Scientific Journal of Impact Factor(SJIF): 3.134

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2,Issue 5, May -2015

WEB PAGES RECOMMENDATION SYSTEM BASED ON K-MEDOID CLUSTERING METHOD

Richa Patel[1] Akshay Kansara[2]

[1]P.G. Student [2] Assistant Professor, P.G Department of Computer Science And Engineering
Saffrony Institute Of Technology,Linch, Mehasana, Gujarat, India

Abstract— *With an expontial growth of World Wide Web, there are so many information overloaded and it become hard to find out data according to need. Web usage mining is a part of web mining, which deal with automatic discovery of user navigation pattern from web log. Web Recommendation System is implemented by using Collaborative Filtering approach. It is a specific type of information filtering system that aims to predict the user browsing activity and then recommended to the user web pages items that are likely to be of interest. In this paper, a new recommendation system is proposed by using K- Medoid clustering approach to predict the user's navigational behavior. The proposed recommendation system based on K-medoid clustering performs well compared to K-Mean clustering algorithm. The performance of the comparative analysis is presented through given experimental results.*

Index Terms— *web mining, web usage mining, Web recommendation system, K-Mean clustering, K-Medoid Clustering, Cosine Similarity, Hamming distance*

I. INTRODUCTION

Introduction In today's world internet has become extremely popular and its growth is very rapid. The information is available on internet for people whom they are using for their different intend. The resources for using internet are growing fast, it is necessary for users to use automatic tools for discover desired information. People require systems at client side and server side for finding out the desired information. Above system used to mine data and extract information from that source. So for a specific user only intresting information of web is useful and rest of the information not important. Several users are interested in content of web and they can browse by using search engines. Using web log files, we can find out information related to web access pattern. These web log files provide the information related to behavior of user. In business area, user's behavior plays an important role for extracting information.

In this area, user navigation patterns are describe as the common browsing behaviors along with a group of users. During navigation, many users may have common interests .so navigation patterns should capture the overloaded information or user's need. In addition, navigation patterns should also be able to differentiate among web pages based on their different meaning to each pattern.

The rest of paper is as follows: Section II presents the overview of web mining and web recommendation system. Section III related to literature work. Section IV presents the block diagram of proposed method and implementation for the usage based recommendation system using K-Mean and K-Medoid clustering algorithms. Experimental result and discussion are revealed in Section V. Finally, Section VI is conclusion and future scope respectively.

II. WEB MINING

A. OVERVIEW

Web mining is a part of data mining techniques to automatically discover and extract knowledge from the web. It can be broadly divided into three domains: web content mining, web usage mining, and web structure mining. Classification of Web Mining can be understood using the given Fig .1

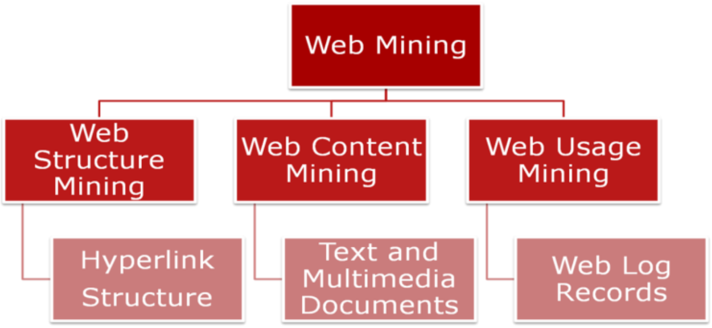


Fig.1: Web Mining Classification

Web Mining is an application of Data Mining which deals with the extraction of interesting knowledge from the World Wide Web. In above figure web mining can be classified into three categories: web structure mining, web content mining and web usage mining. Web Structure Mining is the process of inferring knowledge from the World Wide Web and links between web pages. The structure of a typical web graph contains web pages as nodes and hyperlinks as edges between related web pages. It is the process of using graph theory to examine the node and connection structure of a web site. Web content mining is the process of finding useful information from the available web pages. Generally, the web content mining consists of several types of source data such as textual, image, audio, video, metadata as well as hyperlinks. Web Usage mining is a part of data mining techniques to discover navigation patterns from web log. data is usually collected when user interact with web server such as web/proxy server logs, user queries, registration data. In short, Web usage mining is a process of extracts information from user how to use web sites. Web content mining is a process of extracts information from texts, images and other contents. Web structure mining is a process of extracts information from hyperlinks of web pages.

