TELKOMNIKA, Vol.11, No.1, March 2013, pp. 207~214
ISSN: 1693-6930
accredited by DGHE (DIKTI), Decree No: 51/Dikti/Kep/2010

Combination of Cluster Method for Segmentation of Web Visitors

Yuhefizar[1], Budi Santosa[2], I Ketut Eddy P[3], Yoyon K. Suprapto[4]
[1]Information Technology Department, Politeknik Negeri Padang, Indonesia
[2]Industrial Engineering Department, Institut Teknologi Sepuluh Nopember (ITS), Surabaya
[3,4]Electrical Engineering Department, Institut Teknologi Sepuluh Nopember (ITS), Surabaya
yuhefizar@polinpdg.ac.id[1)],budi_s@ie.its.ac.id[2)] { ketut[3)],yoyonsuprapto[4)]}@ee.its.ac.id

Abstrak

Klasterisasi merupakan salah satu bagian penting dalam web usage mining untuk keperluan segmentasi pengunjung. Hal ini sangat berguna untuk keperluan personalisasi atau modifikasi web. Dalam paper ini, kami melakukan klasterisasi terhadap pengunjung web menggunakan kombinasi metoda klaster hirarki dan non-hirarki terhadap data web log. Metoda klaster hirarki digunakan dalam penentuan jumlah klaster dan non-hirarki digunakan dalam membentuk klaster. Tahapan analisis klaster didahului dengan pra-pengolahan data dan analisis Factor. Dengan pendekatan ini, pemilik web lebih efektif dalam menemukan pola akses pengunjung web dan memberikan pengetahuan baru dalam segmentasi pengunjung. Dari pengujian yang dilakukan terhadap data web log ITS, diperoleh 6 klaster pengunjung web dan klaster ke-3 mempunyai jumlah anggota terbesar. Hal ini menjadi masukan bagi pengelola web untuk memperhatikan pola perilaku anggota klaster ke-3 tersebut baik untuk keperluan personalisasi ataupun modifikasi web. Hal ini juga membuktikan kelayakan dan efisiensi dari penerapan metoda ini.

Kata kunci: *web usage mining, analisis klaster, personalisasi web, modifikasi web, web log*

Abstract

Clustering is one of the important part in web usage miningfor the purpose of segmenting visitors. This action is very important for web personalization orweb modification. In this paper, we perform clustering of the web visitors using a combination of methods of hierarchical and non-hierarchical clustering toward web log data. Hierarchical clustering method used to determine the number of clusters, and non-hierarchical clustering method is used in forming clusters. The stages of cluster analysis are preceded by pre-processing the data and factor analysis. With this approach, the owner of the web is more effective at finding access patterns of web visitors and can have new knowledge about visitors' segmentation. From the test applied on ITS's web log data, 6 clusters of web visitors are resulted. Among the 6 cluster, cluster 3 has the biggest number of members. This information can be useful for web management to pay attention on members' behavioral patterns of the 3rd cluster's either to make personalization or modification on the web. The test results show the feasibility and efficiency of application of this method.

Keywords: web usage mining, cluster analysis, web personalization, web modification, web logs

1. Introduction

The Internet has become a huge information source [1] and an important media in the distribution of current information. This is an integral part of one internet service, namely the World Wide Web (WWW) that is capable of disseminating information in text, image, video, or voice and multimedia. The survey results conducted by Netcraft, in July 2012 states that there are 665,916,461 active sites, and according to internet world stats, in December 2011 there are 2.267.233.742 internet users in the world. This means that the interaction between Internet users with web sites is very high and web servers record every activity of the visitor is in the form of files (web log). Until now, a web log has become the most important part in Web Usage Mining (WUM) to gather the web visitor data, especially in finding patterns of visitors' access, prediction of visitors' behavior [2],[3], to create a user profiles [4],[5].

