ISSN:2320-0790

An International Journal of Advanced Computer Technology

Implementation of Web Page Suggestion in Web mining

Miss. Aparna N. Gupta[1], Mrs. Aarti M. Karandikar[2]

[1] PG Student (CSE), Dept. of CSE, RCOEM Nagpur, Maharashtra India.
[2] Assistant Professor, Dept. of CSE, RCOEM Nagpur, Maharashtra, India.

Abstract: Web is a treasure of information and data, where enormous amount of data is and searching the worthwhile data from the web is a difficult job; therefore the web mining algorithm is employed to recognize the pattern and information from the data. This paper includes Mining Algorithms for evaluation and the implementation of frequent pattern analysis. We implemented pre-processing of web log(stastistical) data and then performed FCM clustering for taking similar interest that is overlapped clustering which allows one data object to belong to two or more clusters. Finally we suggest mostly viewed URLs related to user.

Keywords: Web Mining, Pre-processing, Fuzzy C Mean Algorithm, Fuzzy Logic.

I. INTRODUCTION

During the past few years the World Wide Web has become the largest and most popular way of communication and information broadcasting. It serves as a platform for exchanging various kinds of information. The volume of information available on the internet is increasing rapidly with the explosive growth of the World Wide Web [1]. Web mining is the application of data mining techniques to extract knowledge from web data, including web documents, usage logs of web sites, etc. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining. Web usage mining is the process of extracting useful information from server logs i.e. users history [2]. The Web usage mining process could be classified into two commonly used approaches. The first approach maps the usage data of the Web server into relational tables before an adapted data mining technique is performed. The second approach uses the log data directly by utilizing special pre-processing techniques. Web usage mining focuses on techniques that could predict user behavior while the user interacts with the Web [3].

II. RELATED WORK

Web Mining is technique in data mining to extract knowledge from web data, including web documents, hyperlinks between documents, us-age logs of web sites, etc. There are three kinds of web mining categories: Web Usage Mining-It is the process of tracings useful information from server logs i.e. users history, Web Structure Mining- structure mining is to extract previously unknown relationships between Web pages; Web Content Mining-Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page contents[4].

Clustering analysis aims to group similar web usage sessions into identical clusters. We clustered the pre-processed WUM data using a swarm intelligence based optimization, PSO based clustering algorithm. In this paper, showed the performance of the Particle Swarm Optimization (PSO) algorithm is better than K-means clustering .The result of clustering of server log data based on these parameters: (a) time and request per 30 minutes distribution (b)page viewed and number of user distribution (c) session-number of request distribution (d) session-time distribution [5].

In [6], a cluster optimization technique is proposed to improve web usage mining using ant nest mate approach. As the size of the cluster increases, it will become an inevitable need to optimize the clusters. Cluster optimization methodology is based on ant nest mate recognition ability and is used to eliminate the data redundancies. For clustering ART1-nueral network based approach is used. The accuracy and completeness of the user profiles increases by cluster optimization.

Time aware web users clustering [7] emphasize the help us to discover similarities in usage patterns with respect to the time locality of their visit. Two clustering methods are used for tuning and binding the page and time visiting criteria. The clusters developed by this method presents similar behavior at the same time period, by varying the priority given to page or time visits.

III. OVERALL IDEAS

In order to do Research profiling based on web mining, we performed : (1) Data collection of web log data from the server ; (2)Preprocessing of web log data ; 3)Clustering of URLs;(4) Optimization for reducing the redundancy of clusters;(5) finely perform Rank Calculation of frequently occurred URLs Fig. 1 shows the overall idea for Research Profiling based on web mining.

