate, FPR represent false positive rate
Table 3 shows the result of TPR for Naïve Bayes algorithm is 0.3432% higher than IBk algorithm which indicates a good result.

Conclusion

As a conclusion, this research manage to provide a new clean log dataset to be published to help researchers in data mining research. Besides that, the dataset was presented in nominal data which was compatible to be used directly in WEKA machine leraning algorithm for data mining process. Furthermore, based on the experiment conducted using WEKA software, the True Positive Rate (TPR) of the classified data 100 % is produced. This result can be used as reference and comparison by other researchers with the same interests.

For future work, other bigger and different format of dataset is used. Other than that, different classifier algorithm will be used to be compare with.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to Universiti Sains Islam Malaysia (USIM) for the support and facilities provided. This research paper is supported by Universiti Sains Islam Malaysia (USIM) grants and FRGS grant: [PPP/FST/SKTS/30/12712],[PPP/FST/SKTS/30/12812] and [FRGS/1/2014/I CT04/USIM/02 /1].

REFERENCES

Ahmed MKMH and Raza A (2009). An Automated User Transparent Approach to log Web URLs for Forensic Analysis. *International Conference on IT Security Incident Management and IT Forensics*, Fifth **2**(1).

Allen S (2001). Importance of Understanding Logs from an Information Security Standpoint. *GSEC* v1.2f. SANS Institute.

Bishop M, Wee C and Frank J (1996). Goal-Oriented Auditing and Logging. *ACM Transactions on Computing Systems*, accessed on 10 December 2013. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.160.9479&rep=rep1&type=pdf.

Fayyad U, Piatetsky-Shapiro G and Smyth P (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*.

Hemerly J (2013). Public Policy Consideration for Data-Driven Innovation. *Journal of Computer* **46**(6) 25-31.

Maimon O and Rokach L (2005). Introduction to knowledge discovery in databases. In: *The Data Mining and Knowledge Discovery Handbook*, edited by Maimon O and Rokach L (Springer-Verlag) New York 1–13.

Mertz CJ and Murphy PM (1996). UCI Repository of machine learning databases. University of California (Electronic version). Available: http://www.ics.uci.edu/~mlearn/MLRepository.htm (Accessed 10 December 2013).

Michael K and Miller KW (2013). Big Data: New Opportunities and New Challenges. *Journal of Computer* **46**(6) 22-24.

Patil PU (2012). Preprocessing of Web Server Log File for Web Mining. *World Journal of Science and Technology* **2**(3) 14-18, ISSN: 2231-2587.

Pitt J, Bourazeri A, Nowak A, Roszczynska-Kurasinska M, Rychwalska A, Santiago IR, Sanchez ML, Florea M and Sanduleac M (2013). Transforming Big Data into Collective Awareness. *Journal of Computer* **46**(6) 40-45.

Ratnesh Kumar Jain, Kasana RS and Suresh Jain (2009). Efficient Web Log Mining using Doubly Linked Tree. *International Journal of Computer Science and Information Security* **3**.

Saxena M, Singh NK, Thakur SS and Kumar P (2012). A Review of Computer forensic & Logging System. *International Journal of Advanced Research in Computer Science and Software Engineering*, ISSN: 2277 128X **2**(1).

Stout and Kent (2002). Central Logging with a Twist of COTS in a Solaris Environment. *SANS Institute*. Last accessed on 10 December 2013, Available: http://www.sans.org/rr/papers/52/540.pdf.

Tallon PP (2013). Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost. *Journal of Computer* **46**(6) 32-38.

Valdman J (2011). Log File Analysis. *DCSE*/TR-2001-04. University of West Bohemia in Pilsen. Czech Republic.

Wigan MR and Clarke R (2013). Big Data's Big Unintended Consequences. *Journal of Computer* **46**(6) 46-53.