B. WEB USAGE MINING

Web usage mining is an part of web mining. It contain the two categories, first is general access pattern tracking and customized usage tracking. Now, general access pattern is a mining process using the history of the web page visited by user. And customized usage tracking is targeted on specific user. Web usage mining, the art of analyzing user interactions with a web page, has been dealt by several researchers using different approaches. [1] There are many research area in data mining techniques like association rule mining, classification, clustering and Sequential-pattern-mining-based.

1) WEB USAGE MINING TECHNIQUES

These techniques given below:

1. Sequential-pattern-mining-based: Allows the discovery of temporally ordered Web access patterns,
2. Association-rule-mining-based: Finds correlations among Web pages,
3. Clustering-based: Groups users with similar characteristics,
4. Classification-based: Groups users into predefined classes based on their characteristics [2].

C. WEB RECOMMENDATION SYSTEM

Web Recommendation system is a specific type of information filtering that attempts to predict the user next browsing activity then recommended to the user web pages items that are likely to be of interest to the user. The main goal of recommendation system is to improve the web site usability by knowing the interest of the users. The web recommendation process consists of two components namely online and off –line with respect to web server activity. Offline component build the knowledge base by analyzing historical data, such as server access log file or web logs which are captured from the server. Online component used in capturing the intuition list of the user so recommended page view to the user whenever user comes online for the next time.

There are three types of filtering. Content based Collaborative and Hybrid filtering. Content based filtering is based on the information about the items that are going to be recommended. Collaborative filtering is based on collecting and analyzing a large amount of information on user’s behavior, activity or preferences and predicting what users will liked based on their similarity to other users. Hybrid filtering is a combination of content based and collaborative filtering approach.

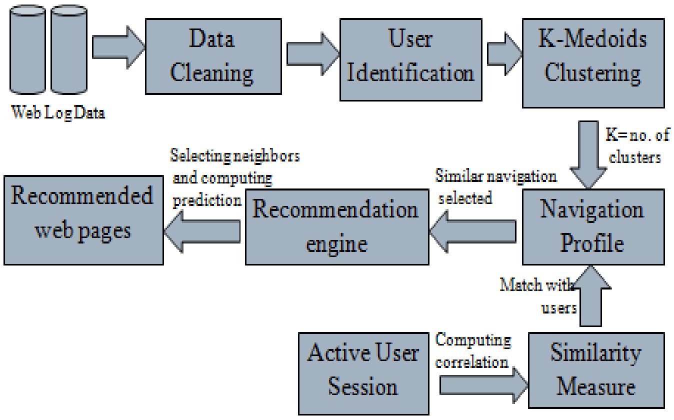
III. LITERATURE REVIEW

The focus of literature review is to study and analyze about available techniques to predict the user's navigation pattern with the web or web site. Recommender system plays a vital role in internet technology for data gathering and rating up a data. There are four types of filtering technique used in Recommender System-demographic, content, collaborative and hybrid [3]. The most widely and popularly used technique is collaborative Filtering [3]. In this paper they also describe the some potential problems with the Collaborative filtering RS. One is the scalability, which is how quickly a recommender system generates recommendation; second is sparsity and also cold start problem and better accuracy. Mehrdad Jalali et al.[4] focused on novel approach for classifying user navigation patterns by using Longest Common Subsequence (LCS) algorithm. They used some evolution methodology that can be used to evaluate the quality of the prediction found. Ping Ni et al. [5] propose a novel method for improving accuracy and effectiveness of news recommendation by combining K-means algorithm. In this paper, reclassification for the current classification through K-mean would be implemented based on the feedback of web usage mining in order to improve the accuracy of news recommendation and convergence of classification. In [6],

