Mining Web Log using Fuzzy C – Mean for Navigational Pattern Prediction

Ankita Dixit
M.Tech Scholar
Oriental Institute of Science and Technology
Bhopal (M.P)

Meena Lakshmi
Project Guide
Oriental Institute of Science and Technology
Bhopal (M.P)

ABSTRACT
The web is an important source of information retrieval in the present scenario and users belonging to various backgrounds access the internet. Internet is in reach of everyone and its users are increasing day by day. So it is very important in this competitive world here in terms of e- commerce, the companies should know the needs and demands of user. The usage information about the users is recorded in logs of web server which are called as web logs. These web log files are analyzed for the purpose of pattern extraction which is useful, this technique of data mining commonly known as web usage mining. Accuracy in the results of mining of logs of web and efficient prediction of patterns of users navigating online are very necessary as these results help in websites tune up for the users. Actually web log at their first place are not meant for this process of web log mining, they are just stored at server for record. The process of Web log mining initiates with cleaning then data preparation which is termed as data pre-processing, it extracts some hidden knowledge which cannot be found out by using any other conventional methods. For good result better quality of input is required so more emphasis is on cleaning and preprocessing of data. Knowledge obtained is then mined and then it is analyzed which predicts out the user's online behavior and activity pattern. In my research work I have implemented the same phases of web usage mining i.e. pre-processing of web log and Fuzzy C - Mean algorithm is applied on the knowledge gained and then analyzed for users navigation results.

General Terms
WUM, FCM

Keywords
Web Mining, Pattern Analysis, Navigation Pattern

1. INTRODUCTION
The growth rate of the Web, propagation of e-commerce, web services, and web based information systems, the volumes of mouse clicked stream and user data gathered by Web based organizations from their daily operations has reached huge proportions. For the time, the considerable increase in the number of websites presents exigent task for webmasters to systematize the contents of the websites to furnish to the needs of users. Analysis and Modeling of the users behavior of web navigation helps in grasping the demand of online users. Subsequent to that, the analyzed outcome can be seen as information to be used in intellectual online applications, refining web site maps, web based personalization system and improving searching accuracy when in quest of information.

Web mining an application of the data mining techniques which are used for knowledge extraction from data on web, of that at least some of structure or data of user which is the usage log data used in the process of mining. There are 3 extensive classes of Web mining [12]: Web content mining: process of discovering accommodating information from text, image, audio or video data in the web. Web structure mining functions on the hyperlinks. Web Usage Mining works on the web logs.

1.1 The Usage Mining On Web
Web usage mining or WUM an application of data mining techniques for the usage pattern discovery from the web data, in order to be familiar with and serve improved needs of Web-based applications [10]. In the same paper, WUM is defined in 3 distinct phases: pre-processing, pattern discovery, and pattern analysis. I think it is an excellent approach to define the usage mining procedure. It also clarified the sub direction of research of WUM, which assists the researchers to focus on each individual process with diverse applications and techniques. With the half of the figure shown in Figure 1 presenting high- level WUM process, which is presented in [9-11], reader may understand the architecture of the Web Usage Mining easily. I will give a detailed introduction as follows, encompassing these three phase processing.

1.1.1 Data Pre-Processing For Mining
WUM which is an application of data mining techniques is used extract usage logs (web data) of bulky Web data repositories. In this stage of WUM the generated results after processing can be used in the task of tuning websites, tuning web server and in navigating throughout a website [13]. However, prior to applying the data mining algorithm, we should carry out data preparation to alter the raw data into the abstracted data necessary for the further process. The data can be collected at the server-side, client-side, proxy servers, or obtained from database. For every type of collection of data, the difference is not only the location, but also the existing data type, the section of population from which the data or information was collected and the implementation method [10]. The information sources obtained for mining purpose comprises of Web usage logs, descriptions of web page, topology of website, registries of users, and questionnaire [15].

The task of preprocessing logs is typically a multifaceted and time demanding. It comprises of 4 tasks:

- The data cleaning.
- Identification of the user's session and its reconstruction.
- The information retrieval of information regarding the content and structure of page.
- The data formatting.