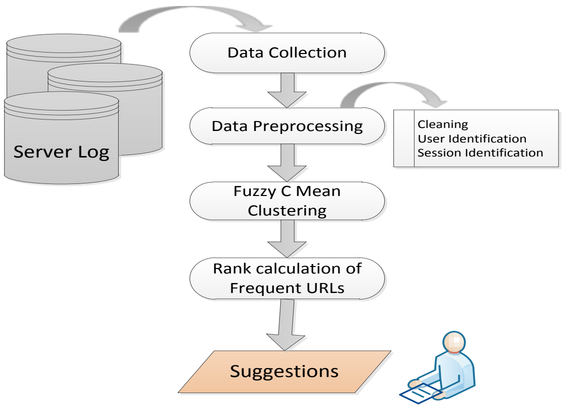
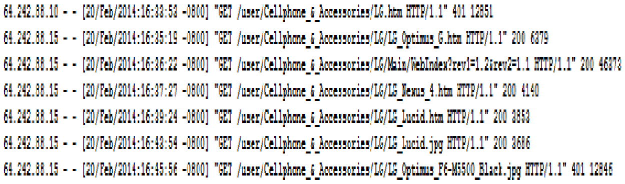


Figure 1: System Architecture

Data Collection: In this architecture, web log data is collected from web server which is navigation history of web site maintained in Web access sequence. Web access sequences are sequence of web pages which express a session of the user (Clicked sequence)and that can be collected by the log file of the web site. It is a huge repository of web pages and links, accesses web sites are recorded in web logs file. In this architecture , web access log (Common log format) is used as data which is generated. Common log format is a standard text file used by server while generating server log file.



```
64.242.88.10 - - [20/Feb/2014:16:33:53 -0800] "GET /user/Cellphone_&_Accessories/LG.htm HTTP/1.1" 401 12851
64.242.88.15 - - [20/Feb/2014:16:35:19 -0800] "GET /user/Cellphone_&_Accessories/LG/LG_Optimus_G.htm HTTP/1.1" 200 6379
64.242.88.15 - - [20/Feb/2014:16:36:22 -0800] "GET /user/Cellphone_&_Accessories/LG/Main/WebIndex?rev1=1.2&rev2=1.1 HTTP/1.1" 200 46373
64.242.88.15 - - [20/Feb/2014:16:37:27 -0800] "GET /user/Cellphone_&_Accessories/LG/LG_Nexus_4.htm HTTP/1.1" 200 4140
64.242.88.15 - - [20/Feb/2014:16:39:24 -0800] "GET /user/Cellphone_&_Accessories/LG/LG_Lucid.htm HTTP/1.1" 200 3853
64.242.88.15 - - [20/Feb/2014:16:43:54 -0800] "GET /user/Cellphone_&_Accessories/LG/LG_Lucid.jpg HTTP/1.1" 200 3686
64.242.88.15 - - [20/Feb/2014:16:45:56 -0800] "GET /user/Cellphone_&_Accessories/LG/LG_Optimus_F6-M5500_Black.jpg HTTP/1.1" 401 12846
```

Figure 2: Generated log data in Common Log format

A. Pre-processing: Pre-processing is necessary step in web mining, because log file contain noisy & ambiguous data which may affect result of mining process [8]. The input of the proposed system is web log file. First, raw data is read from Web server log files and for each HTTP request the following data were distinguished: the IP address of the Web client, the identifier of the Web client, the user identifier, the timestamp, the HTTP method, the URI of the resource requested, the version of HTTP protocol, the HTTP status code, the size of the object sent to the client. Example: 204.31.113.138 - [03/Jul/1996:06:56:12 -0800] "GET PowerBuilder/Compny3.htm HTTP/1.0" 200 5593. The data pre-processing step has data cleaning, user identification and session identification.

A) Data Cleaning- First stage of data cleaning is connected with elimination of useless data. Data cleaning is related to site specific, and involves extraneous references to embedded objects that may or may not be important for purpose of analysis, including references to style files, graphics or sound files. Therefore some of entries are useless for analysis process that is cleaned from the log files. By Data cleaning, errors and inconsistencies will be detected and removed to improve the quality of data[10]. Since our analysis concerns the behavior of users and involves a click-stream analysis, the following requests have been excluded from analysis: hits for embedded objects (e.g. images), automatically generated by Web client browsers, requests generated by Web bots (e.g. Web Crawlers). Thus, data cleaning includes the elimination of irrelevant entries like: 1)Removes requests concerning non-analyzed resources such as images, multimedia files, and page style files.2) Entries with unsuccessful HTTP status codes; HTTP status codes are used to indicate the success or failure of a requested event, and we only consider successful entries with codes between 200 and 299. 3) Entries with request methods except GET and POST.