In the proposed method we extract required patterns by removing noise that is present in the web document. In this paper author proposed a new method web data extraction algorithm which solves the problem like web noisy data, junk mails, spam mails, advertisement etc. This method is used to identify required patterns in an effective manner. Sanjeev Kumar Sharma et al. [7], the proposed SEP architecture consists of three modules such as, Original Recommendation, Semantic Recommendation and Category-based Recommendation. The semantic recommendation will be performed using various data mining techniques such as clustering, association-rule-mining, and similarity measures. Mohammd Hamidi Esfahani et. Al. [8], Many clustering methods used to in recommendation systems but a few of these methods are light or easy to use so they can make the recommendation process and user feedback faster, in the other hand, having a good recommendation is more useful than having too many recommendations that a few of them take the user attention.
In [9], this paper makes analysis on some major recommendation methods based on web data mining such as
Collaborative Filtering and Association Rules mining, and discusses the practical application of these methods in the tourism e-commerce, and then presents a design of web mining based tourism e-commerce recommender system with offline and online modules. In this paper, a framework is generated for capturing recommendations in the form of recommendation list for user using Weighted K-Means clustering [10]. A recommendation list consists of list of pages visited by user as well as list of pages visited by other users of having similar usage profile. R.Thiyagarajan et.al. [11], this paper has paid an attention to group the similar usage behavior of users using K-Means algorithm and new validating measure called MSR is applied to evaluate the cluster's quality.

IV. PROPOSED ARCHITECTUER

The proposed system uses User Based Collaborative Filtering technique which consists of two main components, namely the offline phase and online phase.



The main aim of the proposed model is to predict user navigation to predict the user next browsing activity and then recommend to the user web pages items that are likely to be of interest to the user. User navigation pattern is defined as common browsing characteristics among group of users. Different users have common browsing practices and navigation patterns needs to capture these common interests to identify user needs. Clustering technique is used to group users with similar browsing characteristics and classification technique is used to associate navigation behavior with these groups of users. Clustering has the advantage of grouping common interest users together and the user of classification will be able to classify different user requests.

Here main aim to predict user's navigation behavior. In this proposed work a framework is generated for capturing recommendations in the form of recommendation list for user using clustering techniques. A recommendation list consists of list of pages visited by user as well as list of pages visited by other users of having similar usage profile. In the offline component the three important steps are considered. First step is to clean the web server logs or web usage data by applying data cleaning techniques and then partition into session and also identify potential user from web usage data. In second step, we use usage based recommendation system using K-Medoid Clustering algorithm. Finally, web navigation profile generated based on the performed clusters.

In online phase, active user session match similarity with two measures. One is cosine similarity and second is hamming similarity measure. Based on this two similarity measure recommended web pages to users.

Web Log Data: A log file record contains important information about a request: the client side host name or IP address, the date and time of the request, the requested file name, the HTTP response status and size, the referring URL, and the browser information.

Data Preprocessing: The pre-processing steps include cleaning, user identification and session identification.

K-Mean clustering algorithm:

Here K-Mean clustering algorithm is used to group the web users in this paper. Consider a dataset contain the data to be clustered data points, D={X1... Xn}, first choose from this data points, K initially centroid randomly , where K is user-parameter, the number of clustered desired.

The process of K-means clustering is explained as follows [12]:

(i) The initial seeds with the chosen number of clusters, K, are selected and an initial partition is built by using the seeds as the centroids of the initial clusters.

(ii) Each data point is assigned to the centroid that is nearest, thus forming a cluster.

