ISSN 2348 – 8034
Impact Factor- 3.155

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

I (INTELLIGENT) CLUSTERING: AN APPROACH FOR IMPROVED WEB SITE STRUCTURE USING K- MEANS

Mr.Pramod B. Dhamdhere[*1] and Prof.Pratap Singh[2]
[*1]ME Student
[2]IOKCOE,Department of Computer Engg., India

ABSTRACT

K-Means is one of the most popular partition based clustering technique because of its simplicity and speed of classifying massive data rapidly and efficiently. The output of K- Mean's algorithm highly depends upon the selection of initial cluster centers because the initial Cluster centers are chosen randomly. The proposed system uses K-means Clustering with the slight improvement in the value of K to be considered .To input the required number of clusters requires knowledge of the domain .So the proposed system works on this limitation. Development of websites to facilitate effective user navigation is the challenging task observed these days. Because the way web developers think and design the system is quite different from that of the user. Different methods have been projected to re-link Webpages in order to recover navigability using user direction-finding data. The fully reorganized emerging structure can be highly impulsive, and the cost of disorienting users after the changes remains unanalyzed. The proposed system presents architecture to cluster the usage statistics of all the users to re-link Webpages. The re-ordering or reforming will mostly be based on clusters generated. Therefore an optimum choice of clusters is important step in operation of the method. Hence system uses an enhanced K means clustering algorithm where in the number of clusters (optimal) can be routinely designed and clusters are generated consequently. The system also develops an arithmetical programming model to recover the user navigation on a website. The method is imaginary the transport the functionality of test counter website for data gathering and then regroup it.

Keywords: *K-means Clustering, User navigation, relinking.*

I. INTRODUCTION

Clustering is a data mining technique which helps in grouping or making clusters of data having similar values of Some of the data attributes. Clustering can be used in various Fields like in Health sector for grouping patients with similar Symptoms of the disease, in banking sector to group customers who have dues in their credit card payments, in Market analysis to identify the customers having similar buying patterns. Currently, researchers are exploring the application of this technique in the field of education to better understand students' academic performance and the academic framework in which they learn. Nowadays there has been increasing investments in website design but it is still exposed, however, that finding necessary material in a website is relatively problematic. Designing effective websites is cumbersome task. Palmer indicated that poor website design has been a key element in a number of high profile site letdowns. McKinney et al. also discover that users having difficulty in pinpointing the targets are probably to leave a website even if its information is of good quality. Earlier studies on website has concentrated on a diversity of issues, such as understanding web structures, locating related pages of a given page, mining useful structure of a news website, and removing template from web pages. This work is related to the literature that observes how to recover website navigability through the use of user navigation data. Different works have made an effort to address this question and they can be usually categorized into two types: to help a particular user by animatedly reconstructing pages based on his contour and traversal paths, often denoted as personalization, and to adapt the site structure to simplify the navigation for all users, often stated as transformation. A principal cause of poor website design is that the web developers understanding of how a website should be organized can be considerably diverse from those of the users. Such variances result in cases where users cannot certainly trace the preferred information in a website. This problem is hard to escape because when forming a website, web developers don't have a perfect understanding of users likings and can only form pages based on their own verdicts. However, the degree of website effectiveness should be the approval of the users rather than that of the developers. Thus, Webpages should be structured in a way that generally matches the users model of how pages should be organized. This paper proposed a modified K-means algorithm which classifies the input data set into appropriate clusters without taking number of clusters K as input, as it was required in the case of K-means. The proposed algorithm does not require the number of clusters K as input distinguishing the shortcomings of website reorganization tactics, proposed system addresses the question of how to recover the organization of a website rather than reorganize it