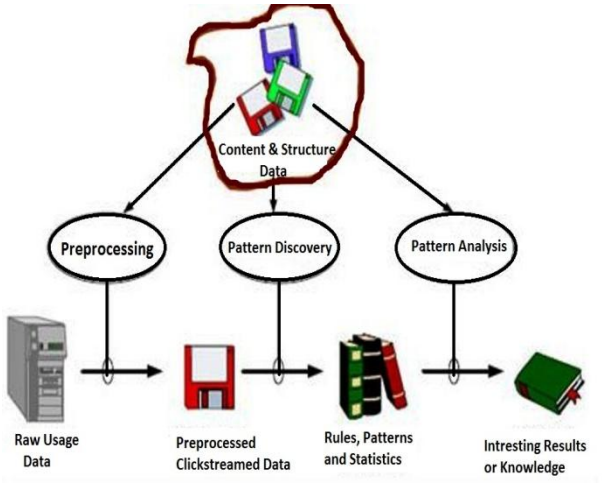


Fig 1: Basic of Web Usage Mining Process

1.1.2 Pattern Discovery

This is the key component of the Web mining. Pattern discovery congregates the algorithm and procedures from numerous research areas like data mining, statistics of machine learning, and recognition of patterns. Various techniques are used for this purpose, but here we have used clustering which I have discussed later.

1.1.3 Pattern Analysis

Pattern Analysis is a final stage of the whole WUM. The target of this procedure is to eradicate the irrelative rules or patterns and to extract the interesting rules or patterns from the pattern discovery process output. The Web mining algorithms output is normally not in the form apposite for direct human being utilization, and consequently needed to be convert to a layout which can be assimilate easily. This can be done with the help of some analysis methodologies and tools. We have used Weka tool whose predefined classes have been imported in java.

1.2 Fuzzy C-Means Clustering

Clustering is a technique for grouping together users or the data items or the pages having the similar characteristics. Clustering of user info or pages can be facilitating the growth and implementation of future marketing approaches [9]. Clustering of users will assist in discovering the user grouping, which has analogous navigation pattern. It is very helpful for assembling user demographics to achieve market dissection in ecommerce applications or deliver personalized web content of the individual or each user. Clustering of pages is valuable for search engines and web service providers, later it could be used to discover the groups of pages having related content. Fuzzy C-Means is an algorithm of clustering which uses the concept of categorizing the data in two or more clusters that belongs to the same group as generated in Fuzzy Logics.

1.3 Data Sources

Web Usage Mining applications are based on data collected from three main sources [15]: web servers, proxy servers, and web clients.

- **The Server Side:** Web servers are confidently the richest and the most communal source of data. They can collect large amounts of information in their log files and in the log files of the databases they use. These logs usually contain basic information e.g. name, remote host IP, request date and request time, and request line is taken as it has came from the client. When exploiting log information from web servers, the key concern is the users sessions identification. Apart from web logs, behaviour of user can also be traced at the side of server with the help of TCP/IP packets. Even in this case the users sessions identification is still an issue, but packet usage provides some advantages [17]. In fact: (i)On real time data is collected.(ii)Information coming from various servers can be simply combined together to form a unique log.(iii)Special buttons (e.g. the stop button) usage can be easily detected so to collect information usually unavailable in log files. Packets are hardly ever used in practice due of rise scalability issue on web servers with huge traffic [17].
- **The Proxy Side:** Many internet service providers (ISPs) give to their customer Proxy Server services to improve navigation speed through caching. In many respects, collecting navigation data at the proxy level is basically the same as collecting data at the server level. The main difference in this case is that proxy servers collect data of groups of users accessing huge groups of web servers
- **The Client Side:** Client side usage data can be easily tracked by using JavaScript, java applets or even on some advanced browsers. These techniques shun the problems such as identification of users session and caching problem (use of the back button).In addition, they provide detailed information about actual user behaviours [17]. Conversely, these approaches rely deeply on the cooperation by the user and rise many issues concerning with the privacy laws, which are quite severe.