(iii) Keeping the same number of clusters, the new centroid of each cluster is calculated.

(iv)Iterate Steps (ii) and (iii) until the clusters stop changing or stop conditions are satisfied.

In K-mean Clustering, Each cluster is represented by center of the cluster and this clustering algorithm is sensitive to outlier and also noisy data. Hence, K-Medoid algorithm has been proposed in this paper and web page recommendation has been done accordingly. In this paper, K-Medoid clustering algorithm is used in the offline phase to generate the similar user groups or user clusters are used to generate the user navigation profile.

K-Medoid clustering algorithm: K-medoids method which is based on representative object techniques. In K-Medoid, each cluster is represented by one of the objects in the cluster rather than center of the cluster. K-medoids is more robust than k-means in presence of noise and outliers because medoids is less influenced by outlier.

The steps of K-Medoid clustering are as given below.

```
Step 1:- Arbitrarily choose 'k' objects as the initial medoids;
Step 2:- Repeat,
    a. Assign each remaining object to the cluster with the nearest medoid;
    b. Randomly select a non-medoid object;
    c. Compute the total cost of swapping old medoid object with newly selected non-medoid object.
    d. If the total cost of swapping is less than zero, then perform that swap operation to form the new set of k- medoids.
Step: - 3 until no change
```

Medoid is replaced with centroid to represent the cluster. Medoid is the most centrally located data object in a cluster. Here, k data objects are selected randomly as medoids to represent k cluster and remaining all data objects are placed in a cluster having medoid nearest (or most similar) to that data object. After processing all data objects, new medoid is determined which can represent cluster in a better way and the entire process is repeated. Again all data objects are bound to the clusters based on the new medoids. In each iteration, medoids change their location step by step. This process is continued until no any medoid move.

Navigation Profile: The clusters produced by the clustering step (previous step) are used to generate the navigation profile with one profile for each cluster by setting the min_sup and min_weight. The web navigation profile contains only those page views that passed certain confidence support and weights values.

Similarity Measure: Here, Cosine similarity and Hamming similarity measures are used to measure the similarity between the active user and the extracted usage profiles in the first phase.

In Cosine Similarity [10], the similarity of the active session with each of the discovered aggregate profile is determined.

Cosine Similarity (d1, d2) = dot(d1, d2)/||d1|| ||d2||dot(d1, d2)

Where d1 is a each cluster and d2 is a active user session and then measure the similarity.

In Hamming Distance measure [10], it is a distance measure and also cannot be negative. If it is zero, then the vectors are identical. Hamming distance is used when the vector are binary; they consists of 0's and 1's only.

Hamming Similarity (ui, uj) = 1- Hamming distance(ui, uj)

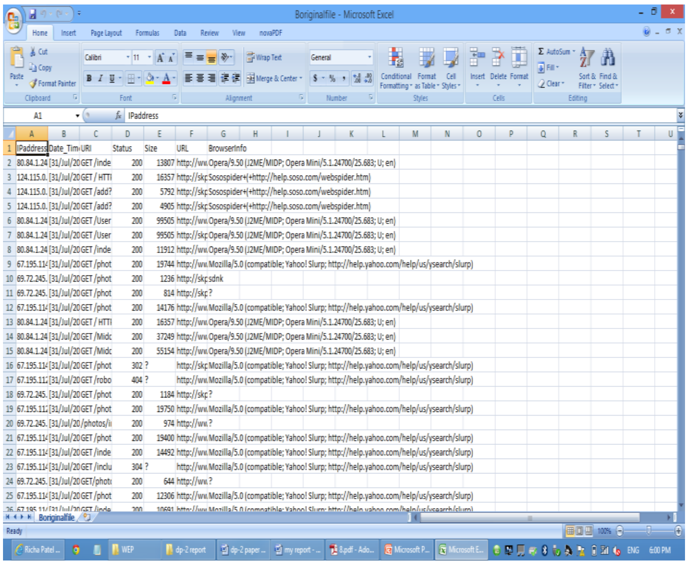
In the first phase navigation profiles are extracted. In second phase, this two measure Cosine similarity and Hamming distance are used to measure the similarity between active user and extracted usage profiles. And then recommendation list is generated from the nearest usage profiles.