ISSN 2348 – 8034
Impact Factor- 3.155

substantially specifically. We develop a mathematical programming (MP) model that simplifies user navigation on a website with slight changes to its present structure. Our model is mostly suitable for informational websites whose matters are static and quite stable over time Examples of informational websites are universities, hospitals, tourist attractions, federal agencies, and sports organizations. Our model, however, may not be federal activities, and sports administrations, Our model, however, for websites that purely use dynamic pages or have volatile contents. This is because a steady state might never be reached in user access patterns in such websites, so it may not be possible to use the weblog data to improve the site structure. The relevancy of web page can be attained by considering the amount of in-links and outlines existing in a particular web page. When the web page has more number of out-links to a pertinent page, then that page can be treated as a central page. From this central page, all remaining web pages are compared for similarity and the most similar pages are grouped together. The combination of most parallel pages together is known as clustering. Clustering can be done based on different algorithms such as hierarchical, k-means, partitioning, etc. The very easiest unverified learning algorithm that solve clustering problem is K- Means algorithm. It is a simple and easy way to classify a given data set through a certain number of clusters.

A) MOTIVATION

Data Security is the science and study of methods of protecting data from unauthorized disclosure and modification as per the technology upgraded, there is need to secure data which is transmitted over the network. Unsecured networks can be hacked into easily, and hackers can do lots of things in short amounts of time. A hacker can search The hard drive of the average PC user in less than a minute. In this short time period a search can be conducted on spread sheets or databases that contain user names and passwords.

II. MY WORK

A principal cause of poor website design is that the web developers understanding of how a website should be organized can be considerably diverse from those of the users. Such variances result in cases where users cannot certainly trace the preferred information in a website. This problem is hard to escape because when forming a Website, web developers don't have a perfect understanding of users likings and can only form pages based on their own verdicts. However, the degree of website effectiveness should be the approval of the users rather than that of the developers. Thus, Webpages should be structured in a way that generally matches the users model of how pages Should be organized Distinguishing the shortcomings of website reorganization tactics, proposed system Addresses the question of how to recover the organization of a website rather than reorganize it substantially. Specially, we develop a mathematical programming (MP) model that simplifies user navigation on a website with slight changes to its present structure. Our model is mostly suitable for informational websites whose matters are Static and quite stable over time. Examples of informational websites are universities, hospitals, tourist attractions, federal agencies, and sports organizations. Our model, however, may not be appropriate for websites that purely use dynamic pages or have volatile contents. This is because a steady state might never be reached in user access patterns in such websites, so it may not be possible to use the weblog data to improve the site structure. The relevancy of web page can be attained by considering the amount of in-links and out-links existing in a particular web page. When the web page has more number of out-links to a pertinent page, then that page can be treated as a central page. From this central page, all remaining web pages are compared for similarity and the most similar pages are grouped together. The grouping of most similar pages together is known as clustering. Clustering can be done based on different algorithms such as hierarchical, k-means, partitioning, etc. The simplest unsupervised learning algorithm that solve clustering problem is K- Mean's algorithm. It is a simple and easy way to classify a given data set through a certain number of clusters.

III. LITERATURE REVIEW

.Personalization is a method of modifying net sides to the necessities of exact customers by the material of the customer's directional performance and outline records [2]. Perkowitz and Etzioni [3] describe an approach that automatically synthesizes index pages which contain links to pages pertaining to particular Topics based on the co-occurrence regularity of sides in user traversals, to simplify operator navigation. File gathering [3] algorithm is more effective in execution the gathering by as each file as first centroid and before combines those papers in a group by considering the relevancy of matters, till completely papers in a group must parallel feature. The best mutual file gathering methods are of two types such as: Agglomerative Ranked gathering and K- Means groups. Web transformation, on the other hand, involves changing the structure of a website to facilitate the navigation for a large

ISSN 2348 – 8034
Impact Factor- 3.155

set of users [4] instead of personalizing pages for individual users. Fu et al. [5] describe an approach to reorganize web pages so as to provide users with their desired information in fewer clicks. However, this approach considers Only local structures in a website rather than the site as a whole, so the new structure may not be necessarily optimal. Gupta et al. [6] propose a heuristic method based on simulated annealing to relink web pages to improve navigability. This method makes use of the aggregate user preference data and can be used to improve the link structure in websites for both wired and wireless devices. However, this approach does not yield optimal solutions and takes relatively a long time (10 to 15 hours) to run even for a small website. The K-means algorithm is considered as one of the most commonly used algorithms for classification of numeric data in data mining [7] Lot of researches and studies are going on to address two of the major limitations of K- means algorithm One to select efficiently the initial centroids and second to remove the need of giving the number of clusters required as input to the algorithm.