2. RELATED WORK

In this paper [4] they have discussed about precise web mining of logs and an efficient online navigational pattern prediction which is in disputably vital for alteration up websites and subsequently helping in visitors' preservation. Similar to any other data mining task, it starts with cleaning of data and its preparation and it ends up determining some hidden knowledge which cannot be pull out using conventional methods. In order for this process to yield good results it has to rely on some good quality input data. For that reason, additional focus in this process is on cleaning of data and it's pre-processing. While on the other side, the confront facing online prediction is scalability. As a result any improvement in the efficiency of online prediction solutions is more than necessary. As retort to the abovementioned concerns they proposed an upgrading to the web mining process of log and to the pattern prediction of navigation online. There contribution contains 3 different components. First, they proposed a sophisticated time-out based heuristic for session identification. Second, they suggested the practice of a precise an algorithm based on density for navigational pattern detection. Finally, they recommended a new loom for efficient online prediction. The demeanor experiments reveal the applicability and efficiency of the projected approach. In this paper [5] they have discussed about log file analysis of web which began with the reason to offer to Web site administrators a method to ensure adequate bandwidth and ability of server for the organization. This analysis field made huge advances with the fleeting of time, and now companies

look for ways to use Web log files to get information concerning visitor profiles and activities of buyer. The investigation of Web log may tender advices regarding aim proved way to advance the offer, information concerning problems occurred to the users, and even about effort for the safety of the site. Traces about hacker attacks or heavy use in exacting intervals of the time may be actually useful to systematize the server and regulate the Website. From the users point, web is a growing collection of bulky amount of information, usually a great portion of time is essential to look for and find the suitable information. Personalization is a option for the achievement of the developing of a web infrastructure. A client or a visitor who finds effortlessly what he was probing for is a client or a visitor that will return. For this reason, Web sites are shaped and modified to made contents more easily reachable, using profiles found to make recommendations or to aim users with ad hoc advertising.

In this paper [6] they discussed about WUM an application of data mining techniques for gaining knowledge for serving better the requirements of web based applications. WUM analysis is done by applying techniques of pattern reorganization on the logs of web data. Pattern recognition is defined as the act of captivating in raw data and making an action based on the category of the pattern. WUM is divided into 3 parts: Preprocessing, Pattern discovery and Pattern analysis. Further, this paper intended with experimental work in which web log data is used. They have taken the log data of the "NASA" from NASA web server which is analyzed by "Web Log Explorer". Web Log Explorer is a web usage mining tool which plays the vital role to carry out this work.

In this paper[7] they converse the most topical loom to log of web data illustration aspire to incarcerate the navigational patterns of users with revere to the on the whole structure of the website. One such depiction is tree structured log files which is the main focus of this entire work. Most accessible technique for analyzing such kind of data are based on the use of frequent mining sub tree techniques to pull out frequent user activity and navigational paths. In this paper they evaluated the use of extra typical data mining techniques facilitating by a recently proposed structure preserving flat data illustration for ordered data. The originally proposed agenda was attuned to better suit the task of log mining. Experimental estimate is executed on 2datasets of real world web log and assessments are prepared with an existing status of the art classifier for tree structured data. The results show the huge potential of the process in facilitating the application of a broad range of data mining or analysis techniques to tree structured web log data.

In this paper [8] In this title, they confer about the speedy expansion of www in its quantity of passage and the range and complication of web sites. In this paper, a new loom is offered based on hybrid clustering methods for WUM. The WUM process has 3 steps: preprocessing of data, data mining and analysis of result. Firstly, they gave a concise depiction of the WUM process and data on web, then the appearance of the preprocessing step and the data warehouse were engaged. The amalgam clustering technique based on FCM clustering are worn for analyzing and the Web logs taken from the real world servers of web. The outcome obtained after applying these technique and the equivalent interpretation are also offered. Furthermore, this paper also described WUM with regards to cloud which is cloud mining.

3. PROPOSED SCHEME

In the proposed methodology first web log data of NASA is taken as input and then pre-processing is done on input dataset. Apply FCM clustering on the preprocessed dataset and then from the clustered values patterns are generated. Lastly useful knowledge is extracted from the generated patterns.

3.1 Proposed Algorithm

STEP 1: Read the web log file.

STEP 2: Preprocessing

[1]Select required attribute from log file like IP Address/ URL, Date and Time, Request Type of User request, Protocol, Port Number and Page Number & remove other attributes if present.

[2]Remove irrelevant entries like robot request.

[3]From cleaned log file identify unique users according to IP Address and unique web pages.

[4]Session identification is performed by taking threshold time value of 30 min and 10 min.

STEP 3: Clustering by Fuzzy C – Means.

The Algorithm consist of the following few steps along with the minimization of the objective function.

[1] First of all the Dataset whose clustering is done is initialized with the objective function as $S = [u_{ij}]$, it is a matrix which contains the values to be clustered of m rows and n columns.