WUM or web log mining [6] is one category in the field of web mining [7], which is the mining conducted on the web based on web log data. Specifically, by [8], states that WUM is the application of data mining techniques to discover the interaction between visitors of a website

Received December 29, 2012; Revised January 9, 2013; Accepted February 7, 2013

through web log data. The mining of web logs is useful for a variety of fields, including for web personalization [9] and web modification [10].

Techniques on WUM is including statistical analysis[11], association rules [12],[13], sequential patterns [14],[15], classification [16],[17] and clustering [18-20]. Clustering is one of the important topics in WUM for visitor segmentation based on access patterns on the web or frequency of visits. by [21], use belief function method to perform the clustering on web log data. They divide web visitors into different groups and find a common access pattern for each group member. However, this approach still requires identify sessions that are less efficient on the pre-processing stage. By [22], conduct the clustering of web visitors with the K-Means method and they only prove that the method of K-Means clustering can be used to web log data without validation of its cluster result.

According to [23], clustering on web sessions includes three stages, namely pre-processing, measurement on the similarity and the application of cluster algorithms. In this research, we perform clustering based on the visiting frequency of visitor on the sites in the given period of time regardless of the web session so it is more efficient at the pre-processing stage and then we perform clustering using a combination of hierarchical and non-hierarchical cluster methods.

This paper is organized as follow: in chapter 1 that explains the background of the research and also the related research, chapter 2 discusses about stages of the research as well as the method used, chapter 3 is about the result and analysis, and chapter 4 is the conclusion of the research.

2. Research Method

Stages of this research in general are shown in Figure 1.

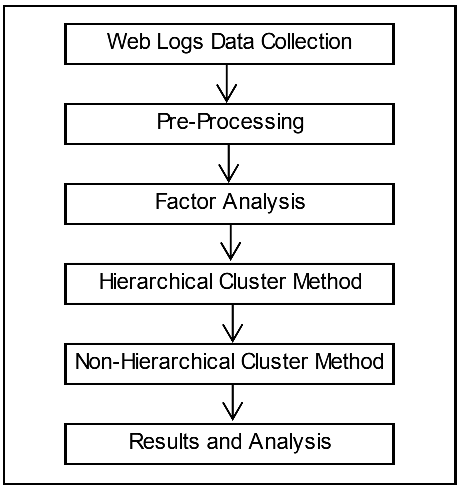

Figure 1. Stages of Research

2.1. Dataset

The dataset used in this research are web log data from web of Tenth of November Institute of Technology Surabaya, with the web address is www.its.ac.id and the period of data collection is from 3 to 16 July 2012.Web log file format used in this research is the Common Log Format (CLF) [24], which is the standard format used by the web server when creating a log. Each line of CFLs consists of host/IP Address, identification, authuser, date and time, method, request, status, and bytes as shown in table 1.

From the first line of Table 1, we obtained information that the visitor with IP address 66.249.69.xxx have accessed a web page index.php on July 15, 2012 at 06:45:13 with a status code of 200 and 15319 file size and so on. This is the kind of information which is to be researched to get web visitor segmentation.

Table1. Common log format

Host/IP Address	Ident, authuser	Date & time	Method	Request	Status	Bytes
66.249.69.xxx	--	15/Jul/2012:06:45:13	GET	/index.php	200	15319
114.79.57.xxx	--	15/Jul/2012:19:08:48	GET	/info.php	200	15582
206.53.148.xxx	--	15/Jul/2012:19:08:50	GET	/media.jpg	200	1324
96.47.225.xxx	--	15/Jul/2012:19:20:20	POST	/berita.php	200	30462
114.79.16.xxx	--	15/Jul/2012:20:00:01	GET	/favicon.ico	200	3798

2.1.1. Pre-Processing

At this stage, we perform the process of cleaning/filtering from web log data from items that are not needed (irrelevant data). Filtering has been done based on:

(i) **The file extension**, the accepted file extensions are .html, .php, .jsp, asp and other extensions that refer directly to a web page. Item data with file extensions such as .jpg, .gif, .ico, .bmp, .cgi, .swf, .css, .txt does not describe the behavior of web visitors so that the data item is removed [25].