B) User Identification- Identifying the individual uses by observing their IP address means user Identification. For Identifying the Unique User, proposed some Rules:1) If there is new IP address, then there is a new user, a reasonable assumption is that each different agent type for an IP address represents a different user.

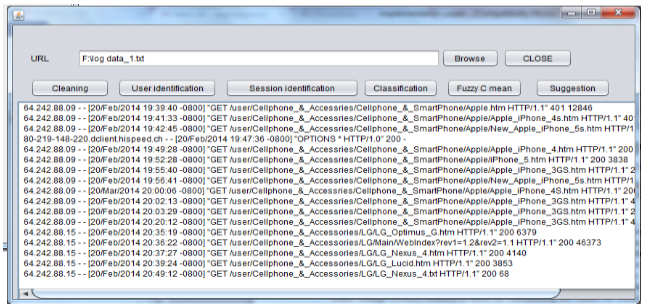
C) Session Identification- After user identification, the pages accessed by each user must be divided into individual session, which is known as session identification . The goal of session identification is to find each user's access pattern and frequently accessed path [12]. Mechanism used in this paper for implementation for time out a)defines a time limit for the access of a particular page and this limit is 30 minutes divided into more than one session. A session refers user's navigation behaviors in a Website, b) to identify the access time of the user for a respective web page.

3. Fuzzy C Means Clustering: Clustering is the process of collecting similar object one another. In this paper, we are using object for implementation as user session as time generated by pre_processing stage .In clustering , grouping performed based on users having similar access sequences. This implementation paper concerns clustering of each web access is composed of page url T time spend on that page. A web page pattern can be denoted by Si={(urli1, ti1),(urli2, ti2),……,(urlik, tik)} where 1<=i<=m, 'm' is number of web access patter derived from web log data.

4. **Rank calculation of frequent URLs:** In this architecture rank calculation is performed for individual cluster of frequently occurred URLs and showing top five URLs having highest rank as output in the form of suggestion.

IV. IMPLEMENTATION AND RESULTS

In implementation and evaluation phase, software specifications are Net Beans IDE 7.2 and my sql for the database. Server log data is generated in common log format. Offline process of pattern analysis on statistical log data that is mining process performed processing to extract meaningful pattern or URL by taking log data from the server. 1000 rows of text file is used for performing the data mining process on the data. GUI of our system looks like as follow:

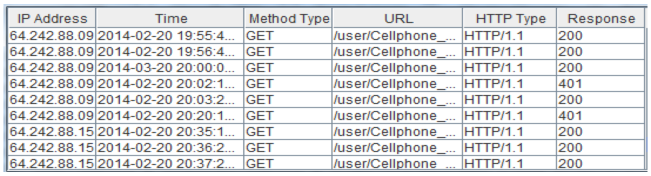


URL F:\log data_1.txt

```
64.242.88.09 - - [20/Feb/2014 19:39:40 -0800] "GET /user/Cellphone_&_Accessries/Cellphone_&_SmartPhone/Apple.htm HTTP/1.1" 401 12846
64.242.88.09 - - [20/Feb/2014 19:41:33 -0800] "GET /user/Cellphone_&_Accessries/Cellphone_&_SmartPhone/Apple/Apple_iPhone_4s.htm HTTP/1.1" 40
64.242.88.09 - - [20/Feb/2014 19:42:45 -0800] "GET /user/Cellphone_&_Accessries/Cellphone_&_SmartPhone/Apple/New_Apple_iPhone_5s.htm HTTP/1
80-219-148-220.dclient.hispeed.ch - - [20/Feb/2014 19:47:36 -0800] "OPTIONS * HTTP/1.0" 200 -
64.242.88.09 - - [20/Feb/2014 19:49:28 -0800] "GET /user/Cellphone_&_Accessries/Cellphone_&_SmartPhone/Apple/Apple_iPhone_4.htm HTTP/1.1" 200
64.242.88.09 - - [20/Feb/2014 19:52:28 -0800] "GET /user/Cellphone_&_Accessries/Cellphone_&_SmartPhone/Apple/iPhone_5.htm HTTP/1.1" 200 3838
64.242.88.09 - - [20/Feb/2014 19:55:40 -0800] "GET /user/Cellphone_&_Accessries/Cellphone_&_SmartPhone/Apple/Apple_iPhone_3GS.htm HTTP/1.1" 2
64.242.88.09 - - [20/Feb/2014 19:56:41 -0800] "GET /user/Cellphone_&_Accessries/Cellphone_&_SmartPhone/Apple/New_Apple_iPhone_5s.htm HTTP/1
64.242.88.09 - - [20/Mar/2014 20:00:06 -0800] "GET /user/Cellphone_&_Accessries/Cellphone_&_SmartPhone/Apple/Apple_iPhone_4S.htm HTTP/1.1" 20
64.242.88.09 - - [20/Feb/2014 20:02:13 -0800] "GET /user/Cellphone_&_Accessries/Cellphone_&_SmartPhone/Apple/Apple_iPhone_3GS.htm HTTP/1.1" 4
64.242.88.09 - - [20/Feb/2014 20:03:29 -0800] "GET /user/Cellphone_&_Accessries/Cellphone_&_SmartPhone/Apple/Apple_iPhone_3GS.htm HTTP/1.1" 2
64.242.88.09 - - [20/Feb/2014 20:20:12 -0800] "GET /user/Cellphone_&_Accessries/Cellphone_&_SmartPhone/Apple/Apple_iPhone_3GS.htm HTTP/1.1" 4
64.242.88.15 - - [20/Feb/2014 20:35:19 -0800] "GET /user/Cellphone_&_Accessories/LG/LG_Optimus_G.htm HTTP/1.1" 200 6379
64.242.88.15 - - [20/Feb/2014 20:36:22 -0800] "GET /user/Cellphone_&_Accessories/LG/Main/WebIndex?rev1=1.2&rev2=1.1 HTTP/1.1" 200 46373
64.242.88.15 - - [20/Feb/2014 20:37:27 -0800] "GET /user/Cellphone_&_Accessories/LG/LG_Nexus_4.htm HTTP/1.1" 200 4140
64.242.88.15 - - [20/Feb/2014 20:39:24 -0800] "GET /user/Cellphone_&_Accessories/LG/LG_Lucid.htm HTTP/1.1" 200 3853
64.242.88.15 - - [20/Feb/2014 20:49:12 -0800] "GET /user/Cellphone_&_Accessories/LG/LG_Nexus_4.txt HTTP/1.1" 200 68
```

Figure 3: GUI of the system

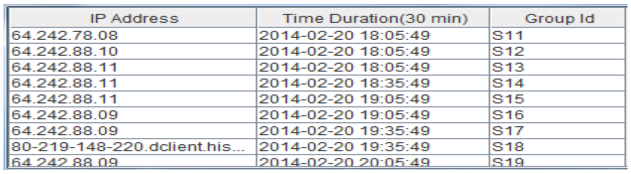
1) **Preprocessing**: In this phase, we have performed three steps of Preprocessing i.e. Cleaning, User Identification and Session Identification. **Cleaning**: Some of entries are useless for analysis process that is cleaned from the log files. Filtered file have stored in relational database for further processing. **User Identification**: In this implementation phase, we Identified the individual user by observing their IP address means user Identification.



IP Address	Time	Method Type	URL	HTTP Type	Response
64.242.88.09	2014-02-20 19:55:4...	GET	/user/Cellphone_...	HTTP/1.1	200
64.242.88.09	2014-02-20 19:56:4...	GET	/user/Cellphone_...	HTTP/1.1	200
64.242.88.09	2014-03-20 20:00:0...	GET	/user/Cellphone_...	HTTP/1.1	200
64.242.88.09	2014-02-20 20:02:1...	GET	/user/Cellphone_...	HTTP/1.1	401
64.242.88.09	2014-02-20 20:03:2...	GET	/user/Cellphone_...	HTTP/1.1	200
64.242.88.09	2014-02-20 20:20:1...	GET	/user/Cellphone_...	HTTP/1.1	401
64.242.88.15	2014-02-20 20:35:1...	GET	/user/Cellphone_...	HTTP/1.1	200
64.242.88.15	2014-02-20 20:36:2...	GET	/user/Cellphone_...	HTTP/1.1	200
64.242.88.15	2014-02-20 20:37:2...	GET	/user/Cellphone ...	HTTP/1.1	200

Figure 4: Result of User Identification Step in Preprocessing

Session Identification: The pages accessed by each user must be divided into individual session that is known as session identification. We have specified 30 minutes time duration for making the session ID for each user. If user appeared on web site and time duration is more than 30 minutes then more than session ID created.