IV. PROPOSED SYSTEM

When User want Surfing on Internet that time user did not get actual information he want He/she has to spend a lot of time on that particular web site . This Paper we suggest a new Data mining algorithm k means i.e. improved k-means algorithm. This Better-quality K-means Work on at database of Web server [12] this algorithm take input as session log with preferences And then transform these input into the number of clusters. The cluster is depends on the no of input so the total no link and the relinking of that all pervious links of particular website.[13] With the help of relinking and linking we find the priority of that particular link and these link come on the very front page of web site. This way we can decrease Time complexity.

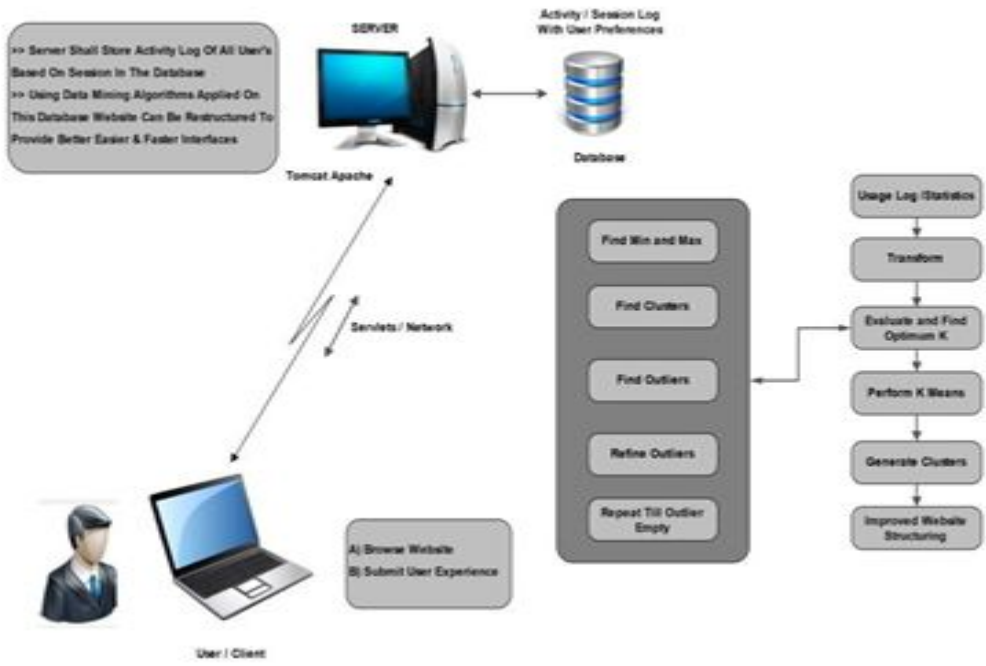

Fig1:I-clustering using K means for improving Website Structure

.

USER: The end user is the actual user who searches for the Relevant data on the web using browser installed in his system. Web Server: It tracks the search request made by the user. Also server maintains the activity session log with the user preferences and stores it in the database. Using improved k means clustering algorithm it improves the navigation of the website so that it provides better, easier and fast interface.