(ii) **Access Method**. Only access that uses the GET method can indicate the behavior of web visitors. Item data with other access methods, such as HEAD and POST are also removed [25].

(iii) **The response code from the webserver**. Web server response with the code of 200 indicates an access request to a web page is granted and displayed by the web server. Therefore, the data item with a code other than 200 is removed [26].

(iv) **The frequency of visitor access**. Only visitors with access >10 were used in this research, as it is assumed that visitors with access<10 can not properly describe the behaviour of visitors.

The final result of pre-processing stage in the form of a matrix vector is as follow [22]:

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

where *m* is the number of web visitors (data), *n* is the number of web pages (variable), and *X* is a vector of observations. Implementation of matrix vector in equation (1) about the web visitor behavior data based on the frequency of visits to the web page is shown in Table 2.

Table 2. Matrix vector

User	Web page											
	p1	*p2*	*p3*	*p4*	*p5*	*p6*	*p7*	*p8*	*p9*	*p10*	…	*pn*
u1	6	9	0	0	0	0	0	0	5	20	…	X_{1n}
u2	0	0	0	35	0	35	0	0	0	0	…	X_{2n}
u3	0	11	0	0	0	0	0	0	14	9	…	X_{3n}
u4	0	1	4	3	2	3	0	4	4	1	…	X_{4n}
u5	0	84	0	0	0	0	0	0	0	0	…	X_{5n}
u6	5	5	5	0	1	5	4	2	6	0	…	X_{6n}
u7	1	37	0	0	3	0	0	1	9	1	…	X_{7n}
u8	2	21	3	0	4	0	2	1	7	0	…	X_{8n}
:	:	:	:	:	:	:	:	:	:	:	:	:
um	X_{m1}	X_{m2}	X_{m3}	X_{m4}	X_{m5}	X_{m6}	X_{m7}	X_{m8}	X_{m9}	X_{m10}	…	X_{mn}

With *p1, p2, p3, pn* are the variable for a web page, for example, *p1* is the web page with the name of index.php. *u1, u2, u3, um* are the variable for the visitors of the web, for example *u1* is a web visitor's with IP address, 72.233.234.xxx. From Table 2, it can be concluded that the visitors with variables *u1* have accessed the web page *p1* 6 times, web page *p2* 9 times and so on

After the pre-processing of the dataset, 165 web visitor data were acquired with 57 variables (accessed web page). This data in the form of this matrix vector that was processed further.

2.1.2. Factor Analysis

The next stage is to conduct a factor analysis on the data resulted from the pre-processing stage. Factor analysis is a multivariate method that is used to describe the pattern of relationships between variables in order to find independent variables that affect the objects called by a factor. In this case, factor analysis aims to reduce the variables into several sets of indicators called factors, with no loss of meaningful information from the initial variable.

The first stage in factor analysis is the process of testing the adequacy of the data and the identification of correlations between variables with Measure of Sampling Adequacy (*MSA*) method in equation (2), Kaiser-Meyer-Olkin (*KMO*) in equation (3) and Bartlett's Test in equation (4) [27].

$$\mathrm{MSA_i} = \frac{\sum_{j=1}^{p} r_{ij}^2}{\sum_{j=1}^{p} r_{ij}^2 + \sum_{j=1}^{p} a_{ij}^2} \quad (2)$$

$$KMO = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2 + \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}^2} \quad (3)$$

where:

i= 1, 2, 3, ..., p dan j = 1, 2, 3, ..., p
r_{ij}= Coefficient of correlation between variables i and j
a_{ij} = Partial correlation coefficient between variables i and j

$$BartlettsTest = -\ln\,|R|\left[n - 1 - \frac{2p+5}{6}\right] \quad (4)$$

where:

$|R|$ = Value of determinan
n = Number of data
p = Number of variabel

Based on this method, a group of data is said to meet the sufficiency of the data and the correlation assumptions when the value of the *MSA, KMO* is greater than 0.5 and a significance value of Bartlett test <0.05. Therefore, variables with *MSA*<0.5 were excluded from the analysis. Output of the analysis in form of factor scores will be used in the cluster analysis. Table 3 shows the test results using *KMO*, *Bartlett's* and *MSA* methods.

Table 3. Results of the testing with KMO, Bartlett and MSA methods

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.757
Bartlett's of Sphericity	Approc. Chi-Square	9872.112
	Df	1596
	Sig	0

As shown in Table 3, the value of KMO and Bartlett's Test is 0.757 with significance value is 0.0. This means that the variable and the data can be received and analyzed further because the value of KMO and Bartlett's Test received is > 0.5 and significance value <0.05. Variables with MSA <0.5 were excluded in this research. Table 4 shows the variables with MSA <0.5.

After testing the adequacy of the data, then a factor analysis was performed with results as shown in figure 2.As shown in Figure 2 that there are 14 factors formed (eigenvalues ≥ 1) of 57 baseline variables. With the distribution of the variable and the percentage of variable ability explained by factor shown in table 5 and table 6.

The last step in factor analysis is to make factors score, this is a score for factors that are formed to replace the value of the original variable by naming variable *f1* to factor 1, *f2* to factor 2, and so on. The results from the factor scores operation are used for cluster analysis.

Table 4. Variables with *MSA*<0.5

Variables	*MSA Values*
p50	0.415
P7	0.394
P36	0.416
P1	0.401
P33	0.471
P30	0.491

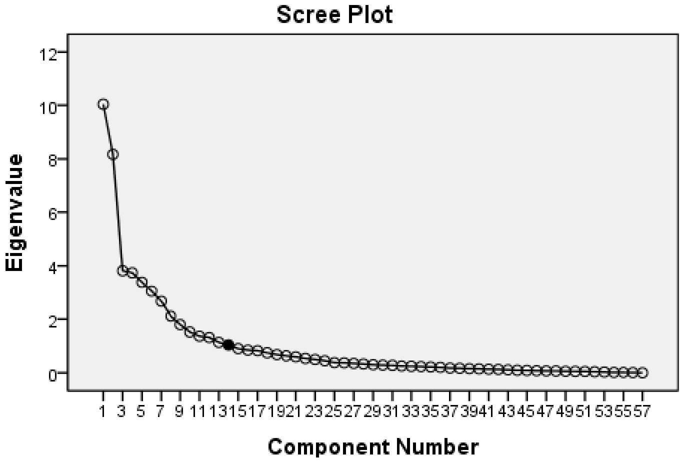

Figure 2. Scree plot factorization results

Table 5. Distribution and percentage of variable ability explained by resulted factor

Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
p10(67.8%)	*p40(81%)*	*p4(89.7%)*	*p19(79.3%)*	*p32(69.4%)*	*p11(72.9%)*	*p45(82.7%)*
p12(88.4%)	*p42(76.9%)*	*p6(96.8%)*	*p31(83.7%)*	*p34(82.6%)*	*p13(79.2%)*	*p51(80%)*
p15(87.8%)	*p44(73.6%)*	*p27(98.4%)*	*p38(87.7%)*	*p35(78.3%)*	*p14(86.1%)*	*p61(91.1%)*
p16(78.2%)	*p55(71.8%)*	*p28(98.6%)*	*p57(92.2%)*	*p37(87.8%)*	*p26(82.6%)*	*p63(34%)*
p17(92.7%)	*p59(75%)*					
p20(82.8%)	*p60(81.1%)*					
p21(87.7%)	*p62(66.4)*					
p22(90.4%)						
p24(92.4%)						
p25(94.8%)						
p49(81.1%)						

Table 6. Distribution and percentage of variable ability explained by resulted factor (continue)

Factor 8	Factor 9	Factor 10	Factor 11	Factor 12	Factor 13	Factor 14
p3(67.3%)	*p46(84.5%)*	*p23(74%)*	*p52(95.4%)*	*p47(89.9%)*	*p39(63.8%)*	*p2(71.8%)*
p5(66.5%)	*p48(63.5%)*	*p29(53.7%)*	*p58(96%)*	*p56(89.2%)*		*p9(55.8%)*
p8(69%)	*p53(70.8%)*	*p41(71.6%)*				
p18(76%)	*p54(55.2%)*	*p43(77.1%)*				

2.1.3. Cluster Analysis

Cluster analysis is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar to each other than to those in other clusters. This is non-parametric techniques which is very much applicable in the real world. Cluster analysis in this study was carried out by combining the hierarchical clustering method and the non-hierarchical clustering method. Result of the factor analysis in the form of factor scores were used as input to the cluster analysis.

2.1.3.1. Hierarchical Cluster

The first phase of the hierarchical cluster is calculating the distance between objects with euclidean distance method and cluster formation using the single linkage method. Based on the results of the agglomeration schedule from this method, the number of clusters based on the rules of the elbow were determined, as shown in Table 7.

Table 7. Agglomeration schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
:	:	:	:	:	:	:
158	3	24	61.638	157	0	159
159	3	10	**100.274**	158	0	160
160	3	6	115.043	159	0	161
161	3	4	116.709	160	0	162
162	3	78	133.669	161	0	163
163	1	3	147.534	0	162	164
164	1	2	175.524	163	0	0

Table 7 shows a difference in co-efficient in where co-efficient in stage 159 is bigger than the other. Thus, based on elbow rule, with the amount of data as 165, 165 – 159 = 6 (resulted 6 clusters). These result are used as input for the non-hierarchy cluster analysis.

2.1.3.2. Non-Hierarchy Cluster

Non-Hierarchical Cluster is used to determine web's visitor segmentation. In this case, K-Means method [22] was used with the following algorithm:

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

3. Results and Analysis

Based on the implementation of Non-Hierarchy Cluster method with 6 cluster of web visitor, membership of every cluster was gotten, as shown in Table 8.

Table 8. The number of clusters' members

Cluster	Member
1	2
2	1
3	143
4	13
5	3
6	3

Valid data: 165

Table 8 informs the grouping of 165 web's visitor with cluster 1 consists of two members, cluster 2 with one, cluster 3 with one hundred forty three, cluster 4 with thirteen, and cluster 5 and 6 with three members each. The detail information can be seen in Table 9.

It can be concluded from Table 9 that web visitors (*u*1, *u*2, *u*3…*u*165) within the same cluster have the same access or visiting pattern toward ITS web page so that this information can be used as an input for the web personalization and modification, including cluster 3 which has the most member.

The last part of cluster analysis is to produce the final cluster centers. As informed by Table 10, the amounts of clusters produced are six and each cluster has its own characteristic which is different from one another. This information can be seen from the value of the final cluster center of each variable in where the positive sign (+) represents the values which are above average and the negative sign are the value below average. Here, the value of *f1* has a positively big value in cluster 1 but has negative value in other clusters. It means that the web page in factor 1 is visited by more members in cluster 1 comparing to the other clusters. Based on the clusters, it can be concluded that cluster 1 is the visitors who dominantly access the web page within *f1* and *f14*, cluster 2 consists of visitors who dominantly access the web page within *f*3, and so on.

Table 9. Cluster membership

Cluster	Member
1	u1, u3
2	u2
3	u5, u7, u9, u10, u12, u13, u14, u15, u17, u18, u19, u20, u21, u22, u23, u26, u27, u29, u32, u33, u34, u35, u36, u37, u38, u39, u40, u41, u42, u43, u44, u45. U46, u47, u48, u49, u51, u52, u53, u54, u55, u57, u58, u59, u60, u61, u62, u63, u64, u65, u67, u70, u71, u72, u73, u74, u75, u76, u77, u79, u81, u82, u83, u84, u85, u86, u87, u88, u89, 90, u91, u92, u93, u94, u95, u96, u97, u98, u99, u100, u101, u102, u103, u104, u105, u106, u107, u108, u109, u110, u111, u113, u114, u115, u116,u117, u118, u119, 120, u121, u122, u123, u124, u125, u126, u127, u128, u129, u130, u131, u132, u133, u134, u135, u136, u137, u138, u139, u140, u141, u142, u143, u145, u146, u147, u148, u149, u150, u151, u152, u153, u154, u155, u156, u157, u158, u159, u160, u161, u162, u163, u164, 165
4	u4, u8, u11, u24, u25, u28, u50, u56, u66, u68, u69, u112, u144
5	u30, u78, 80
6	u6, u16, u31

Table 10. The final clusters centre

Var	Cluster					
	1	2	3	4	5	6
f1	7.89195	-.15046	-.09949	-.04319	-.14271	-.13891
f2	-.14383	-.35827	.03727	-.16429	-.72473	-.12447
f3	.03924	12.59114	-.08314	-.04008	-.16055	.07413
f4	-.02047	-.19888	-.13353	-.03349	.13025	6.45976
f5	.02678	-.39458	-.04784	.66457	-.42887	-.05686
f6	.00977	-1.06264	-.00937	.17408	-.60885	.64886
f7	-.08475	-.07753	-.15246	1.83025	-.47691	-.10470
f8	.18684	-.69466	-.00275	-.07509	.68261	-.11938
f8	-.19577	-.15362	-.03678	.34198	.26607	.18699
f10	-.07120	-.07587	.01548	-.11539	-.25520	.09032
f11	-.37024	-.01147	-.15697	1.95009	-.22823	-.48951
f12	-.08222	-.00952	-.13323	.18111	5.60284	.02113
f13	.20360	.04535	-.04079	.40478	.14716	-.10767
f14	1.08553	.11134	-.04146	.23921	.01811	.16078

4. Conclusion

Based on the application of combined method of hierarchy and non-hierarchy cluster toward the web log data, it can be summed up that this method can give new information about a web visitors' pattern or behavior so that the information can be used for web personalization and web modification. From the test applied on ITS's web log data, 6 clusters of web visitors are resulted. Among the 6 cluster, cluster 3 has the biggest number of members (143 members). This information can be useful for web management to improve the service on the web page which is frequently visited or accessed by member of 3rd cluster, especially if the management wants to do the web personalization and web modification.

References

[1] Yohanes BW, Handoko, Wardana HK. Focused Crawler Optimization Using Genetic Algorithm.*TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2011; 9(3): 403 - 410.

[2] Fong ACM, Baoyao Z, Hui SC, Hong GY, Do TA. Web Content Recomender System Based On Consumer Behavior Modelling. *IEEE Transactional on Consumer Electropnics*. 2011; 57(2): 962 – 969.

[3] Awad MA, Khalil I. Prediction of User's Web-Browsing Behaviour: Application of Markov Model. *IEEE transaction on Systems, Man, And Cybernetics, Part B: Cybernetics.* 2012; 42(4): 1131 – 1142.

[4] Nasraoui O, Soliman M, Saka E, Badia A, Germain R. A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites. *IEEE Transaction on Knowledge and Data Engineering*. 2008; 20(2): 202 – 215.

[5] Godoy D, Amandi A. User Profiling for Web Page Filtering. *IEEE Internet Computing*. 2005; 9(3): 56–64.

[6] Wang Y-T, Lee AJT. Mining Web Navigation Patterns With a Path Traversal Graph. *Experts System with Application.* 2011; 38(6): 7112 – 7122.

[7] 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)*. Karachi. 2010: 1–6.

[8] 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. Washington 2006: 325–328.
[9] Chang CC, Chen P-L, Chiu F-R, Chen Y-K. Application of Neural Networks and Kano's Method to Content Recommendation in Web Personalization. *Journal Expert Systems with Application*. 2009; 36 (3); 5310 – 5316.
[10] Kumar R. Mining Web Logs: Applications and Challenges. *KDD'09 Proceedings of the 15th ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining*. New York. 2009.
[11] Srivastava J, Cooley R, Deshpande M, Tan P.-N. Web Usage Mining: Discovery and Applications of Usage Patterns From Web Data. *SIGKDD Explorations*. 2000; 1(2): 12-23.
[12] Lee CL, Lee S. Interpreting The Web-Mining Results by Cognitive Map and Association Rule Approach*. Information Processing & Management*. 2011; 47(4); 482 – 490.
[13] Nagi M, ElSheikh A, Sleiman I, Peng P, Rifaie M, Kianmehr K,Karampelas P, Ridley M, Rokne J,Alhajj R. *Association Rules Mining Based Approach for Web Usage Mining*. IEEE International Confrence on Information Reuse and Integration (IRI). Las Vegas, NV. 2011: 166–171.
[14] Lee Y-S, Yen S-J. Incremental and Interactive Mining of Web Traversal Patterns. *Information Sciences*. 2008; 178(2): 287-306.
[15] [15] Wu H-Y, Zhu J-J, Zhang X-Y. The Explore of the Web-Based Learning Environment Base in Web Sequential Pattern Mining. *International Conference on Computational Intelligence and Software Engineering (CISE)*. Wuhan. 2009: 1–6.
[16] Chen C-M, Lee H-M, Chang Y-J.*Two* Novel Feature Selection Approaches For Web Page Classification*. Expert Systems with Applications.* 2009; 36(1): 260 – 272.
[17] Yu JX, Yuming O, Zhang C, Zhang S. Identifying Interesting Visitors Throught Web Log Classification*. IEEE Intelligent Systems*. 2005; 20(3): 55 – 59.
[18] Sudhamathy G, Venkateswaran JC. Web Log Clustering Approaches – A Survey. *International Journal on Computer Science and Engineering (IJCSE)*. 2011; 3(7): 2896–1903.
[19] Shi P. An Efficient Approach for Clustering Web Access Patterns from Web Logs. *International Journal of Advanced Science and Technology*. 2009; 5: 1–14.
[20] Martiana E, Rosyid N, Agusetia U. Mesin Pencari Dokumen dengan Pengklasteran Secara Otomatis. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2010; 8(1): 41 - 48.
[21] Xie Y, Phoha VV. *Web User Clustering From Access Log Using Belief Function*. Proceedings of the ACM K-CAP'OI. First International Conference on Knowledge Capture. Victoria. 2001: 202-208.
[22] Xu HJ, Liu H. *Web User Clustering Analysis Based on KMeans Algorithm*. International Conference on Information, Networking and Automation (ICINA). Kunming. 2010; 2: V2-6 – V2-9.
[23] Chaofeng L. *Research on Web Session Clustering.* Journal of Software. Academy Publisher. 2009; 4(5): 460–468.
[24] Tanasa D, Trousse B. Advanced Data Preprocessing for Intensitas Web Usage Mining*. IEEE Intelligent System*. 2004; 19(2); 59 – 65.
[25] Lee C-H, Lo Y-L, Fu Y-H. A Novel Prediction Model Based on Hierarchical Characteristic of Web Site. *Expert Systems with Application.* 2011; 38 : 3422–3430.
[26] Liu B. Web Data Mining : Exploring Hyperlinks, Contents, and Usage Data. Berlin: Springer. 2007.
[27] Niu J, He Y, Li M, Zhang X, Chao C, Zhang B. *A Comparative Study on Application of Data Mining Technique in Human Shape Clustering: Principal Component Analysis VS. Factor Analysis.* IEEE Conference on ICIEA. 2010; 2014 – 2018.