Recommendation Engine: The main objective of Recommendation engine in this part of architecture is to generate list of recommendation web pages to user and as well as list of pages visited by other users of having similar usage profile.

V. EXPERIMENTAL RESULT AND DISCUSSION

A real dataset is used for this experiment. The dataset is taken from the Boriginal log file. It contain the details such as IP address, URL field, Status code, Date_time, Size field, URI field and BrowserInfo. The dataset is given below.

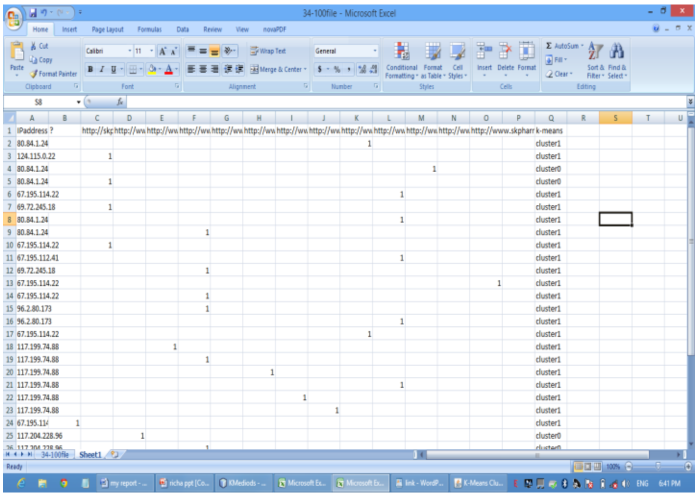
Fig.2 Dataset used in the experiment



	A	B	C	D	E	F	G
1	IPaddress	Date_Tim	URI	Status	Size	URL	BrowserInfo
2	80.84.1.24	[31/Jul/20	GET /inde	200	13807	http://ww	Opera/9.50 (J2ME/MIDP; Opera Mini/5.1.24700/25.683; U; en)
3	124.115.0.	[31/Jul/20	GET / HTTI	200	16357	http://skp	Sosospider+(+http://help.soso.com/webspider.htm)
4	124.115.0.	[31/Jul/20	GET /add?	200	5792	http://skp	Sosospider+(+http://help.soso.com/webspider.htm)
5	124.115.0.	[31/Jul/20	GET /add?	200	4905	http://skp	Sosospider+(+http://help.soso.com/webspider.htm)
6	80.84.1.24	[31/Jul/20	GET /User	200	99505	http://ww	Opera/9.50 (J2ME/MIDP; Opera Mini/5.1.24700/25.683; U; en)
7	80.84.1.24	[31/Jul/20	GET /User	200	99505	http://skp	Opera/9.50 (J2ME/MIDP; Opera Mini/5.1.24700/25.683; U; en)
8	80.84.1.24	[31/Jul/20	GET /inde	200	11912	http://ww	Opera/9.50 (J2ME/MIDP; Opera Mini/5.1.24700/25.683; U; en)
9	67.195.114	[31/Jul/20	GET /phot	200	19744	http://ww	Mozilla/5.0 (compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp)
10	69.72.245.	[31/Jul/20	GET /phot	200	1236	http://skp	sdnk
11	69.72.245.	[31/Jul/20	GET /phot	200	814	http://skp	?
12	67.195.114	[31/Jul/20	GET /phot	200	14176	http://ww	Mozilla/5.0 (compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp)
13	80.84.1.24	[31/Jul/20	GET / HTTI	200	16357	http://ww	Opera/9.50 (J2ME/MIDP; Opera Mini/5.1.24700/25.683; U; en)
14	80.84.1.24	[31/Jul/20	GET /Midc	200	37249	http://ww	Opera/9.50 (J2ME/MIDP; Opera Mini/5.1.24700/25.683; U; en)
15	80.84.1.24	[31/Jul/20	GET /Midc	200	55154	http://ww	Opera/9.50 (J2ME/MIDP; Opera Mini/5.1.24700/25.683; U; en)
16	67.195.114	[31/Jul/20	GET /phot	302	?	http://skp	Mozilla/5.0 (compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp)
17	67.195.112	[31/Jul/20	GET /robo	404	?	http://ww	Mozilla/5.0 (compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp)
18	69.72.245.	[31/Jul/20	GET /phot	200	1184	http://skp	?
19	67.195.112	[31/Jul/20	GET /phot	200	19750	http://ww	Mozilla/5.0 (compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp)
20	69.72.245.	[31/Jul/20	/photos/i	200	974	http://ww	?
21	67.195.114	[31/Jul/20	GET /phot	200	19400	http://ww	Mozilla/5.0 (compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp)
22	67.195.114	[31/Jul/20	GET /inde	200	14492	http://ww	Mozilla/5.0 (compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp)
23	67.195.114	[31/Jul/20	GET /inclu	304	?	http://ww	Mozilla/5.0 (compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp)
24	69.72.245.	[31/Jul/20	GET/photo	200	644	http://ww	?
25	67.195.114	[31/Jul/20	GET /phot	200	12306	http://ww	Mozilla/5.0 (compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp)

