Clustering of User Behaviour based on Web Log data using Improved K-Means Clustering Algorithm

S.Padmaja [#1], Dr.Ananthi Sheshasaayee [#2]

[#1] Research Scholar, Research and Development Centre,
Bharathiar University, Coimbatore, India.

[#2] Associate Professor and Head, Department of Computer Science,
Quaid-e-Millath Government College for women, Chennai , India.

[#1] spadmaja.research@gmail.com

[#2] ananthiseshu@gmail.com

Abstract: **The proposed work does an improved K-means clustering algorithm for identifying internet user behaviour. Web data analysis includes the transformation and interpretation of web log data find out the information, patterns and knowledge discovery. The efficiency of the algorithm is analyzed by considering certain parameters. The parameters are date, time, S_id, CS_method, C_IP, User_agent and time taken. The research done by using more than 2 years of real data set collected from two different group of institutions web server .this dataset provides a better analysis of Log data to identify internet user behaviour.**

Keywords: User behaviour, k-Means, Web log Data, behaviour analysis, clustering

I. INTRODUCTION

Clustering is an unsupervised learning, it finds natural grouping of instances given unlabelled data. Clustering is the process of grouping related objects into classes. Different applications of clustering are

- Market research done in economic source.
- Document classification.
- Cluster of web log data to discover groups of similar access pattern.
- Pattern Recognition.
- Spatial data analysis.
- Image processing etc…

Three different categories of web mining are web content, web structure and web usage mining. Web Usage Mining addresses the problem of extracting behavioural patterns from one or more web access logs [1,2]. the entire process can be divided into three major steps. The first step, pre-processing, is the task of accurately identifying pages accessed by web visitors. This is a very difficult task because of page caching and accesses by web crawlers. The second step, pattern discovery, involves applications of data mining algorithms to the pre-processed data to discover patterns. The last step, pattern analysis, involves analysis of patterns discovered to judge their interestingness [2]. Web usage mining is one of the applications of data mining techniques to discover user access patterns from web data. Web usage data captures web-browsing behaviour of users from a web site. Web usage mining can be classified according to kinds of usage data examined. In our context, the usage data is Access logs on server side, which keeps information about user navigation. Our work is mainly focused on web usage mining including discovery of user navigation patterns from access logs. In the next section, the structure of access logs as our main data source is introduced [3].

II. RELATED WORK

A. Analysis of User's Behaviours and Growth Factors

Dong Woo Kim, Tae Gu Kang, Guozhong Li and Seong Taek Park have discussed, systematically analyzes behaviour of internet shopping mall users through big data analysis and proposes a strategic operation plan using this analysis. Analyzing customers' behaviours using various methodologies of analysis with the secured data and providing them what they want, timely, through that would allow efficient operation of shopping malls [4].

B. Extracting Users'Navigational Behavior

Maryam Jafar1, Farzad SoleymaniSabzchi and Shahram Jamali have explained existing WUM techniques and it is shown that how WUM can be applied to Web server logs. and focused on methods that can be used for the task of pattern extraction from Web log files. After discovering patterns, the result will

be used for pattern analysis phase. Analyzing of the Web users‟ navigational patterns can help understand the user behaviours and Web structure, therefore the design of Web components and Web applications will be improved [5].

C. Efficiency Calculation of Mined Web Navigational Patterns

L. K. Joshila Grace* and V. Maheswari presented the parameters like frequency, Utility, downloads, book mark, selection are considered for each web page and efficiency for the web navigation pattern is found. The work is done by using real data set extracted from an e-commerce web site and with synthetic data set. This provides a better analysis of the web site. The result provided by this proposed work can be used or various application in developing the web site contents [6].

D. Access Patterns in Web Log Data

Mohammed Hamed Ahmed Elhiber and Ajith Abraham discussed like this in their paper the process of Web Usage Mining consisting steps: Data Collection, Pre-processing, Pattern Discovery and Pattern Analysis. It has also presented several approaches such as statistical analysis; clustering, association rules and sequential pattern are being used to discover patterns in web usage mining. The pattern analysis phase means applying data mining techniques such as SQL and OLAP on the pattern discovery data to filter insignificant information to obtain the valuable information[7].