[2] After each kth step compute the centroid of the matrix or vector matrix of the dataset denoted as C,

$C^{(k)} = [c_j]$, and contains the vector Matrix S,

[3] Compute objective function and minimize the effect of the objective function using,

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} . xi}{\sum_{i=1}^{N} u_{ij}^{m}}$$

[4] Update each value of S[k] with next S [k+1]

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

[5] To check the computer value of S[k] and S[k+1], means $\|S^{k+1} - S^{k}\| < \epsilon$, stop the process otherwise go to Step –[2].

STEP 4: Prediction: On various criteria prediction is done by analyzing the discovered patterns from the NASA dataset.

3.2 Proposed Flowchart

The flow chart of the algorithm of the proposed work is as shown in the Figure

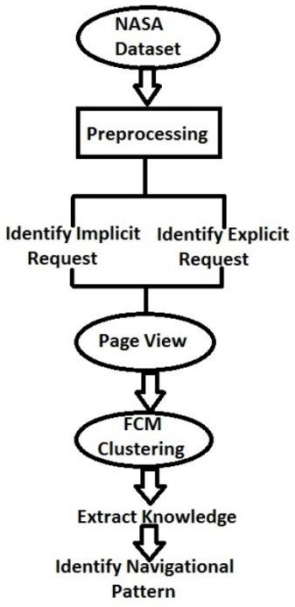


Fig 2: Flow Chart of Proposed Algorithm

4. WEB SERVER LOG

For the research work we have taken web server log or dataset of NASA available freely on the internet for use. The Log files contain 23 days data from 1st Aug 1995 to 23rd Aug 1995 but we have taken 1 week data for research purpose dated from 01/Aug/1995:00:00:01 -0400 to 08/Aug/1995:22:56:20 -0400.The dataset contain 32598 entries.

TABLE I. Sample of NASA Dataset

"uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] ""GET /images/USA-logosmall.gif HTTP/1.0"" 304 0"
"ix-esc-ca2-07.ix.netcom.com - - [01/Aug/1995:00:00:09 -0400] ""GET /images/launch-logo.gif HTTP/1.0"" 200 1713"
"133.43.96.45 - - [01/Aug/1995:00:00:16 -0400] ""GET /shuttle/missions/sts-69/mission-sts-69.html HTTP/1.0"" 200 10566"
"kgtyk4.kj.yamagata-u.ac.jp - - [01/Aug/1995:00:00:17 -0400] ""GET / HTTP/1.0"" 200 7280"

The data set contains following entries:

- User Name: It is normally IP address or URL; it identifies the user who visited website.
- Time Stamp: It is the time spent by user on each page having format day/month/year: Hour: Min: Sec.
- Request Type: It is the method that is used by the user to send the request to the server and it can be either GET or POST method.
- HTTP Reply Code: it tells the status codes of hypertext transfer protocol.
- Bytes in Reply: number of bytes consumed in user's request.

4.1 Status Code Used in Dataset

The table below shows the list of the status code of HTTP used in the dataset.

TABLE II. HTTP Status Code

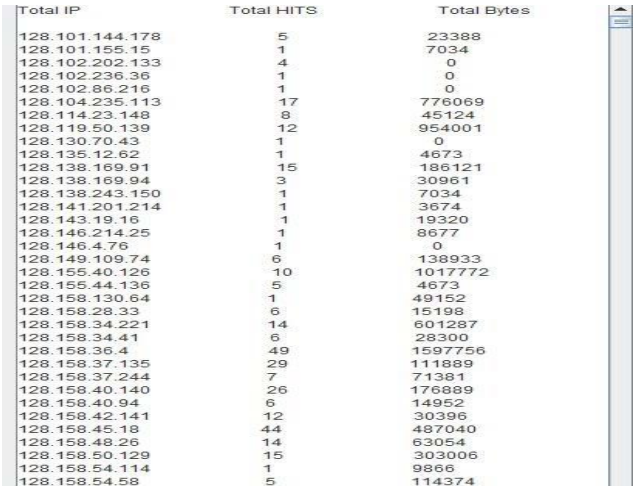
Status Code	Description
200	OK
201	Created
202	Accepted
203	Non-Authoritative Information
204	No Content
302	Found
304	Not Modified
400	Bad Request
401	Unauthorized
403	Forbidden
404	Not Found
415	Unsupported Media Type
500	Internal Server Error
501	Not Implemented
502	Bad Gateway
503	Service Unavailable
504	Gateway Timeout
505	HTTP Version Not Supported

5. EXPERIMENTAL RESULTS

In this research work after pre-processing following results we have got.

