Journal of Theoretical and Applied Information Technology
20th April 2014. Vol. 62 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TWO LEVEL CLUSTERING APPROACH FOR DATA QUALITY IMPROVEMENT IN WEB USAGE MINING

[1]YUHEFIZAR, [2]BUDI SANTOSA, [3]I KETUT EDDY P., [4]YOYON K. SUPRAPTO

[1]Lecturer, Dept. of Information Technology, Polytechnic State of Padang, Indonesia
[2]Prof., Dept. of Industrial Engineering, Institute of Technology Sepuluh Nopember, Indonesia
[3]Assoc. Prof., Dept. of Electrical Engineering, Institute of Technology Sepuluh Nopember, Indonesia
[4] Assoc. Prof., Dept. of Electrical Engineering, Institute of Technology Sepuluh Nopember, Indonesia

E-mail: [1]ephi.lintau@gmail.com , [2]budi_s@ie.its.ac.id , [3]ketut@ee.its.ac.id , [4]yoyonsuprapto@ee.its.ac.id

ABSTRACT

Web Usage Mining (WUM) is a term related to the extraction of knowledge from web log data. Web log data has a lot of irrelevant data to proceed WUM**.** Therefore, it requires several steps to get a good quality of data, because the final result of WUM depends on the quality of the input data. Therefore, in this paper we propose a new approach to overcome these problems, hence, it is called the two level clustering approach. The first level clustering is performed on the data in the form of access frequently and use non-hierarchical cluster method, followed by a second level clustering on the web log data in the form of user access. At the second level clustering, it combines cluster hierarchical and non-hierarchical methods. From the experiments, 90.78% on web log data quality is reached

Keywords: *Data Quality Improvement, Two Level Clustering, Web Log Data, Web Mining, Web Usage Mining*

1. INTRODUCTION

The development of the internet today is inseparable from the use of websites as a medium for information dissemination. In fact, every access which is done in every web automatically will be saved by the web server as web log data[1].

Web log data has a set of transactions (clickstream) done by web visitors, so it will describe the visitor behavior pattern. The higher the amount of access to a web, the greater the number of web log data recorded [2]. The Web log data has become an essential part in the research of Web Usage Mining (WUM) [3], to be investigated further, especially in terms of understanding the patterns of visitors' behavior of the web [4]-[5]-[6].

In the e-business/e-commerce web for example, the existence of web log data is very important in order to obtain detailed information on customer behavior patterns, transaction history and even predict the next transaction, as well as in the web of e-learning, e-news, e-banking, e-government and others. The access patterns can be used for many purposes. Therefore, web log data become the main data source in the research of WUM [7].

WUM is one of the important research field of web mining which is associated with the extraction of knowledge from web log data [8]-[9]. Web log data recorded by web server contains many data items that are not needed in the data mining process, therefore, in the process of WUM, pre-processing of web log data must be done in advance to get good data quality[10]-[11]. Some references have clarified the importance of pre-processing activities in this WUM [12].

Pre-processing is an attempt to improve the quality of data such that the data used in WUM is objective, representative and have a small standard of error. Quality of data is main problem in the research of WUM, therefore the final result of WUM depends on the quality of the input data

In this paper, we propose a new method for improving the quality of data in the WUM process, which is the two-level clustering. The first level clustering is done in the form of data frequently user access using non-hierarchical clustering method. At this stage, the cluster that has the lowest members will be deleted. The second level clustering is done by first changing the form of web log data into user access behavior patterns. At this stage a combination of methods applied hierarchical and non-hierarchical cluster. The end result of this second level cluster web log data obtained is ready for further processing using WUM.

Journal of Theoretical and Applied Information Technology
20th April 2014. Vol. 62 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

This paper is organized as follow; chapter 1 explains the background of the research, chapter 2 will explore the literature that is relevant to this research, chapter 3 discusses about the proposed method as well as the stages of the research, chapter 4 is about the experiment and results, and chapter 5 is the conclusion of the research.