IP Address	Time Duration(30 min)	Group Id
64.242.78.08	2014-02-20 18:05:49	S11
64.242.88.10	2014-02-20 18:05:49	S12
64.242.88.11	2014-02-20 18:05:49	S13
64.242.88.11	2014-02-20 18:35:49	S14
64.242.88.11	2014-02-20 19:05:49	S15
64.242.88.09	2014-02-20 19:05:49	S16
64.242.88.09	2014-02-20 19:35:49	S17
80-219-148-220.dclient.his...	2014-02-20 19:35:49	S18
64.242.88.09	2014-02-20 20:05:49	S19

Figure 5: Result of User Identification Step in Preprocessing

2) **Clustering**: After pre-processing phase, we performed clustering to take similar interest by grouping Session id's and urls. A hard clustering obtained from a fuzzy partition by using a threshold of the membership value. The most popular fuzzy clustering algorithm is the fuzzy c-means (FCM) algorithm. Fuzzy C-means (FCM) clustering is of overlapped clustering which allows one data object to belong to two or more clusters. In this algorithm, we used objects as user session ids generated by session

tracking stage and choose random centroid . Each web access is composed of page url and T time spend on that page. A web page pattern can be denoted by Si={(urli1, ti1),(urli2, ti2),……,(urlik, tik)} where 1<=i<=m, m is number of web access patter derived from web log data. The distance between two data object calculated using Euclidean distance which decided membership value for cluster on the basis of fuzzy logic which is shown by equation number (2)as given bellow:

$$U_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \quad ......(2)$$

$$C_j = \frac{\sum_{K=1}^{C} U_{ij}^{m} x_{ki}}{\sum_{K=1}^{C} U_{ij}^{m}} \quad ....(3)$$

Where,

U ij is the degree of membership of Xi in the cluster j,

Cj is the centre of the cluster,

C is the total number of clusters,

N is the total number of user sessions,

Xi is the feature vector.

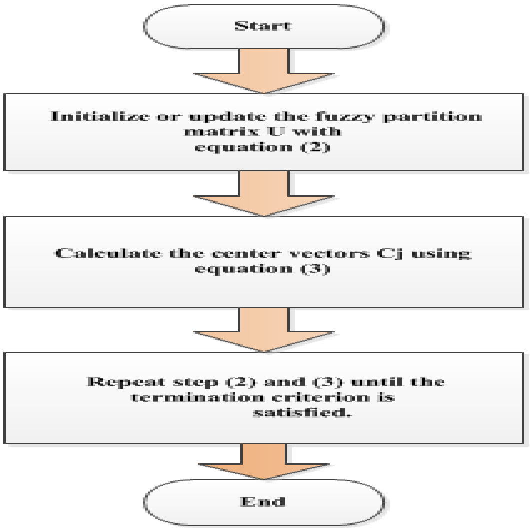
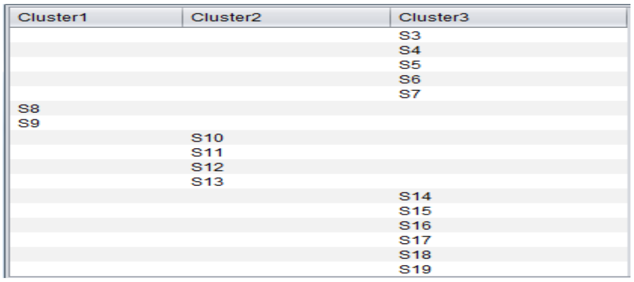