E. Clustering of Navigation Patterns using BW Theorem

V. Chitraa and Antony Selvadoss Thanamanihave have taken different parameters for comparison were Rand Index, Sum of Squared Errors, F-measure. The method was implemented after data cleaning in all the three data sets, session construction step. Clustering was done twice in all three data sets algorithms with their proposed clustering algorithm, without selecting features and after selecting features. When applied in full data set with irrelevant features, it was observed that the clustering results are poor. The performance was increased after relevant features were selected. The result of the optimized clustering proves its significance and there is an increase in similarity of intra clustering and dissimilarity in inter clustering than the existing methods [8].

F. Algorithm for Tracing Visitors' On-Line Behaviors

S. Umamaheswari and S. K. Srivatsa have focused on the newly proposed algorithm is Visitors' Online Behaviour (VOB) which identifies user behaviour, creates user cluster and page cluster, and tells the most popular web page and least popular web page. This paper brings into discussion about the basic concepts of web mining, web usage mining, general data pre-processing, how to pre-process the web data, what are the various existing pre-processing techniques and the proposed VOB algorithm[9].

III. WEB LOG FILE

There are three types of log files that can be used for Web usage mining. Log files are stored on the server side, on the client side and on the proxy servers. By having more than one place for storing the information of navigation patterns of the users makes the mining process more difficult. Really reliable results could be obtained only if one has data from all three types of log file. The reason for this is that the server side does not contain records of those Web page accesses that are cached on the proxy servers or on the client side. Besides the log file on the server, that on the proxy server provides additional information. However, the page requests stored in the client side are missing. Yet, it is problematic to collect all the information from the client side. Thus, most of the algorithms work based only the server side data. Some commonly used data mining algorithms for Web usage mining are association rule mining, sequence mining and clustering [10,11].

IV. PREPROCESSING

Web usage mining has three different types of activity [12,13]:

1. Preprocessing activities that make a review to the web log data prior to processing.

2. Discovery Activity Pattern (Pattern Mining), which spent most of all mining activities because these activities do a search to find hidden patterns in the data log.

3. Pattern Analysis (Analyzing Pattern) which is a process to study and conduct an analysis of the results obtained from the search behaviour patterns.

In the Preprocessing, data log / record of the use of web pages are usually not in a format that could be used by the mining applications. If data will be used in mining applications, the data format must be changed into another form that can be used by the mining application. Some of the steps that are part of the preprocessing is Cleansing, User Identification, Session Identification, Path Completion, and Formatting[.13]