2. LITERATURE REVIEW

Some previous researchers have filed several methods in order to improve the quality of web log data. In [12], it is proposed a method using the association rule algorithm to identify unique users and user sessions. N. K. Tyagi et. al[10] have proposed those two algorithms for data cleaning and data reduction. A new approach to find frequent item sets employing through set theory that can extract association rules for each homogenous cluster of transaction data records and relationships between different clusters is proposed [13]. The paper applied an algorithm to reduce a large number of item sets to find valid association rules. They used the most suitable binary reduction for log data from web database. Web Log data pre-processing, the first step of a common WUM process was proposed in [14]. In the working scheme of LODAP four main modules are used, namely data cleaning, data structuring, data filtering and data summarization.

Two techniques of pre-processing web log data were used in [15]. First is data cleaning by erasing picture data from log and second user identification by using 3 attributes (IP Address, Operating System and User Agent). In [16], using a programming approach (algorithm) to move the log data into a database and delete the log data items that are not useful. Similar method was also performed by Theint Aye [8] with an emphasis to improve the quality of log data in web log extraction field, by deleting useless data.

Based on the previous research, it seems that the most of researchers o improve the quality of data (pre-processing) by deleting unnecessary data items based on standard form of the web log data (see Table 1), Meanwhile, the web log data, can be seen by two different forms, i.e. level of frequently user access and user access behavior patterns. This encouraged us to try to make improvements to the quality of web log data based on the two-level view.

3. PROPOSED METHOD

In this paper, we conduct research with several stages as shown in Figure 1.

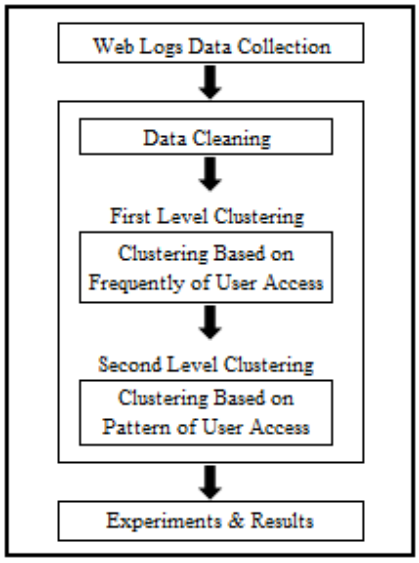

Figure 1: Research Methods

3.1 Web Log Data Format

File format of web log data used is the Common Log Format (CLF) [8]-[17], which is a standard format used by the web server when creating a log file. Each line of CFLs contains the IP address, user identifier, date and time, method, http request, http status code and transfer volume/size, as shown in table 1.

The first row of table 1 informs that the visitor with IP address 66.249.69.xxx have access to a web page news.html on September 4, 2011 at 06:45:13 with 200 status code and a file size of 15319 and so on.

Table 1: Common Log Format

IP Address	Date & time	Me-thod	Request	Sta-tus	Size
66.249.69.xxx	04/Sep/2011:06:45:13	GET	/news.html	200	15319
valley.net	05/Sep/2011:19:08:48	GET	/info.html	200	15582
206.53.148.xxx	05/Sep/2011:19:08:50	GET	/media.jpg	200	1324
Evins.net	05/Sep/2011:19:20:20	POST	/button.gif	200	30462
114.79.16.xxx	06/Sep/2011:20:00:01	GET	/favicon.ico	200	3798

3.2 Data Cleaning

In this stage, the cleaning process of *weblog* data is conducted in order to remove the unwanted data items (*irrelevant data*). This data cleaning process is done based on four criteria.

Journal of Theoretical and Applied Information Technology
20th April 2014. Vol. 62 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The first is based on the file extension. The accepted file extensions are .html, .php, .jsp, asp and other extensions which directly referring to a web page. Data items with file extensions such as: .jpg, .gif, .ico. .bmp, .cgi, .swf, .css, .txt does not show the behavior of web visitor so this data items will be deleted [18].

The second is based on respond code from web server. Web server response with code 200 indicates that the request of accessing a web page is accepted and displayed by the web server. Therefore, all data items unless code 200 will be deleted [5].

The third is based on access methods. Only access with GET method that shows the behavior of web visitors. Data items with other access method, such as HEAD and POST will be deleted [18].

The fourth is based on user access frequency. Only users with access more than five times were used in this research, because it is assumed that visitors with access less than 5 times cannot properly describe the behavior of users.

3.3 Frequently Access

After the data cleaning, the data format is converted in the form of frequently of user access. Here, each user is shown by the number of access to any web page visited. It is presented in a matrix vector [19], as in equation (1).

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix} \quad (1)$$

With *m* is the number of web visitors, *n* is the number of web pages, and X is a vector of observations. Implementation matrix vector in equation (1) shown in table 2.

Table 2: Matrix Vector

User	Web page						
	p1	*p2*	*p3*	*p4*	*p5*	...	*pn*
u1	6	9	0	0	0	…	X_{1n}
u2	0	0	0	35	0	…	X_{2n}
u3	0	11	0	0	0	…	X_{3n}
u4	0	1	4	3	2	…	X_{4n}
u5	0	84	0	0	0	…	X_{5n}
:	:	:	:	:	:	:	:
Um	X_{m1}	X_{m2}	X_{m3}	X_{m4}	X_{m5}	...	X_{mn}

With *p1, p2, p3, pn* are variables for web page, for example *p1* is web page named news.html. *u1, u2, u3, um* are variables for web visitors, such as *u1* is a web visitor with IP address 66.249.69.xxx. From table 2 we can conclude that visitor with varible *u1* has accessed webpage *p1* six times, webpage *p2* nine times and so on.

3.4 Clustering Based On Frequently Access

Clustering based on web log data in form of frequently access format will be done in this stage. The purpose is for data reduction, so data inside cluster with the smallest number will be removed.

In this case, K-Means method [19] was used with the following steps:

1) Determine the number of *k* as many as the number of cluster which is formed. This is also intended to represent the starting centroid.
2) Data are allocated randomly into cluster based on the nearest centroid.
3) Recalculate the *centroid* k position.
4) Repeat step 2 and 3 until inter-cluster object moving no longer exist.

3.5 User Access

Data from the result of clusterization in the first stage, (some of data are already removed), will be proccessed again based on pattern of user access[20]. User access is visiting activity of user to a web based on time sequence. This makes the data format changed, as seen in table 3.

With u_1, u_2, u_3, *m* are the variable for a user access. 1,2,3 and so on are the variable of the visited web page, for example, variabel 1 is referred to the web page news.php, 2 is referred to the web page sport.php and so on. From table 3, it can be concluded that user with id='u_1' have accessed the web page 1, then web page 2, then web page 4 and so on, so it will be shown user behaviour pattern. Then the data is done with the pattern of the second stage clustering.

Table 3: User Access

ID	Web page					
u_1	1	2	4	2	5	3
u_2	2	4	6	2	5	2
u_3	3	5	2	5	4	1
:	:	:	:	:	:	:
m	3	4	6	6	8	4

3.6 Clustering Based On User Access

In this case, combination of clustering hirearchy with single linkage method is used with non hierarchy method. Single linkage method cluster the data based on the closest distance between objects. The distance between D(X,Y)

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

with cluster X and Y is calculated with equation (2).

$$D(X,Y) = \min_{x \in X, y \in Y} d(x,y) \quad (2)$$

where X and Y are any two sets of elements considered as clusters, and $d(x,y)$ denotes the distance between the two elements x and y[2].

Amount of best clusters are determined based on elbow rule in where the result will be used to apply non-hierarchical clustering method.

4. EXPERIMENTS AND RESULTS

This part explains about the experiments and result from two level clustering method.

4.1 Web Log Data Collection & Data Cleaning

Web log data used in this study comes from this Clark Net web log data (http://ita.ee.lbl.gov/html/traces.html), by selecting ClarkNet web log data on September. Total size of web log data used is 168 MB. After the data cleaning process, results are obtained as shown in table 4.

Table 4: Information Web Log Data

	The number of unique users	The number of web pages	Total record
Raw data	72.622	32.041	1.673.798
After cleaning	9.791	2,646	123.692

From table 4, we obtained information that from 1,673.798 recorded data, only 123.692 recorded data or 7.4% of the data that can be processed further. When it is viewed from the unique user side, it is only 9,791 (13.5%) users and 2,646 (8.3%) of web pages.

From the above results it is proved that this stage is crucial in the process of data mining.

4.2 First Level Clustering

After the process of data cleaning, data is transformed as shown in Table 2. Then, clustering using non-hierarchical cluster method (K-Means) applied by dividing the data into 2 clusters, as seen in table 5.

Table 5: Number Of Cases In Each Cluster

Cluster	1	12
	2	9.779
Valid data		9.791
Missing data		0

Clustering objective at this stage is to reduce the data based on the level of access to the website. Cluster members residing on the cluster with the smallest amount of data will be removed at this stage.

4.3 Second Level Clustering

Data results from first level clustering are transformed as shown in Table 3. So, data information is changed as shown in table 6.

Table 6: Information Web Log Data

	The number of unique users	The number of web page	Total record
After cleaning	9.791	2.646	123.692
After first cluster	9.779	2.646	121.658

It can be seen that the number of users are decreased by 12 users and this has effects to the total record.

At this stage, the method clustering is done by combining hierarchical and non-hierarchical cluster. Method begins by using a cluster hierarchy and the elbow rules obtained by the number of cluster. (See table 7).

Table 7: Agglomeration Schedule

Stage	Cluster Combined Cluster 1	Cluster 2	Coefficients	Next Stage
:	:	:	:	:
9775	1	1661	1070774417	9778
9776	2952	9693	1184831043	9777
9777	2952	7307	1649439017	9778
9778	1	2952	2068619188	0

It can be seen on the stage 9776 that the biggest coefficient of increment/leap occurs compared with other stages. Based on elbow rules, then the optimal number of clusters is 9779 - 9776 = 3 cluster.

The results of this elbow rules are used as the input for non-hierarchical clustering method, in order to obtain results as table 8.

Table 8: Number Of Cases In Each Cluster Data

Cluster	1	1
	2	9.738
	3	40
Data Valid		9.779
Data Missing		0

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Based on the results in table 8, so the best data mining process to be done next is the data contained in cluster 2. So the end result of data is as shown in table 9.

Table 9: Information Web Log Data

	Raw data	Final data	% reduce of data
The number of web page	32.041	2.628	91.80
The number of unique users	72.622	9.738	86.59
Total record	1.673.978	101.272	93.95

From Table 9, we obtained information that from 1,673.798 recorded data (raw data), only 101.272 data that can be processed further. When it is viewed from the unique user side, it is only 9,738 users and 2,628 of web pages.

5. CONCLUSION

Based on the results of two level clustering method on web log data, it can be concluded that this method can improve the quality of data web log up to 90.78%. A conclusion can also be taken that of 1.673.798 raw recorded data, only 101.272 web log data can be processed further (6.05 %). From user point of view, from 72.622 users listed, only 9.738 users that can be processed in next stage or 13.41%. Similar thing happens for the amount of web page in where only 8.20% which will continue to be processed. Therefore, there is an increase of 90.78% on data quality based on reduced data.

REFRENCES:

[1] L. Haibin, K. Vlado, *Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future request, Data & Knowledge Engineering Journal, Vol. 61*, 2007, pp. 304-330.

[2] Yuhefizar, B. Santosa, I Ketut E.P, Y. K. Suprapto, *Combination of Hierarchical and Non-Hierarchical Cluster Method for Segmentation of Web Visitors, Telkomnika Journal, Vol. 11* No. 1, 2013, pp. 207-214.

[3] Pani S. K., Panigrahy L., Sankar V. H., Ratha B. K., Mandal A. K., Padhi S. K.., *Web Usage Mining: A Survey on Pattern Extraction from Web Logs, International Journal of Instrumentation, Control & Automation (IJICA), Volume1,* Issue 1, 2011, pp 15–23.

[4] Mamoun A, Awad, I. Khalil, *Prediction of User's Web-Browsing Behavior: Application of Markov Model, IEEE Transactions on System, Man, Anda Cybernetics, Vol. 42,* No. 04, 2012, pp 1131-1142.

[5] Bing Liu, *Web Data Mining :Exploring Hyperlinks, Contents, and Usage Data*, Chicago, Springer, 2007.

[6] Khasawneh N., Chan C-C., *Active User-Based and Ontology-Based Web Log Data Preprocessing for Web Usage Mining*, International Conference on Web Intelligence, IEEE/WIC/ACM, 2006, pp. 325 – 328.

[7] J. Srivastava, R. Cooley, M. Deshpande, P-N Tan, *Web Usage Mining: Discovery and Application of Usage Patterns from Web Data, SIGKDD Exploration, Vol. 1* Issue 2, 2000, pp. 1-12.

[8] Aye T.T., *Web Log Cleaning For Mining of Web Usage Patterns*. International Conference on Computer Research and Development (ICCRD). (page: 490-494 Year of Publication : 2011 ISBN: 978-1-61284-840-2).

[9] Chen H. W., Zong X, Wei L. C., Haw Y. J., *World Wide Web Usage Mining Systems and Technologies*, *Journal Systemic, Cybernetic and Informatics, Volume ,1* No.4, 2004, pp53–59.

[10] N. K. Tyagi, A.K. Solanki , S. Tyagi, *An Algorithmic Approach to Data Preprocessing in Web Usage Mining. International Journal of Information Technology and Knowledge Management, Volume 2*, No. 2, 2010, pp. 279-283.

[11] Raju G. T., Satyanarayana P.S., *Knowledge Discovery from Web Usage Data: Complete Preprocessing Methodology*, *IJCSNS International Journal of Computer Science and Network Security, Vol.8,* No.1, 2008, pp179–186.

[12] R. Cooley, B. Mobasher, J. Srivastava, *Data preparation for mining World Wide Web browsing patterns, Knowledge and Information Systems, Vol. 1*, No. 1, 1999, pp. 5-32.

[13] Youquan H., *Decentralized Association Rule Mining on Web Using Rough Set Theory, Journal of Communication and Computer, Volume 2*, No.7, 2005, ISSN1548-7709.

[14] G. Castellano, A. M. Fanelli, M. A. Torsello, *Log Data Preparation for Mining Web Usage Patterns, IADIS International*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Conference Applied Computing, pp 371-378,2007.

[15] Suneetha K. R., Krishnamoorthi D. R., *Identifying User Behavior by Analyzing Web Server Access Log File*, *IJCSNS International Journal of Computer Science and Network Security, Vol. 9,* No. 4, 2009.

[16] Wahab M. H. A., Mohd M. N. H., *Data Preprocessing on Web Server Logs for Generalized Association Rules Mining Algorithm*, *World Academy of Science, Engineering and Technology.*

[17] Hussain T, Asghar S, Masood N. *Web Usage Mining: A Survey on Preprocessing of Web Log File*. International Conference on Information and Emerging Technologies (ICIET), 2010, pp. 1 – 6.

[18] Lee C-H, Lo Y-L, Fu Y-H., *A Novel Prediction Model Based on Hierarchical Characteristic of Web site*. Expert Systems with Application, Vol. 38, 2011, pp. 3422 – 3430.

[19] Xu HJ, Liu H., Web User Clustering Analysis Based on KMeans Algorithm. International Conference on Information, Networking and Automation (ICINA) , 2010, pp. V2-6 – V2-9.

[20] Y-S. Lee, S-J. Yen, *Incremental and Interactive mining of web traversal patterns, Information Sciences Journal, Vol. 178*, 2008, pp. 287-306.