ISSN 2348 – 8034
Impact Factor- 3.155

V. ALGORITHM

Suitable number of clusters with n tuples distributed properly
1) Compute sum of the attribute values of each tuple (to find the points in the data set which are farthest apart).
2) Take tuples with minimum and maximum values of the sum as initial centroids.
3) Create initial partitions (clusters) using Euclidean distance between every tuple and the initial centroids.
4) Find distance of every tuple from the centroid in both the initial partitions. Take d=minimum of all distances. (Other than zero)
5) Compute new means (centroids) for the partitions created in step 3.
6) Compute Euclidean distance of every tuple from the new means (cluster centers) and fmd the outliers depending on the following objective function: If Distance of the tuple from the cluster mean ¡d then not an Outlier.
7) Compute new centroids of the clusters.
8) Calculate Euclidean distance of every outlier from the new cluster centroids and find the outliers not satisfying the objective function in step 6.
9) Let B=Yl, Y2,. ....Y p) be the set of outliers obtained in step 8 (value of k depends on number of outliers).
10) Repeat until I (B==¡D)
a) Create a new cluster for the set B, by taking mean value of its members as centroid.
b) Find the outliers of this cluster, depending on the objective function in step 6.
c) If No.of outliers = p then
i) Create a new cluster with one of the outliers as its member and test every other outlier for the objective function as in step 6.
ii) Find the outliers if any
d) Calculate the distance of every outlier from the centroid of the existing clusters and adjust the outliers the existing which satisfy the objective function in step 6. e) B= ZI,Z2 .... Zebu the new set of outliers. (value of q depends on number of outliers).

VI. EXPRIMENTAL RESULT

Here we use the data set that is S-Set of System W-Set of Web site, H=is set of Hit. F=functionality of the Find the Outliers C= is set of the centroid.