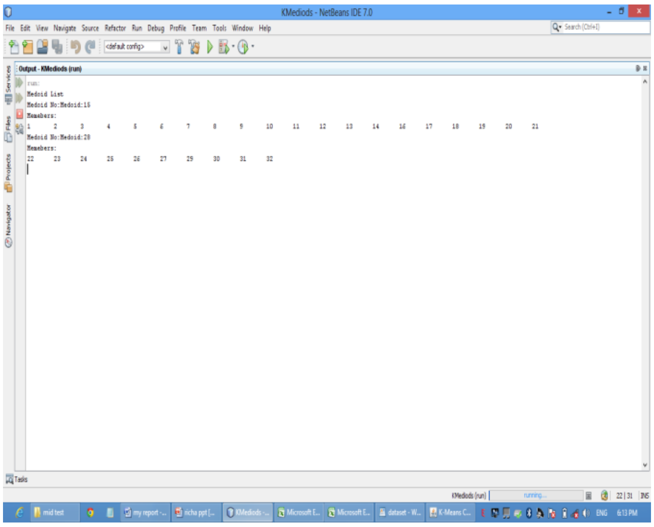
In the offline phase, first the preprocess the data on given dataset. We first clean the data using data cleaning techniques and then remove all duplicates and find potential users from given data set. Next, we use K-Mean clustering techniques and K-Medoid clustering techniques are applied to Boriginal dataset with k=2.

Fig.3 Result of K-mean Clustering Algorithm



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	IPaddress	?	http://skp	http://ww	http://ww	http://ww	http://ww	http://ww	http://ww	http://ww	http://ww	http://ww	http://ww	http://ww	http://www.skpharr		k-means
2	80.84.1.24										1						cluster1
3	124.115.0.22		1														cluster1
4	80.84.1.24												1				cluster0
5	80.84.1.24		1														cluster0
6	67.195.114.22											1					cluster1
7	69.72.245.18		1														cluster1
8	80.84.1.24											1					cluster1
9	80.84.1.24					1											cluster1
10	67.195.114.22		1														cluster1
11	67.195.112.41											1					cluster1
12	69.72.245.18					1											cluster1
13	67.195.114.22														1		cluster1
14	67.195.114.22					1											cluster1
15	96.2.80.173					1											cluster1
16	96.2.80.173											1					cluster1
17	67.195.114.22										1						cluster1
18	117.199.74.88				1												cluster1
19	117.199.74.88					1											cluster1
20	117.199.74.88							1									cluster1
21	117.199.74.88											1					cluster1
22	117.199.74.88								1								cluster1
23	117.199.74.88									1							cluster1
24	67.195.114	1															cluster1
25	117.204.228.96			1													cluster0