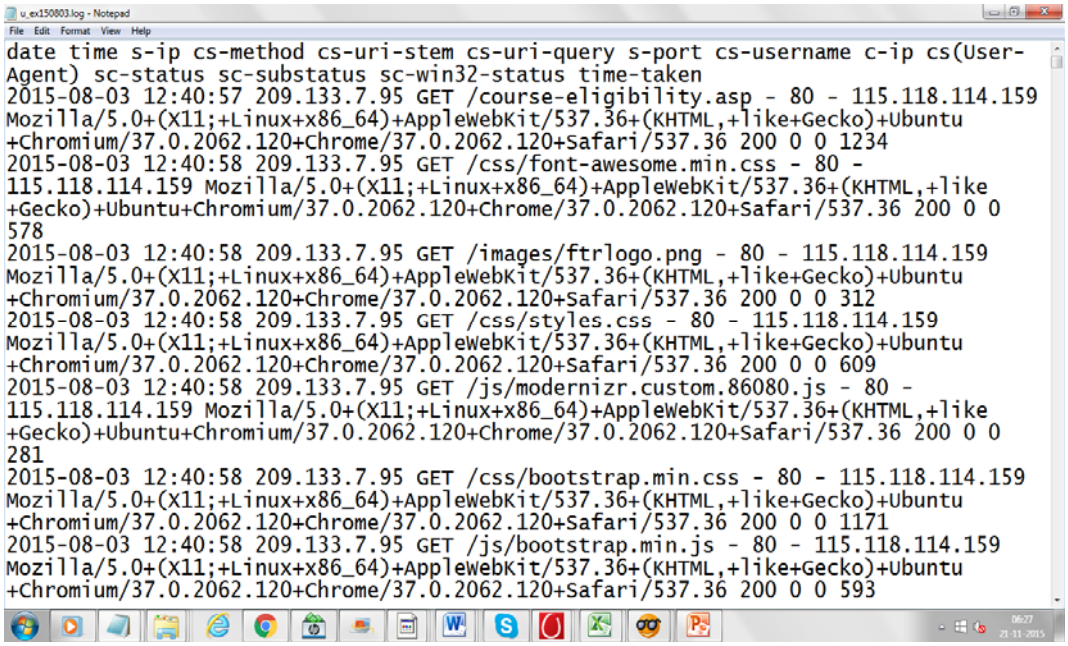

Fig 1. Raw Web Server Log Data

E5 /css/styles.css

	A	B	C	D	E	F	G
1	date	time	s-ip	cs-method	cs-uri-stem	c-ip	cs(User-Agent)
2	03-08-2015	12:40:57	209.133.7.95	GET	/course-eligibility.asp	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
3	03-08-2015	12:40:58	209.133.7.95	GET	/css/font-awesome.min.css	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
4	03-08-2015	12:40:58	209.133.7.95	GET	/images/ftrlogo.png	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
5	03-08-2015	12:40:58	209.133.7.95	GET	/css/styles.css	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
6	03-08-2015	12:40:58	209.133.7.95	GET	/js/modernizr.custom.86080.js	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
7	03-08-2015	12:40:58	209.133.7.95	GET	/css/bootstrap.min.css	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
8	03-08-2015	12:40:58	209.133.7.95	GET	/js/bootstrap.min.js	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
9	03-08-2015	12:40:59	209.133.7.95	GET	/images/logo2.png	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
10	03-08-2015	12:40:59	209.133.7.95	GET	/images/bg-links.png	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
11	03-08-2015	12:40:59	209.133.7.95	GET	/js/jquery-2.1.4.min.js	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
12	03-08-2015	12:40:59	209.133.7.95	GET	/images/fiversns.png	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
13	03-08-2015	12:40:59	209.133.7.95	GET	/images/bg-black.png	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
14	03-08-2015	12:40:59	209.133.7.95	GET	/images/1.jpg	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
15	03-08-2015	12:40:59	209.133.7.95	GET	/images/mapped_popup.png	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
16	03-08-2015	12:41:00	209.133.7.95	GET	/images/favicon.ico	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
17	03-08-2015	12:43:23	209.133.7.95	GET	/images/calendar.gif	66.102.6.172	Mozilla/5.0+(Linux;+Android+4.2.1;+en-us;+Nexus+5+Build/JOP40D)+AppleWebKit/535.19+(K
18	03-08-2015	12:46:07	209.133.7.95	GET	/marine-science-engineering.asp	108.171.159.41	Mozilla/5.0+(Windows+NT+6.1;+WOW64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Chron
19	03-08-2015	12:46:07	209.133.7.95	GET	/images/logo2.png	108.171.159.41	Mozilla/4.0+(compatible;)
20	03-08-2015	12:46:08	209.133.7.95	GET	/css/styles.css	108.171.159.41	Mozilla/4.0+(compatible;)
21	03-08-2015	12:46:09	209.133.7.95	GET	/js/jquery-2.1.4.min.js	108.171.159.41	Mozilla/4.0+(compatible;)
22	03-08-2015	12:46:09	209.133.7.95	GET	/images/9.jpg	108.171.159.41	Mozilla/5.0+(Windows+NT+6.1;+WOW64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Chron