Start
Initialize or update the fuzzy partition matrix U with equation (2)
Calculate the center vectors Cj using equation (3)
Repeat step (2) and (3) until the termination criterion is satisfied.
End

Figure 6: Flowchart of FCM Algorithm

The Result of clustering on 300 rows of text file in which three clusters were formed because cluster head value can take randomly which is as shown follow:



Cluster1	Cluster2	Cluster3
		S3
		S4
		S5
		S6
		S7
S8		
S9		
	S10	
	S11	
	S12	
	S13	
		S14
		S15
		S16
		S17
		S18
		S19

Figure 7: Result of FCM Clustering Algorithm

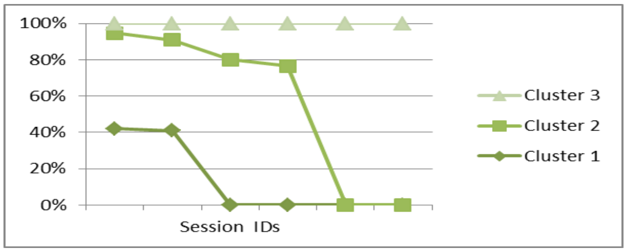
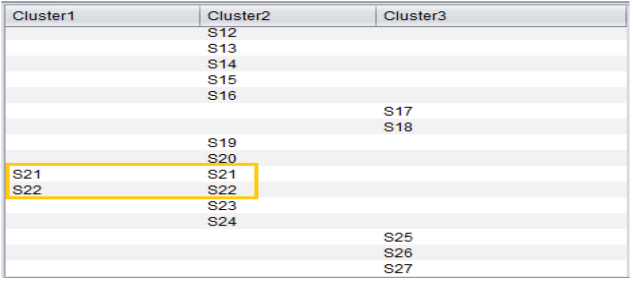


Figure 8: Result analysis of FCM algorithm on 300 rows of text file

When we increased the data redundancy found that result and graphical result analysis as follow:



Cluster1	Cluster2	Cluster3
	S12	
	S13	
	S14	
	S15	
	S16	
		S17
		S18
	S19	
	S20	
S21	S21	
S22	S22	
	S23	
	S24	
		S25
		S26
		S27

Figure 9: Result of FCM Clustering Algorithm on increased data

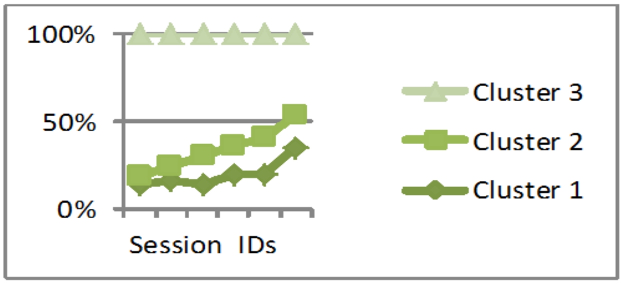
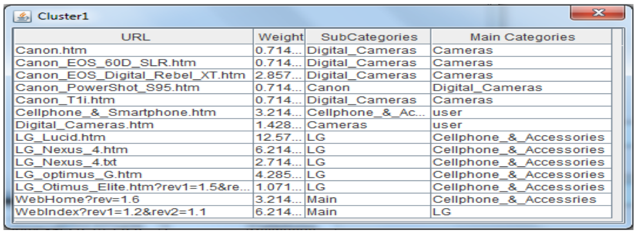


Figure 10: Result Analysis of FCM Clustering Algorithm on increased data

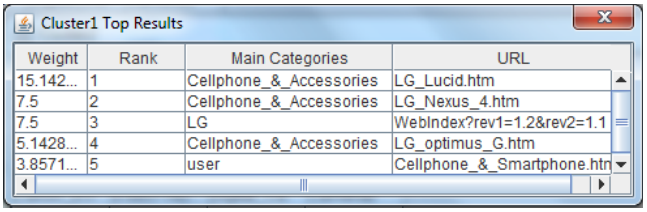
3) **Rank Calculation and Suggestions:** We calculated the rank of all URLs for categories and sub categories wise. We calculated rank using frequently appeared URLs divided by total number of URLs in one cluster and result of rank calculation for cluster 1 and top five results of a cluster as suggestions which is as shown below in figure 11

and 12 respectively. Similarly we calculated rank and their top results for cluster 2 and cluster 3 respectively