Fig.4 Result of K-Medoid Clustering Algorithm


KMediods - NetBeans IDE 7.0
File Edit View Navigate Source Refactor Run Debug Profile Team Tools Window Help
<default config>
Output - KMediods (run)
Services
Files
Projects
Navigator
Tasks

```
run:
Medoid List
Medoid No:Medoid:15
Memebers:
1       2       3       4       5       6       7       8       9       10      11      12      13      14      16      17      18      19      20      21
Medoid No:Medoid:28
Memebers:
22      23      24      25      26      27      29      30      31      32
```

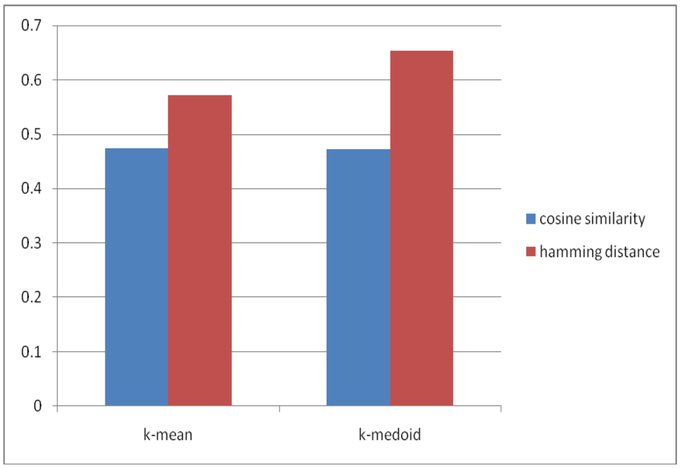
In given Fig.3 and Fig.4 clusters are generated using K-mean and K-medoid in the usage profile for the corresponding clusters. Now in second phase, active user visits the pages suppose 1 and 2 from page view categories. It is symbolically denoted as

A= {1,1,0,0,0,0,0,0,0,0,0,0,0,0,0}

Here, 0 means not pages visited by active user and 1 means pages visited by active user session.

After generating cluster active user session match similarity with two measures which is cosine similarity and hamming distance measures. Result of two similarity is given below.

Fig.5 Comparisons between K-mean and K-medoid clustering algorithm using similarity measures



In Fig.5, comparison of cosine similarity and hamming distance measure for the clusters generated using k-mean and K-medoid clustering techniques. It is clearly depicts that Hamming similarity value is higher than the cosine similarity value. This is because of the binary representation of the web data.

After measure the similarity with active user session, now we generate the recommendation web pages using K-mean and K-medoid clustering algorithm. To calculate cluster quality equation is given below.

Percentage of Recommendation Quality =Number of correctly recommended pages/ (Total Number of Visited pages - Number of Pages in the Active User Session) *100

Fig.6 Comparison of Recommendation Quality using K-mean and K-medoid clustering algorithm