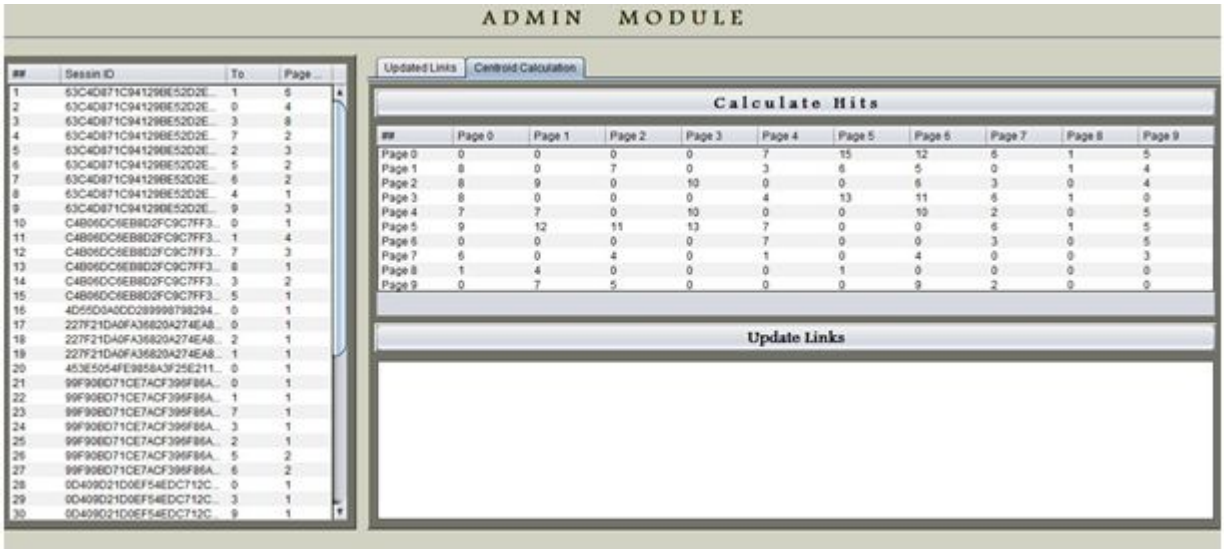

Figure 2. Count no.of hits

ISSN 2348 – 8034
Impact Factor- 3.155

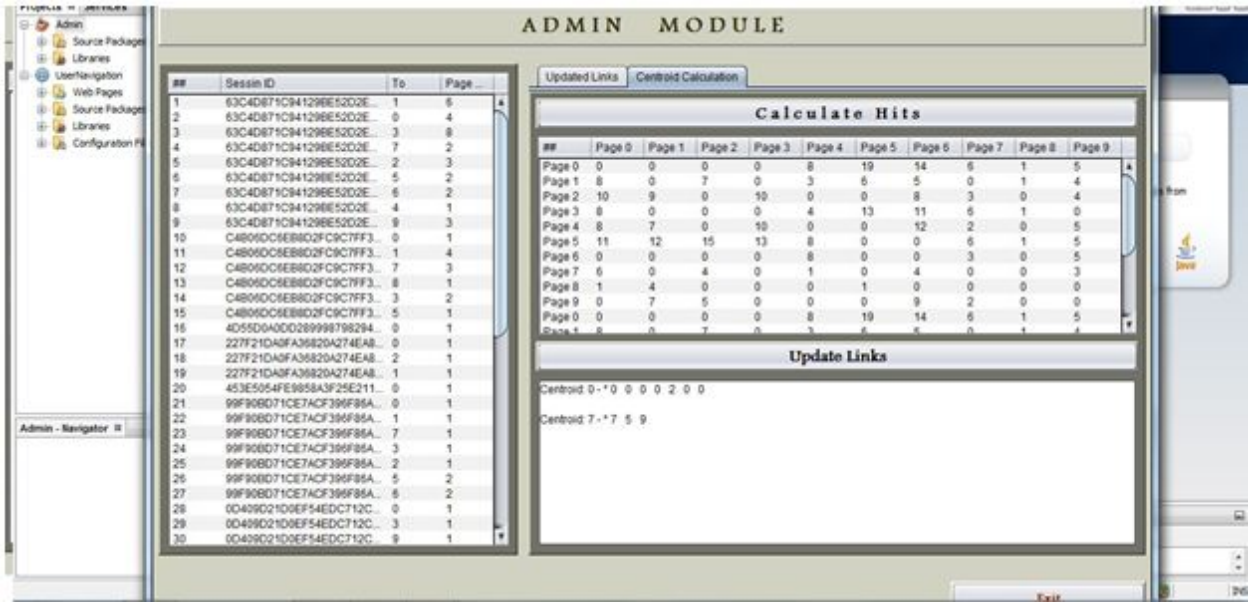

Figure3. Find Centroid input take as no of hits

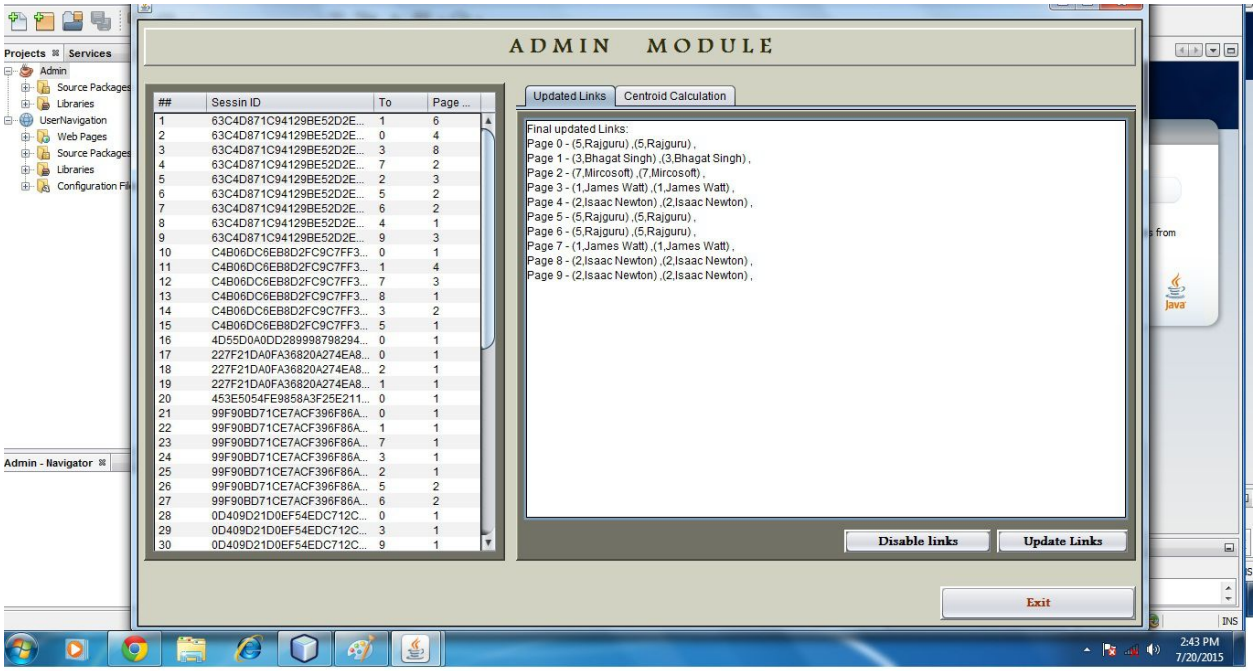

Figure4. Updated Link information as per input