Fig 2. Web Server Log Data before pre-processing

Algorithm: Data Preprocessing

1. Start
2. To collect raw log data generated from Web server.
3. To clean the data by filtering unwanted data like audio and picture file with extension .jpg, .gif,.jpeg,.css etc..
4. To identify the number of users from preprocessed web log data.
5. Identify the number of sessions.
6. Identify time taken in each session.
7. Goto step 3 until it completes all raw server data
8. End

E1 cs-uri-stem

	A	B	C	D	E	F	G
1	date	time	s-ip	cs-method	cs-uri-stem	c-ip	cs(User-Agent)
2	03-08-2015	12:40:57	209.133.7.95	GET	/course-eligibility.asp	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
3	03-08-2015	12:40:58	209.133.7.95	GET	/images/ftrlogo.png	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
4	03-08-2015	12:40:58	209.133.7.95	GET	/js/modernizr.custom.86080.js	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
5	03-08-2015	12:40:58	209.133.7.95	GET	/js/bootstrap.min.js	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
6	03-08-2015	12:40:59	209.133.7.95	GET	/images/logo2.png	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
7	03-08-2015	12:40:59	209.133.7.95	GET	/images/bg-links.png	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
8	03-08-2015	12:40:59	209.133.7.95	GET	/js/jquery-2.1.4.min.js	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
9	03-08-2015	12:40:59	209.133.7.95	GET	/images/fiversns.png	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
10	03-08-2015	12:40:59	209.133.7.95	GET	/images/bg-black.png	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
11	03-08-2015	12:40:59	209.133.7.95	GET	/images/mapped_popup.png	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
12	03-08-2015	12:41:00	209.133.7.95	GET	/images/favicon.ico	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
13	03-08-2015	12:46:07	209.133.7.95	GET	/marine-science-engineering.asp	108.171.159.41	Mozilla/5.0+(Windows+NT+6.1;+WOW64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Chron
14	03-08-2015	12:46:07	209.133.7.95	GET	/images/logo2.png	108.171.159.41	Mozilla/4.0+(compatible;)
15	03-08-2015	12:46:09	209.133.7.95	GET	/js/jquery-2.1.4.min.js	108.171.159.41	Mozilla/4.0+(compatible;)
16	03-08-2015	12:46:12	209.133.7.95	GET	/css/bootstrap.min.css	108.171.159.41	Mozilla/4.0+(compatible;)
17	03-08-2015	12:46:15	209.133.7.95	GET	/images/favicon.ico	108.171.159.41	Mozilla/4.0+(compatible;)
18	03-08-2015	12:46:16	209.133.7.95	GET	/bsc-nautical-science.asp	108.171.159.41	Mozilla/5.0+(Windows+NT+6.1;+WOW64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Chron
19	03-08-2015	12:46:57	209.133.7.95	GET	/college-admission-procedure.asp	106.215.173.185	Mozilla/5.0+(Linux;+U;+Android+4.1.2;+en-gb;+GT-S6312+Build/JZO54K)+AppleWebKit/534.3(
20	03-08-2015	12:47:07	209.133.7.95	GET	/course-eligibility.asp	115.118.114.159	Mozilla/5.0+(X11;+Linux+x86_64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Ubuntu+Chron
21	03-08-2015	12:49:29	209.133.7.95	GET	/part-time-courses-offered.asp	117.217.88.95	Mozilla/5.0+(Windows+NT+10.0;+WOW64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Chro
22	03-08-2015	12:49:30	209.133.7.95	GET	/js/modernizr.custom.86080.js	117.217.88.95	Mozilla/5.0+(Windows+NT+10.0;+WOW64)+AppleWebKit/537.36+(KHTML,+like+Gecko)+Chro

Fig 3. Preprocessed Web Server Log Data

V. IMPROVED K-MEANS CLUSTERING ALGORITHM

Clustering is the process of organizing data into classes in the form of high intra-class similarity and low inter-class similarity. In K-Means clustering algorithm randomly choose K data items from X as initial centroids. It assigns each data point to the cluster which has the closest centroid. Calculates new cluster centroids and have to continue the process until the convergence criteria is met. The improved K-Means clustering algorithm minimizes the total of squares of the gaps.

Algorithm: Improved K-Means clustering

1. Start
2. Input preprocessed web log data.
3. Generate clusters. Number of cluster generated is set of input points x,x(k).
4. Initialize k cluster centroids c. place centroids ci ….ck at random location.
   μi= some value ,i=1,...,k
5. Grouping each object to its closest cluster center.
   $ci=\{j:d(xj,\mu i)\leq d(xj,\mu l),l\neq i,j=1,...,n\}$
   $\mu i=1|ci|\sum j\in cixj,\forall i$
6. Repeat the step 4 and 5 until no change in the cluster group.
7. End

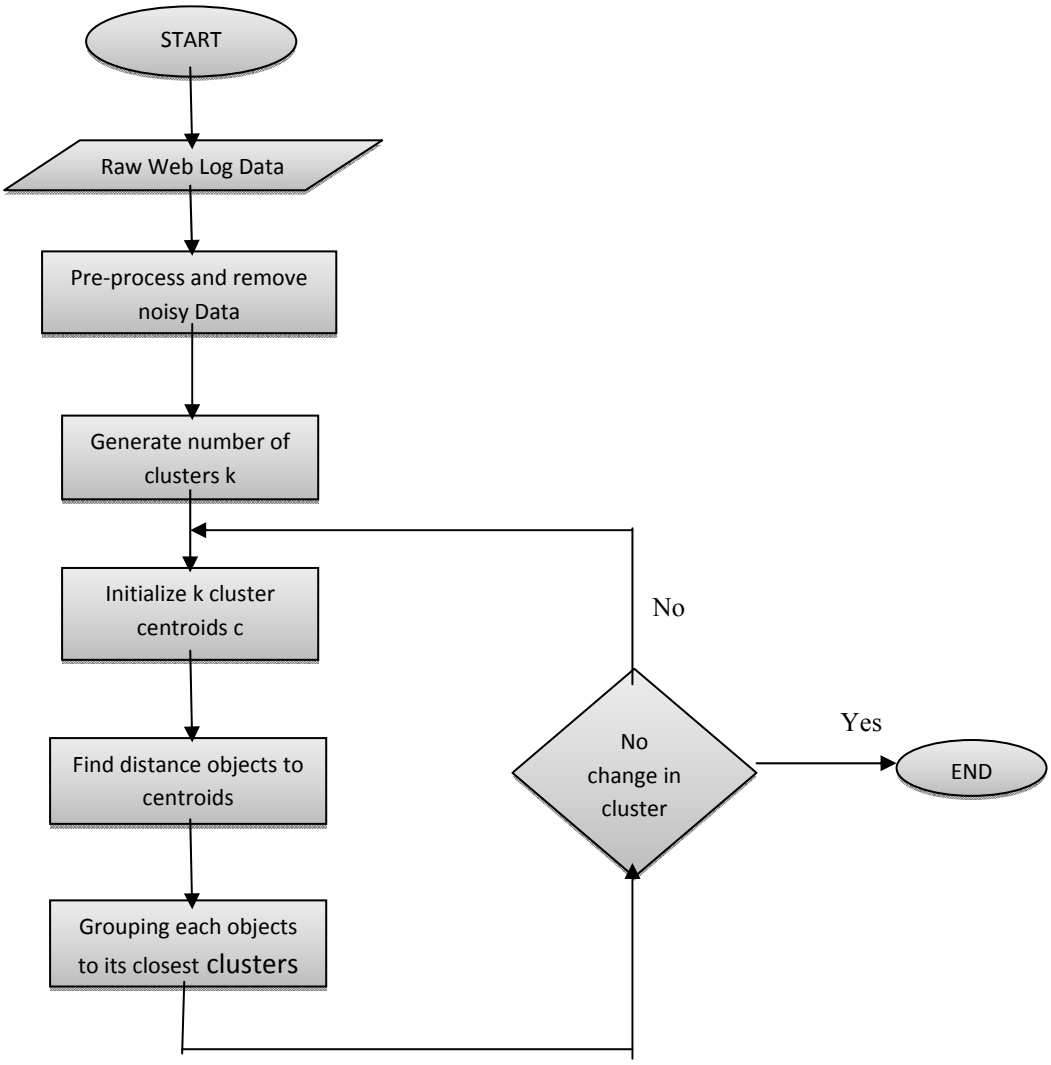

Fig 4 A Combined Flowchart for Data Pre-processing and Enhanced K-Means clustering Algorithm

VI. CONCLUSION

In this article clustering is done group web users into same cluster based on users browsing behaviour during a specific time interval. The process cannot be completed until Pre-process is done properly. Preprocessed web log data is clustered using improved K-Means clustering algorithm to identify internet users behaviour.

REFERENCES

[1] F. Masseglia, P. Poncelet, and M. Teisseire. Using data mining techniques on web access logs to dynamically improve hypertext structure. In ACM SigWeb Letters,8(3): 13-19, 1999.
[2] Sana Siddiqui and Imran Qadri. International Journal of Advanced Research in Computer Science and Software Engineering. Volume 4, Issue 6, June 2014, pp 794-802.
[3] Murat Ali Bayir, Ismail H. Toroslu and Ahmet Cosar. Performance Comparison of Pattern Discovery Methods on Web Log Data.IEEE conference, 2006. Pp 445-551.
[4] Dong Woo Kim, Tae Gu Kang, Guozhong Li**3** and Seong Taek Park**, "** "Analysis of User's Behaviors and Growth Factors of Shopping Mall using Bigdata", Indian Journal of Science and Technology, Vol 8(25), IPL0444, October 2015, PP 1-8.
[5] Maryam Jafari, Farzad SoleymaniSabzchi,* and Shahram Jamali, "Extracting Users'Navigational Behavior", Journal of Computer Sciences and Applications, 2013, Vol. 1, No. 3, PP 39-45.
[6] L. K. Joshila Grace* and V. Maheswari, "Efficiency Calculation of Mined Web Navigational Patterns", Indian Journal of Science and Technology, Vol 7(9), 1350–1354, September 2014. PP 1-5.
[7] Mohammed Hamed Ahmed Elhiber and Ajith Abraham, "Access Patterns in Web Log Data: A Review", Journal of Network and Innovative Computing ,, Volume 1 (2013) pp. 348-355.
[8] V. Chitraa and Antony Selvadoss Thanamani, "Clustering of Navigation Patterns using Bolzwano_Weierstrass Theorem", Indian Journal of Science and Technology, Vol 8(12), 69283, June 2015. PP 1-9.
[9] S. Umamaheswari and S.K.Srivasta, "Algorithm for Tracing Visitors' On-Line Behaviors for Effective Web Usage Mining", International Journal of Computer Application, Volume 87 – No.3, February 2014. Pp 22-28.
[10] R. Cooley, B. Mobasher, and J. Srivastava, "Data preparation for mining world wide web browsing patterns," Knowledge and Information Systems, Vol. 1, No. 1, pp. 5-32, 1999.

[11] Renáta Iváncsy and István Vajk, Frequent Pattern Mining in Web Log Data, Acta Polytechnica Hungarica Vol. 3, No. 1, 2006,pp 77 – 90.
[12] Han, J. and Kamber, M. 2000. Data Mining: Concepts and Techniques.Morgan Kaufmann Publishers.
[13] Ford Lumban Gaol, Exploring The Pattern of Habits of Users Using WebLog Squential Pattern, IEEE, Second International Conference on Advances in Computing, Control, and Telecommunication Technologies,2010, pp 161-163.
[14] K. Selvakumar, L. Sai Ramesh and A. Kannan, "Enhanced K-Means Clustering Algorithm for Evolving User Groups", Indian Journal of Science and Technology, Vol 8(24), IPL0371, September 2015. PP 1-8.