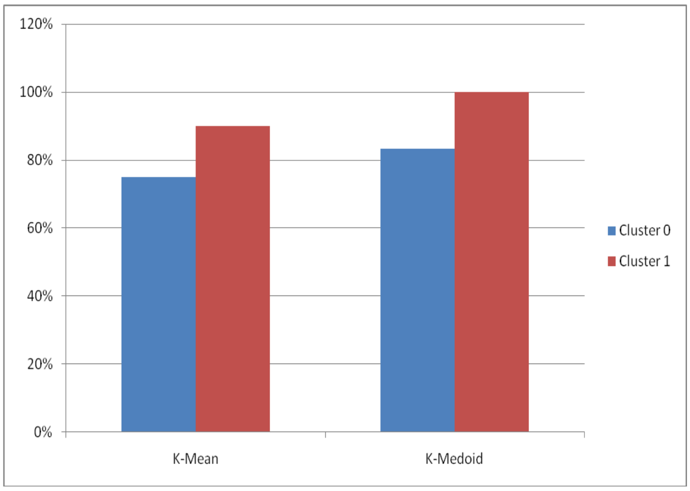


Fig.6 shows the recommendation list for the active user as given above using K-mean clustering and K-medoid clustering respectively. From given result, we observed that Hamming similarity using K-medoid clustering gives better recommendation quality than cosine similarity measure for the binary web usage data. The recommendation quality is given in figure.

VI. CONCLUSION AND FUTURE SCOPE

Here, in this paper a usage navigation pattern prediction system was presented. The system consists of four stages. cleaning stage, identify potential user, k- medoids clustering algorithm used to discover the navigation pattern and based on similarity measure with active user session, recommendation engine generate recommendation to target user. To group the similar usage behavior of users using K-Medoid algorithm for aggregated usage profile is applied to evaluate the cluster's quality. The results of this clustering approach are compared with the results of traditional clustering called K-Means. This approach improved the quality of clustering for user navigation pattern and the quality of recommendations. In future, the overlapping clusters may be obtained and these clusters may be used for usage profile generation. Hence more pages visited by users can be considered for recommendation process.

REFERENCES

[1] Mrs.Niranjana.Kannan and Dr (Mrs).Elizabeth Shanthi, "Classification and Clustering of Web Log Data to Analyze User Navigation Patterns", Volume 1, No. 1, August 2010
[2] Yew-Kwong Woon Wee-Keong Ng Ee-Peng Lim , "Web Usage Mining: Algorithms and Results
[3] Atisha Sachan, Vineet Richariya, "A Survey on Recommender System based on Collaborative Filtering Technique", International Journal of Innovation in Engineering and Technology(IJEIT), Volume 2,Issue 2, April 2013
[4] Mehrdad Jalali ,Norwati Mustapha, Ali Mamat, Md. Nasir B Sulaiman , "A new classification model for online predicting users' future movements", IEEE 2008
[5]PingNi, Jianxin Liao, Xiaomin Zhu,Keyan Ren , "News Contents Recommendation Model Based on Feedback of Web Usage", IEEE 2009
[6]V. Shanmuga Priya, S. Sakthivel, "An Implementation of Web Personalization Using Web Mining Techniques", International Journal of Computer Science and Mobile Computing(IJCSMC,) Vol. 2, Issue. 6, June 2013, pg.145 – 150
[7]Sanjeev Kumar Sharma, Dr. Ugrasen Suman, "Design and Implementation of Architectural Framework of Recommender System for e-Commerce" , International Journal of Computer Science and Information Technology & Security (IJCSITS), Vol. 1, No. 2, December 2011
[8]Mohammad Hamidi Esfahani, Farid Khosh Alhan, "New Hybrid Recommendation System Based On C-Means Clustering Method" , IEEE 2013
[9] Xuesong Zhao, Kaifan Ji, "Tourism E-Commerce Recommender System Based on Web Data Mining" , IEEE 2013
[10] R. Thiyagarajan, K. Thangavel, R. Rathipriya , " Recommendation of Web Pages using Weighted K-Means Clustering" , International Journal of Computer Applications (0975 – 8887) Volume 86 – No 14, January 2014
[11] R. Thiyagarajan, K. Thangavel, R. Rathipriya ," Usage Profile based Recommendation system" , IEEE 2014
[12] Kyoung-jae Kim, Hyunchul Ahn, A Recommender system using GA K-means clustering in an online shopping market.., Expert Systems with Applications (20070, doi:10,1016/j.eswa.2006.12.025.