ISSN 2348 – 8034
Impact Factor- 3.155

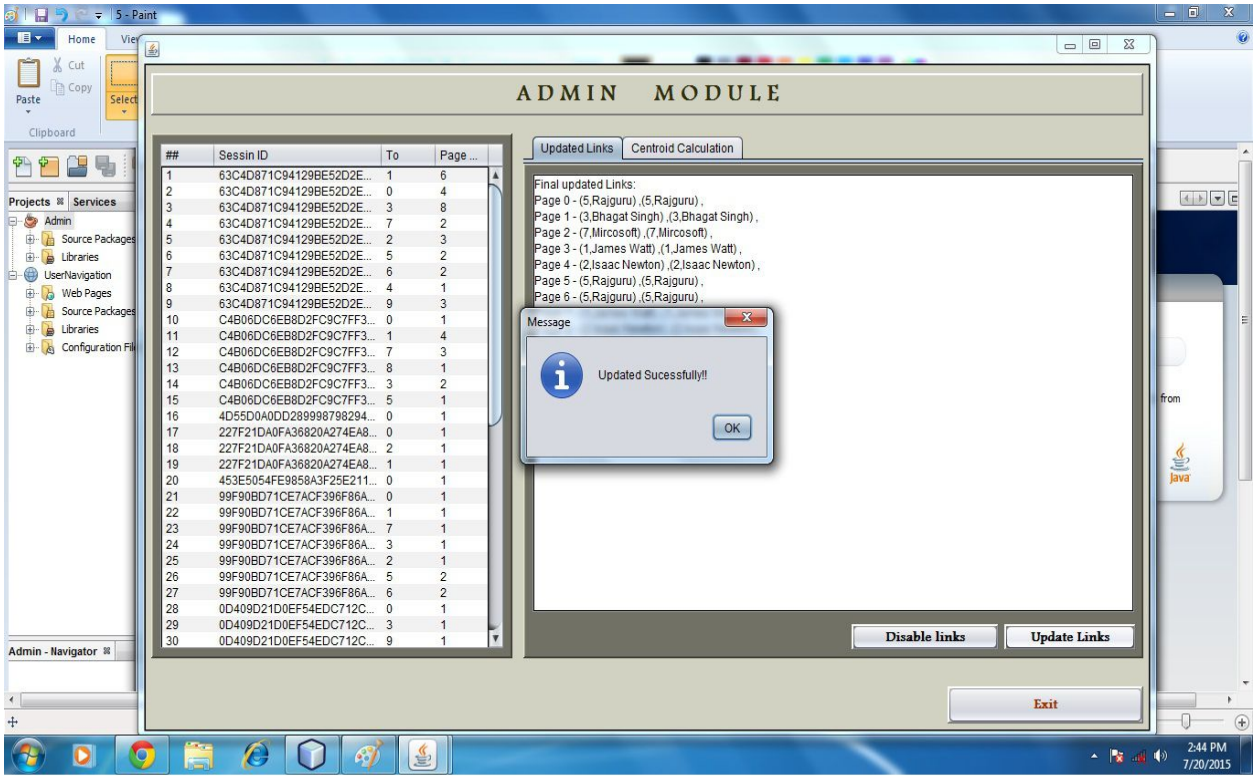

Figure 5.Update website structure Successfully

Websites	Websites Navigation Time (From - to To) Count	Retrieve Click Time	Relevant Click Time
Min Click	2	2	1
Max Click	8	6	5

	Precision	Recall
Min Click	0.5	0.5
Max Click	0.83333	0.625

Figure 2. Precision Recall

ISSN 2348 – 8034
Impact Factor- 3.155

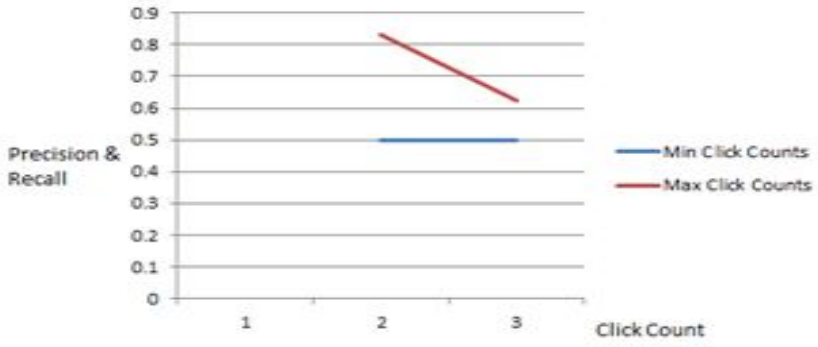

Figure 2. Precision Recall graph

VII. CONCLUSION

The Proposed system is efficient system for navigating website structure. Using the data mining algorithm i.e. Improved K means algorithm it helps to restructure the link of website. It provides the easier, fastest and better interface to retrieve information from the website. The major benefit of this system is that it performs the process of retrieving.

VIII. ACKNOWLEDGEMENTS

We are thankful to Mr. Pratap Singh Professor, Faculty of Computer Engineering, IOKCOE, Pune for her Guidance and in the successful completion this study. I would also like to thank all my colleagues who have directly or indirectly guided and helped me in the preparation of this report and also for giving me an unending support right from the stage this idea was conceived. I also acknowledge the research work done by all researchers in this field.

REFERENCES

1. *Y.Fu,M.Y.Shih,M.Credo, andC.Ju, "Reorganizing WebSites Based on UserAccess Patterns,"IntelligentSystems in Accounting,Finance andManagement, vol.11,no.1, pp.39-53,2002.*
2. *T.Nakayama,KatoAndY. Yamane,Discovering the Gap between WebSite Designers Expectations and Users Behavior, Computer Networks,vol.33,pp.811-822,2000.*
3. *M.Perkowitz and O.Etzioni,Towards Adaptive WebSites:Conceptual Framework and CaseStudy,Artificial Intelligence,vol.118,pp.245- 275,2000.*
4. *Facilitating Effective User Navigation through Website Structure Improvement, MinChenand Young U.Ryu,Knowledge an dData Engineering, Vol.25,No.3, March2013*
5. *Shehroz, Ahmad, "Cluster center initiation algorithm for K-means clustering", Pattern Recognition Letters 25, pages 1293-1302,2004*
6. *J.Palmer,WebSiteUsability, Design,and PerformanceMetrics, InformationSystems Research,vol.13,no.2,pp.151-67,2002.*
7. *J.HouandY.Zhang,Effectively Finding Relevant Web Pages from Linkage Information, IEEE Trans. Knowledge and Data Eng.,vol.15,no.4,pp.940-951,July/Aug.2003.*
8. *M.Eirinaki and M.Vazirgiannis, Web Mining for Web Personalization,ACM Trans.Internet Technology,vol.3, no.1, pp.1-27,2003*
9. *B M AhamedShafeeq, K S Hareesha, " Dynamic Clustering of Data with Modified K-Means Algorithm", International Conference on Information and Computer Networks (ICICN 2012), IPCSIT, vol. 27,pages 221-225, 2012*
10. *Mohammed EI Agha, Wesam M. Ashour, " Efficient and Fast Initializtion Algorithm for K-means Clustering" Intelligent Systems and Applications, vol. 4, issue 1,pages 21-31, 2012*