- Total Number of Request Made : 32008
- Total Number of Unique Users : 2592

The figure below shows the list of uniquely identified user, total number of hits and total number of bytes consumed in the requests.



Total IP	Total HITS	Total Bytes
128.101.144.178	5	23388
128.101.155.15	1	7034
128.102.202.133	4	0
128.102.236.36	1	0
128.102.86.216	1	0
128.104.235.113	17	776069
128.114.23.148	8	45124
128.119.50.139	12	954001
128.130.70.43	1	0
128.135.12.62	1	4673
128.138.169.91	15	186121
128.138.169.94	3	30961
128.138.243.150	1	7034
128.141.201.214	1	3674
128.143.19.16	1	19320
128.146.214.25	1	8677
128.146.4.76	1	0
128.149.109.74	6	138933
128.155.40.126	10	1017772
128.155.44.136	5	4673
128.158.130.64	1	49152
128.158.28.33	6	15198
128.158.34.221	14	601287
128.158.34.41	6	28300
128.158.36.4	49	1597756
128.158.37.135	29	111889
128.158.37.244	7	71381
128.158.40.140	26	176889
128.158.40.94	6	14952
128.158.42.141	12	30396
128.158.45.18	44	487040
128.158.48.26	14	63054
128.158.50.129	15	303006
128.158.54.114	1	9866
128.158.54.58	5	114374

Fig 3: Result generating the list of total number of users

5.1 HTTP Status Code

The table below shows the HTTP response code along with the total number of visitors making the requests.

TABLE III. List of HTTP Request Made By Users In Their Visit

Response Code	No. of Visitors
200	29071
304	2349

404	168
403	0
501	0
400	0
Other	420

The figure below shows the graphical representation of this.

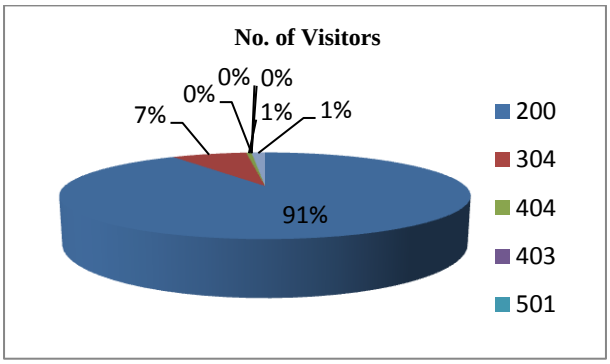


Fig 4: Result generating the list of total number of users

5.2 Accessed File Type

The table below shows the various types of file accessed by the users on their visit in the total duration of one week.

TABLE IV. List of File Type Accessed by the Users

File Type	Total Users
TXT	413
GIF	19901
JPG	560
HTML	6795
OTHERS	25213

The figure below shows the graphical representation of the different file type accessed within one week duration

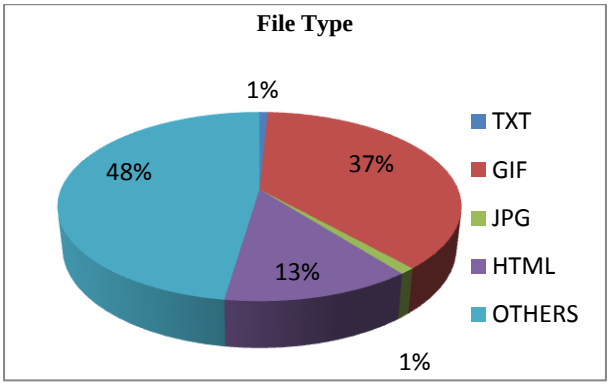


Fig 5: Percentage of different types of file accessed by the users

5.3 Clustering Results

The figure below shows the clustering of attributes of dataset after the pre-processing. Three Cluster 1, 2, 3 are taken. The dataset is passed to FCM and plotting is done on the basis of some attributes to other.

- In the first 3 box each of the attributes "User", "File Access Response Code" and "Access Bytes" are clustered individually according to the serial no.
- In the next 3 boxes clustering is done according to "User" and "File Access Response Code" in fourth box. In fifth box is plotted on the basis of "User" and "Access Bytes" it shows how many values in the dataset is access by the individual user and last box shows the plotting of clusters between the "File Access Response Code" and "Access Bytes".

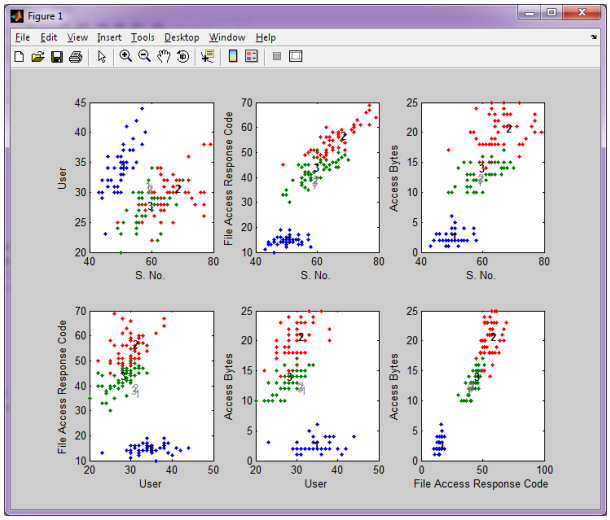


Fig 6: Percentage of different types of file accessed by the users

5.4 Weekly Error Analysis

The table below shows the weekly error report of request made by the users.

TABLE V. Weekly Error report

S. NO	Error	Hit
1	404 Not Found	168
2	403 Forbidden	0
3	503 Service Unavailable	195
Total		**363**

The figure below shows the graphical representation of the distribution of various protocols which results in some error.

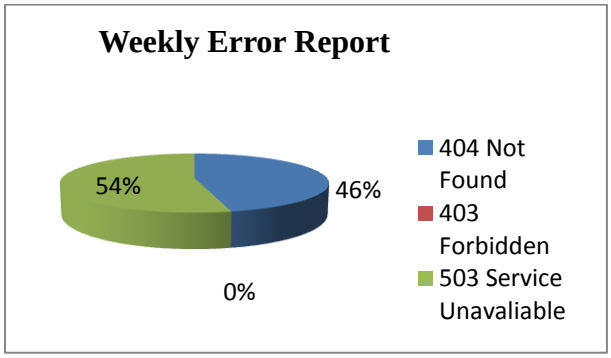


Fig 7: Percentage of different Protocols causing the Error

5.5 Weekly Accessed File Ratio Session Wise

The figure below shows the ratio of file accessed by users in total sessions in a day for duration of one week.

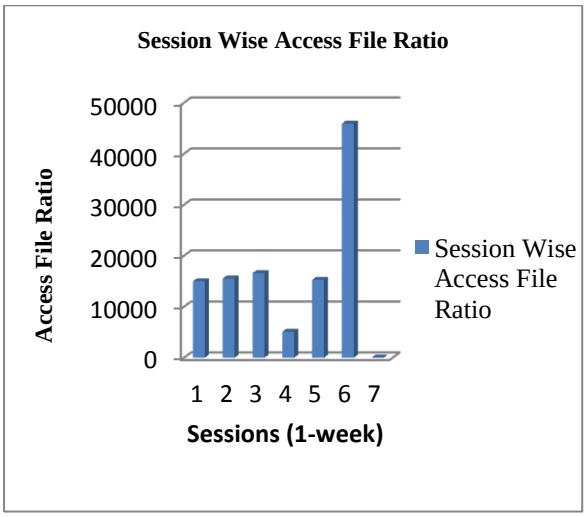


Fig 8: Session Wise Access File Ratio for a week

6. CONCLUSION

The methodology implemented here for the generation of online navigational patterns is proposed in which preprocessing is done and later applying Fuzzy C- Means clustering algorithm to cluster the patterns generated. The proposed methodology implemented here provides the patterns about user's navigation on various parameters which are beneficial for the company owner or organizations for tuning up of their websites according to requirements of the users.

The result analysis shows that the session calculation done by our proposed scheme are better than existing work and also has less false positive values. The methodology provides more generation of pattern from the web log database and true positive value which makes the performance of the proposed methodology better than the existing methodology.

7. FUTURE WORK

Although the technique implemented over here is efficient and predicts the navigational pattern of the users efficiently but there is always room for some improvements so further enhancements can be done in the field of Patterns generation. Since the patterns generated here are based on Clustering of the Dataset some other method can be used in future work.

8. ACKNOWLEDGMENTS

I am grateful for the valuable guidance and support of my Project Guide Mrs. Meena Lakshmi without whom all my effort would have been directionless and fruitless. I sincerely thank her, for assisting me in completing the work.

9. REFERENCES

[1] Bowman Abedelghani Guerbas, Omar Addam "Effective web log mining and online navigational pattern predictions", IEEE transactions on parallel and distributed systems, vol. 23, no. 10, October 2012.

[2] Yuefeng Lia, Ning Zhong "Web mining model and its applications for information gathering", Knowledge-Based Systems 17 (2004) 207–217

[3] Magdalini Eirinaki and Michalis Vazirgiannis "Web Site Personalization Based on LinkAnalysis and Navigational Patterns", ACM Trans. Intern. Tech. 7, 4, Article 21 (October 2007)

[4] Fedja Hadzic, Michael Hecker, "Alternative Approach to Tree Structured Web Log Representation and Mining", 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent

[5] Philipp Singer "T Understanding, Leveraging and Improving Human Navigation on the Web", International World Wide Web Conference Committee (IW3C2)

[6] Arlind Kopliku, Karen Pinel-Sauvagnat, and Mohand Boughanem "Aggregated Search: A New Information Retrieval Paradigm", ACM Comput. Surv. 46, 3, Article 41 (January 2014).

[7] Nanhay Singh, Achin Jain and Ram Shringar Raw "Comparison analysis of web usage mining using pattern recognition techniques", International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 4, Issue 1, 2013, pp1-8

[8] Arvind K. Sharma and P.C. Gupta "Analysis of web server log files to increase the effectiveness of the website using web mining tool" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.4, July 2013.

[9] V.Chitraa, Antony Selvadoss Thanamani "Web Log Data Analysis by Enhanced Fuzzy C Means Clustering", International Journal on Computational Sciences & Applications (IJCSA) Vol.4, No.2, April 2014.

[10] Maratea, A. and Petrosino, A., "An Heuristic Approach to Page Recommendation in Web Usage Mining", Ninth International Conference on Intelligent Systems Design and Applications, pp. 1043-1048, 2009.

[11] Maristella Agosti and Giorgio Maria Di Nunzio "Web Log Mining: A Study of User Sessions", http://www.theeuropeanlibrary.org/

[12] Surbhi Anand, Rinkle Rani Aggrawal "An Efficient Algorithm for Data Cleaning of Log File using File Extensions", International Journal of Computer Applications (0975 – 888) Volume 48– No.8, June 2012.

[13] Subhagata Chattopadhyay, Dilip Kumar Pratihar and Sanjib Chandra De Sarkar. "A Comparative Study Of Fuzzy C-Means Algorithm And Entropy-Based Fuzzy Clustering Algorithms", Computing and Informatics, Vol. 30, 2011, 701–720

[14] T. Velmurugan and T. Santhanam. "Implementation Of Fuzzy C-Means Clustering Algorithm For Arbitrary Data Points ",International Conference on Systemics, Cybernetics and Informatics

[15] G. Castellano, A. M. Fanelli and M. A. Torsello. "Log Data Preparation For Mining Web Usage Patterns", IADIS International Conference Applied Computing 2007

[16] K.Suresh, R.Madana Mohana and A.Rama MohanReddy. "Improved FCM algorithm for Clustering on Web Usage Mining ", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 1, January 2011

[17] Jian Pei, Jiawei Han, Behzad Mortazavi-asl, and Hua Zhu. "Mining access patterns efficiently from web logs", In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pages 396–407, 2000. 143

[18] M.Rathamani, Dr. P.Sivaprakasam, "Cloud Mining: Web usage mining and user behavior analysis using fuzzy C-means clustering", IOSR Journal of Computer Engineering (IOSRJCE) Volume 7, Issue 2 (Nov-Dec. 2012), PP 09-15

[19] M Ming-Syan Chen, Jong Soo Park "Efficient Data Mining for Path Traversal Patterns", International Journal of Computer Trends and Technology-