URL	Weight	SubCategories	Main Categories
Canon.htm	0.714...	Digital_Cameras	Cameras
Canon_EOS_60D_SLR.htm	0.714...	Digital_Cameras	Cameras
Canon_EOS_Digital_Rebel_XT.htm	2.857...	Digital_Cameras	Cameras
Canon_PowerShot_S95.htm	0.714...	Canon	Digital_Cameras
Canon_T1i.htm	0.714...	Digital_Cameras	Cameras
Cellphone_&_Smartphone.htm	3.214...	Cellphone_&_Ac...	user
Digital_Cameras.htm	1.428...	Cameras	user
LG_Lucid.htm	12.57...	LG	Cellphone_&_Accessories
LG_Nexus_4.htm	6.214...	LG	Cellphone_&_Accessories
LG_Nexus_4.txt	2.714...	LG	Cellphone_&_Accessories
LG_optimus_G.htm	4.285...	LG	Cellphone_&_Accessories
LG_Otimus_Elite.htm?rev1=1.5&re...	1.071...	LG	Cellphone_&_Accessories
WebHome?rev=1.6	3.214...	Main	Cellphone_&_Accessries
WebIndex?rev1=1.2&rev2=1.1	6.214...	Main	LG

Figure 11: Result of Rank Calculation



Weight	Rank	Main Categories	URL
15.142...	1	Cellphone_&_Accessories	LG_Lucid.htm
7.5	2	Cellphone_&_Accessories	LG_Nexus_4.htm
7.5	3	LG	WebIndex?rev1=1.2&rev2=1.1
5.1428...	4	Cellphone_&_Accessories	LG_optimus_G.htm
3.8571...	5	user	Cellphone_&_Smartphone.htm

Figure 12: Result of Suggestions of top five results of a cluster

Conclusion and future scope:

In this implementation paper we implemented pre-processing on statistical Web log data in local host machine, apply clustering to get similar interest of particular user and finally provided suggestions by identifying top five URLs from a cluster and suggestions are in the form of frequently viewed URLs by the web user. There are lots of challenges in web Mining to manage big data of the server and we need to solve them by apply different techniques on big log data .Our future work includes the suggestions on Big Data.

References

[1] V. Losarwar, Dr. Madhuri Joshi, "Data Pre-processing in Web Usage Mining", International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, 2012 Singapore

[2] M. Yadav, Mr. P. Mittal, "Web Mining: An Introduction", International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 3, March 2013.Sannella, M. J. 1994 Constraint Satisfaction and Debugging for Interactive

[3] Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE.

[4] Cyrus Shahabi • Farnoush Banaei-Kashani, ―Efficient and Anonymous Web-UsageMining for Web Personalization‖, INFORMS Journal on Computing © 2003 INFORMSVol. 15, No. 2, Spring 2003, pp. 123–147.

[5] S. Alam, G. Dobbie, P. iddle, ―Particle Swarm Optimization Based Clustering Of Web Usage Data‖,IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology 2008.

[6] Anna Alphy , S. Prabakaran , "Cluster Optimization for Improved Web Usage Mining using Ant Nestmate Approach", IEEE-International Conference on Recent Trends in Information Technology(ICRTIT), MIT, Anna University, Chennai. June 3-5, 2011

[7] Petridou, S.G.; Koutsonikola, V.A.; Vakali, A.I.; Papadimitriou, G.I.,‖time aware web users clustering‖,Knowledge and ata Engineering, IEEE Transactions on Volume:20,Issue:5,Page(s): 653 - 667 ,2008.

[8] Y.T. Yu, M.F. Lau, "A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions", Journal of Systems and Software, 2005, in press.

[9] Spector, A. Z. 1989. Achieving application requirements. In Distributed Systems, S. Mullende Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.

[10] Fröhlich, B. and Plate, J. 2000. The cubic mouse: a new device for three-dimensional input. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

[11] Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems.