An Efficient Data Transformation Technique for Web Log

Madihah Mohd Saudi, Farida Ridzuan, *Member, IAENG*, and Hasan Al-Banna Hashim

***Abstract*—Growth of data over time especially in term of volume, velocity, value, veracity and variety led to many challenges especially in extracting useful information from it. Furthermore, managing and transforming raw data into a readable format is crucial for subsequent analysis. Therefore, this paper presents a new web server log file classification and an efficient way of transforming raw web log files by using knowledge database discovery (KDD) technique into a readable format for data mining analysis. An experiment was conducted to the raw web log files, in a controlled lab environment, by using KDD technique and k-nearest neighbor (IBk) algorithm. Based on the experiment conducted, the IBk algorithm generates 99.66% for true positive rate (TPR) and 0.34% for false positive rate (FPR) which indicates the significant efficiency of the new web log file classification and data transformation technique used in this paper.**

***Index Terms*— Log analysis, Data transformation, Knowledge database discovery, Big Data.**

I. Introduction

Nowadays, large amount of data are generated with different formats and an efficient technique to transform these data is very important. The 5Vs (variety, velocity, volume, veracity and value) of data in an organization continue to expand than its ordinary level [1]. For examples in term of volume increase, Boeing 787 has created half a terabyte of data per flight and Parkinson disease dataset has created 150GB that tracked 12 patients using smartphone sensors for eight weeks [2,3]. As for variety, for examples are the traffic dataset which contains numeric information about vehicles such as speed, volume, travel times and textual information about events and in medical domain, patients' feedbacks and various physiological, chemical measurements and physical can be combined and abstracted to obtain their well-being and health [4,5]. While example of the velocity, can be referred as the data for high speed especially for social media messages and credit card transactions [6]. On the other hand, veracity of data refers to the data trustworthiness [7]. Table 1 shows a summary on the challenges in handling big data.

TABLE I
SUMMARIZATION ON THE CHALLENGES OF HANDLING BIG DATA

Authors' Name	Challenges
Michael K. and Miller K.W. [8]	Amount of data to be stored, whether the data will be secure, maintenance duration and cost.
Hemerly J. [9]	Regulation that could preclude economic and social benefits to achieve the maximum benefits from data-driven innovation.
Tallon, P.P [10]	Developing governance mechanisms, policies and structures become the challenges to the organizations. Balance between reward and risk in the face of growing quantities of data and innovation that deliver cheaper storage technology, faster and better in needed.
Pitt, J., Bourazeri, A., Nowak, A., Roszczynska-Kurasinska, M. et. al. [11]	Numerous challenges revolving around justice including distributive justice, procedural justice, interactional justice, natural justice, retributive justice and resisting the spread of misinformation.
Wigan, M.R. and Clarke, R. [12]	Data quality, semantic coherence and legality appear to be of little concern to those responsible for national security applications.

While in a computer, a log file is created to record any events, messages, activities and software that run in an operating system of the computer. It is an auxiliary text files that software application often produce [13]. Logging is a process of recording events or statistics to provide information about performance or system use [14]. Log is used to 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 an entry that consists of information related to a specific event that has happened within a network or system [15].

As for web server, a log file is created to record all information related to the user such as host name, IP address, date and time, URL and request type [16]. It comes in different kind of formats such as National Center for Supercomputing Application (NCSA) common log, The World Wide Web Consortium (W3C) Extended Log File or

Manuscript received July 18, 2016; revised August 15, 2016. This work was supported by Ministry of Higher Education (Malaysia), FRGS grant: [FRGS/1/2014/I CT04/USIM/02 /1].

Madihah Mohd Saudi is with the Faculty of Science and Technology (FST), Universiti Sains Islam Malaysia (USIM), as a lecturer and a research fellow in Islamic Science Institute, Universiti Sains Islam Malaysia (USIM), Bandar Baru Nilai, Nilai, 71800, Malaysia. (e-mail: madihah@usim.edu.my).

Farida Ridzuan is with the Faculty of Science and Technology (FST), as a lecturer and a research fellow in Islamic Science Institute, Universiti Sains Islam Malaysia (USIM), Bandar Baru Nilai, Nilai, 71800 (email: farida@usim.edu.my).

Hasan Al-Banna Hashim is with the Information Security and Assurance (ISA) programme, Faculty of Science & Technology (FST), Universiti Sains Islam Malaysia (USIM), 71800 Nilai, Negeri Sembilan, Malaysia. Malaysia.

ISBN: 978-988-14047-4-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Internet Information Server (IIS) log file format. If anything happened to the computer, a user can always refer to the web log file to obtain information on any activities carried out earlier. However, due to the mass volumes growth of the log files, it is really a daunting task and time consuming to analyse these log files. Nowadays, log is used for many purposes within organizations such as to optimize network and system compared to previous era where it was used primarily for problems troubleshooting [17]. Log is useful for audit and forensics analysis, establishing baselines, supporting internal investigations and recognizing trends of operational and long-term problems [18]. Log file is also used to investigate a user's query behavior [19]. Understanding user's navigational preferences is the way to improve query behavior. The user access patterns information help service providers to modify, adjust their web interface for users individually and to improve the site's static structure within the wider hypertext system. Log files can be complex and often very large. Analyzing log files could be a difficult task although the process of generating log files is straightforward and simple. Table 2 shows the comparative study on web log analysis.

TABLE II
COMPARATIVE STUDY ON WEB LOG ANALYSIS

Title	Method	Strengths	Weaknesses
Efficient Web Log Mining using Doubly Linked Tree. [20]	Doubly Linked Tree	Low run time which is 100 seconds for low support threshold and large web access sequence database	Old dataset retrieved in 2004
Efficient Web Log Mining Using Enhanced Apriori Algorithm with Hash Tree and Fuzzy. [21]	-Enhanced Apriori with Hash Tree and Fuzzy	High efficiency with 97%	N/A
Analyzing Log Files using Data-Mining. [22]	-Classification algorithm -Regression algorithm -Segmentation algorithm -Association algorithm -Sequence analysis algorithm	Applied data mining tool	Small size of dataset
Mining Log Files for Data-Driven System Management. [23]	-Naïve Bayes -Modified Naïve Bayes -Hidden Markov Model (HMM)	Both Modified Naïve Bayes and HMM enhance the classification accuracy	Constructing HMM requires more computational costs

Therefore, an efficient technique or tool to assist user prior analyzing and extracting useful information from the log file is needed to make sure the information retrieved is accurate and in a faster mode. Prior that, the preprocessing steps and techniques should be very clearly defined. Table 3 shows a comparative study on existing data pre-processing techniques.

TABLE III
COMPARATIVE STUDY ON PRE-PROCESSIG TECHNIQUES

Title	Source of Log File	Pre-processing Technique	Algorithm Applied
A Novel Technique for Web Log Mining with Better Data Cleaning and Transaction Identification [24]	College Web Site	-Data cleaning -User identification -Session identification -Path completion -Transaction identification	MFR RL & Time Window
Research on Path Completion Technique in Web Usage Mining [25]	English Study Web site Log File	-Data cleaning -User identification -Session identification -Path completion -Transaction identification	Maximal Forward References (MFR), Reference Length
Hierarchical Sessionization at Preprocessing Level of WUM Based on Swarm Intelligence [26]	Server Log File	-Data cleaning -Log file filtering -User identification -Session identification	NA
Optimized Data Preprocessing Technology For Web Log Mining [27]	IIS Server Log File	-Data cleaning -User identification -Session identification -Path completion	Based on referred web page and fixed priori threshold
Advanced Data Preprocessing for Intersites Web Usage Mining [28]	Log Files from INRIA web sites	-Data fusion -Data cleaning -Data structuration -Data summarization	NA
Web Log Data Preprocessing Based on Collaborative Filtering [29]	Web server Log file	-Data pre-processing	Based on Collaborative Filtering
A Session Identification Algorithm Based on Frame Page and Page threshold [30]	Chizhou College Website	-Data filtering -Session identification	Frame page and Page Threshold

Based on the comprehensive review analysis of the existing works related with the preprocessing techniques and data transformation, knowledge database discovery (KDD) is seen as one of the promising techniques to make the whole processes especially related with data transformation from raw data to useful information faster and more efficient [31]. The KDD processes comprises of nine steps which are domain understanding and KDD goal, selection and addition, pre-processing, data transformation, selecting the appropriate data mining task, selecting the data mining algorithm, employing the data mining algorithm, evaluation

ISBN: 978-988-14047-4-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

and utilizing the discovered knowledge [32]. The process of transforming data into an appropriate form for subsequent analysis is known as data transformation. The data are presented in certain representation in order to be transformed into nominal data.

Hence, this paper's objectives are to produce a new web server log file classification and to introduce an efficient technique to transform the raw web server log files into understandable format using the KDD technique and for subsequent data mining analysis using k-nearest neighbor (IBk) algorithm. The new web server log file classification is used as the input and is created prior doing the data transformation to ensure a better accuracy result for the retrieved information can be achieved.

This paper is organized into four sections. Section 2 presents the methodology used in this research. Section 3 presents the results discussion. Section 4 concludes and summarizes the potential future work of this research paper.

II. Methodology

Different web servers maintain different types of information in the log files [33]. The activity is recorded in web log file when user submit request to a web server. The information in the log files as listed in the following.

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

User Name (x): Detects who had visited the web and user identification 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 search engine, typing Uniform Resource Locator (URL) directly or by clicking on the link.
Path Traversed (z): Detects the path taken by user using several links within the web site.
Time Stamp (a): The session or time spent by the user in web site while browsing.
Page Last Visited (b): The latter page visited by the user before he or she leaves the website.
Success Rate (c): The rate of success for the website which is determined by the number of copying activity done by the user and the number of downloads.
User Agent (d): User's browser information.
URL (e): The resources accessed by the user.
Request Type (f): The information transfer method such as GET and POST.

Fig. 1 shows the procedure to transform log dataset into understandable format. The first step is to clean up the dataset to remove outliers, inconsistencies, incomplete and missing values.

Fig. 2 shows the log format. Based on the log, the data is separated into 15 information columns 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, the log dataset is cleaned up by removing cells with missing values and incomplete. Information that will be extracted along with its values are website link, service port and URL type for subsequent analyses. The other information columns are removed because they contain repetitive values, using dynamic host IP and incomplete data. Next, the dataset were classified into based on the new web log file classification. The details can be referred in section 3. This new web log file classification consists of three categories which are website link, service port and URL type.

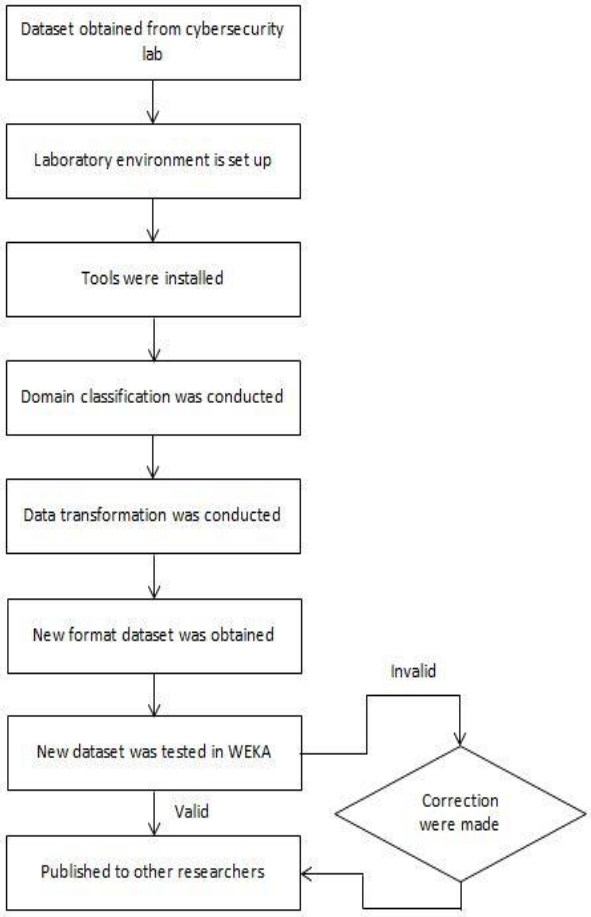

Fig. 1. Research processes involved in transforming log dataset into understandable format

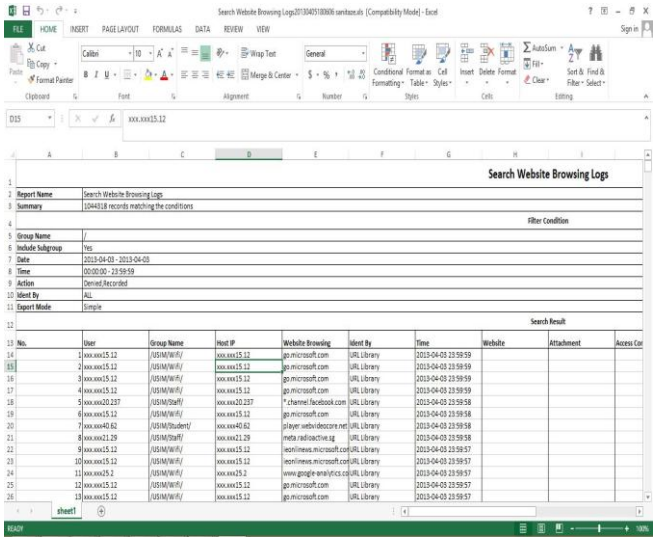

Fig. 2. Log file in xlsx format.

The controlled lab environment is using a PC with virtual private network for categorizing the domain. Table 4 shows the list of tools used in this research. All applications used in this research were free basis and open source except Microsoft Excel.

ISBN: 978-988-14047-4-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE IV
TOOLS LIST

Function	Application	Description
Display log file	Excel	Display logs dataset in xlsx format.
Data mining tool	Waikato Environment for Knowledge Analysis (WEKA)	Cluster and classify the new format dataset.
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.

III. RESULTS AND DISCUSSION

A new web log file classification is produced to ease the data transformation processes. The output of the new clean dataset is transformed into nominal dataset, which can be used for the subsequent analysis using data mining algorithm [34]. Waikato Environment for Knowledge Analysis (WEKA) software is used to conduct the mining analysis. WEKA is a machine learning software written in Java and developed at the University of Waikato, New Zealand.

The log files dataset have been classified based on the proposed web log file classification. These are based on website link, service port and URL type as displayed in Fig.3.

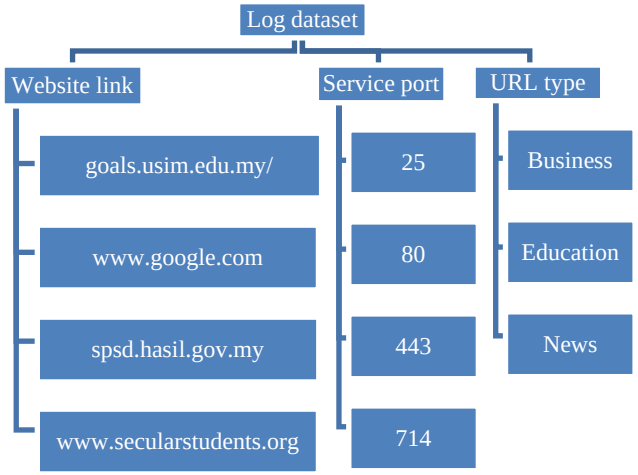

Fig. 3: New Web Log File classification

The nominal data in the dataset have been transformed as displayed in Fig. 4, where it shows the data transformation details.

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

Fig. 4. Data Transformation (using certain number representation).

The clean dataset were obtained after conducting the data transformation processes. Data transformation is essential to produce a dataset that is compatible with the WEKA. Fig. 5 shows the new format of the dataset. Then the dataset is saved in arff format and used as input in the WEKA for validation process. Fig. 6 shows the arff file being uploaded into WEKA machine.

The purpose of this figure is to show how the data can be mined in WEKA software. Useful information can be extracted from the raw log data through the data mining. Later, the data can be analyzed using different kind of machine learning algorithms provided by WEKA.

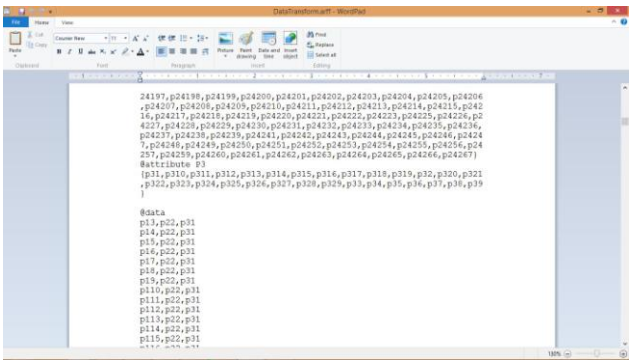

Fig. 5. Arff format of dataset.

Fig. 6. Arff file was uploaded in WEKA and being clustered.

For validation, the new dataset were tested with k-nearest neighbour (IBk) algorithm. IBk can be used to select appropriate value of K based on cross-validation and value for distance weighting [35]. True positive rate (TPR) and false positive rate (FPR) are calculated based on the experiment.

TABLE V
MACHINE LEARNING ALGORITHM RESULT

	IBk (%)
TPR	99.6568
FPR	0.3432

TPR represents true positive rate, FPR represent false positive rate

Table 5 shows that the classification using IBk algorithm managed to achieve 99.6568% TPR and 0.3432% FPR. This result indicates a good result and the classification is

ISBN: 978-988-14047-4-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

successful. Detail results obtained from WEKA is shown in Fig. 7.

```
=== Classifier model (full training set) ===

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0.01 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        7549               99.6568 %
Incorrectly Classified Instances        26                0.3432 %
Kappa statistic                          0.992
Mean absolute error                      0.0035
Root mean squared error                  0.0586
Relative absolute error                  0.8104 %
Root relative squared error             12.6368 %
Total Number of Instances             7575

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.995     0          1        0.995     0.997      0.997    cluster0
                 1         0.005      0.99     1         0.995      0.997    cluster1
Weighted Avg.    0.997     0.002      0.997    0.997     0.997      0.997

=== Confusion Matrix ===

    a    b   <-- classified as
 5182   25 |    a = cluster0
    1 2367 |    b = cluster1
```

Fig. 7. Summary of results generated by WEKA.

IV. CONCLUSION

As a conclusion, this research shows that the new proposed web log file classification and data transformation technique used in this research have successfully produced clean public log datasets for faster analysis and better accuracy result. This new classification and technique discussed in this paper, can be used as reference by other researchers with the same interest. With IbK algorithm, it generates 99.66% true positive rate (TPR). For future work, these clean dataset can be clustered and classified using different machine learning algorithms in WEKA. Furthermore, bigger and different format of dataset can be used for future work.

REFERENCES

[1] Demchenko, Y., Ngo, C., and Membey, P. (2013). Architecture Framework and Components for Big Data Ecosystem. SNE technical report, SNE-UVA.

[2] Finnegan, M. (2013). Boeing 787s to create half a terabyte of data per flight, says Virgin Atlantic. http://www.computerworlduk.com/news/infrastructure/3433595/boeing-787s-to-create-half-a-terabyte-of-data-per-flight-says-virgin-atlantic/. (Accessed on 15/8/2016).

[3] Thirunarayan, K., and Sheth, A. (2013). Semantics-empowered Approaches to Big Data Processing for Physical-Cyber-Social Applications. *Presented in Association for the Advancement*

[4] Anantharam, P., Thirunarayan, K., Sheth, A. (2013). Traffic Analytics using Probabilistic Graphical Models Enhanced with Knowledge Bases. *2nd International Workshop on Analytics for Cyber-Physical Systems, of Artificial Intelligence Fall Symposum Series*, p. 68-75.

[5] Sheth, A., and Thirunarayan, K.. (2012). Semantics Empowered Web 3.0: Managing Enterprise, Social Sensor and Cloud-based Data and Services for Advanced Applications. Synthesis Lectures on Data Management, Morgan & Claypool Publishers 2012.

[6] Sabia and Arora, L. (2014). Technologies to Handle Big Data: A Survey. *International Conference on Communication, Computing & Systems*, p. 6-11.

[7] Maier, M. (2013). Towards a Big Data Reference Architecture. A Master's Thesis.

[8] Michael K. and Miller K.W. (2013). Big Data: New Opportunities and New Challenges. *Journal of Computer*, vol. 46, no. 6, pp. 22-24.

[9] Hemerly J. (2013). Public Policy Consideration for Data-Driven Innovation. *Journal of Computer*, vol. 46, no. 6, pp. 25-31.

[10] Tallon, P.P.(2013). Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost. *Journal of Computer*, Vol. 46, no. 6, pp. 32-38.

[11] Pitt, J., Bourazeri, A., Nowak, A., Roszczynska-Kurasinska, M., Rychwalska, A., Santiago, I.R., Sanchez, M.L., Florea, M., Sanduleac, M. (2013). Transforming Big Data into Collective Awareness. *Journal of Computer*, Vol. 46, no. 6, pp 40-45.

[12] Wigan, M.R. and Clarke, R. (2013). Big Data's Big Unintended Consequences. *Journal of Computer*, vol. 46, no. 6, pp. 46-53.

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

[14] Bishop, M., Wee, C. and Frank, J. (1996). Goal-Oriented Auditing and Logging. *ACM Transactions on Computing Systems*, doi: 10.1.1.160.9479. (Accessed on 15/8/2016). URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.160.9479&rep=rep1&type=pdf.

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

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

[17] Stout and Kent. (2002). Central Logging with a Twist of COTS in a Solaris Environment. *SANS Institute*. (Accessed on 15/8/2016). URL: http://www.sans.org/rr/papers/52/540.pdf.

[18] Ahmed, M.K., M.H. 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 Vol. 2, Issue 1.

[19] Saxena, M., Singh, N.K., Thakur, S.S., 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, Vol. 2, Issue 1.

[20] Jain, R.K., Kasana, R.S. and Jain, S. (2009). Efficient Web Log Mining using Doubly Linked Tree. *International Journal of Computer Science and Information Security*, Vol 3(1).

[21] Veeramalai, S., Jaisankar, N., and Kannan, A. (2010). Efficient Web Log Mining Using Enhanced Apriori Algorithm with Hash Tree and Fuzzy. *International Journal of Computer & Information Technology*, Vol 2(4), pp. 60-74.

[22] Mihut, M. (2008). Analyzing Log Files using Data-Mining. *Journal of Applied Computer Science*, No. 2(2), pp. 32-34.

[23] Peng, W., Li, T., and Ma, S. (2005). Mining Log Files for Data-Driven System Management. *ACM SIGKDD Explorations Newsletter – Natural Language Processing and Text Mining*, Vol. 7(1), pp. 45-51.

[24] Vellingiri, J. and Pandian, S.C. 2011. A Novel Technique for Web Log Mining with Better Data Cleaning and Transaction Identification. *Journal of Computer Science*. Pp. 683-689.

[25] LI, Y., FENG, B. and MAO, Q. 2008. Research on Path Completion Technique in Web Usage Mining. IEEE International Symposium on Computer Science and Computational Technology. Pp. 554-559.

[26] Hussain, T., Dr. Asghar, S. and Masood, N. 2010. Hierarchical Sessionization at Preprocessing Level of WUM Based on Swarm Intelligence. *6th International Conference on Emerging Technologies (ICET)* IEEE. Pp. 21-26.

[27] Zheng, L., Gui, H. and Li, F. 2010. Optimized Data Preprocessing Technology For Web Log Mining. *IEEE International Conference on Computer Design and Applications*. Pp. VI-19-VI-21.

[28] Tanasa, D. and Trousse, B. 2004. Advanced Data Preprocessing for Intersites Web Usage Mining. *IEEE Intelligent Systems* 19, 2. IEEE Computer Society. Pp. 59-65.

[29] Chang-bin, J. and Li, C. 2010. Web Log Data Preprocessing Based on Collaborative Filtering. *IEEE 2nd International Workshop on Education Technology and Computer Science*. Pp.118-121.

[30] Yuankang, F. and Zhiqiu, H. 2010. A Session Identification Algorithm Based on Frame Page and Page threshold. *IEEE Conference*. Pp. 645-647.

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

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

[33] Ratnesh Kumar Jain, Dr. R. S. Kasana1, Dr. Suresh Jain, (2009). Efficient Web Log Mining using Doubly Linked Tree. *International Journal of Computer Science and Information Security*, vol. 3.

[34] Mertz, C.J. and Murphy, P.M. (1996). UCI Repository of machine learning databases. University of California (Electronic version).

ISBN: 978-988-14047-4-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Available from: http://www.ics.uci.edu/~mlearn/MLRepository.htm (Accessed on 15/8/2016).
[35] D. Aha, D. Kibler (1991). Instance-based learning algorithms. Machine Learning. 6:37-66.

ISBN: 978-988-14047-4-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)