An Approach for Improved Website Structure using Smart K-Means

Shukla Bharati[1],Wakhare Sagar[2], Khule Sujit[3],Deokate Vanita[4], Prof. Deepti Varshney[5]
[1, 2, 3, 4, 5]Department of Computer Science
[1, 2, 3, 4, 5]Shree Ramchandra College of Engineering ,India

Abstract-*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. Hence an optimal selection of clusters is significant step in implementation of the system. 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 a arithmetical programming model to recover the user navigation on a website. The system is imagined the deliver the functionality of a test bench website for data collection and then reorder it based on statistics collected to present the effectiveness of our model.*

Keywords-Clustering,Websitedesign, Usernavigation, Web mining.

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.

II. PROBLEM STATEMENT

The network security is the main problem now a days.To provide a data security there isneed of some extra security primitives. We are tackling this problem by using Armstrong numbers and color method.

III. MOTIVATIONS

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

IV. OBJECTIVES

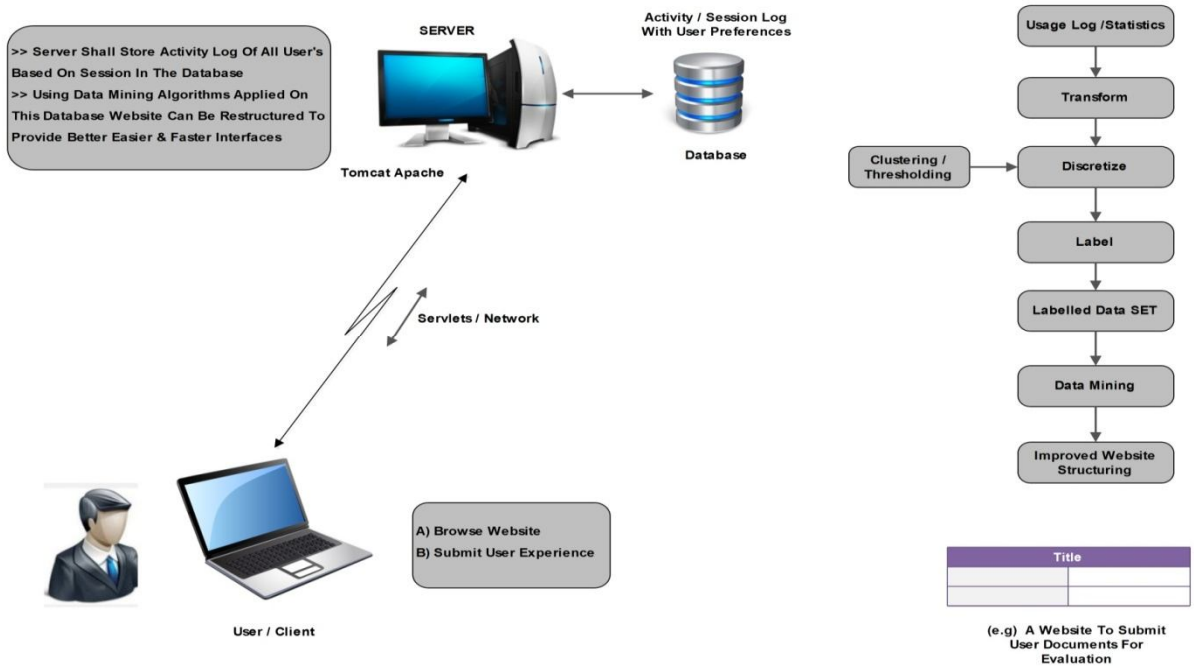
To provide the Improved Web Site Structure Using K-means Algorithem.

.

Existing System

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, WebPagesshould be structured in a way that generally matches the user's 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. 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 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- Means algorithm. It is a simple and easy way to classify a given data set through a certain number of clusters.



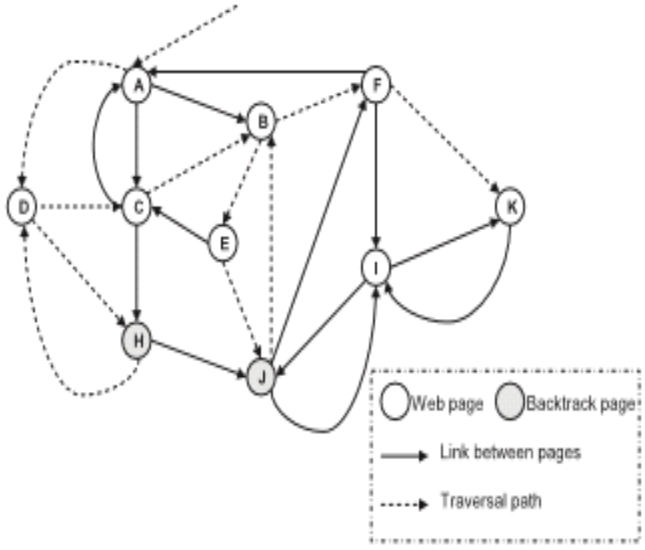


Fig3 . Example of a mini session.

Our model allows Webmasters to specify a goal for user navigation that the improved structure should meet. This goal is associated with individual target pages and is defined as the maximum number of paths allowed to reach the target page in a mini session. We term this goal the path threshold for short in this paper. In other words, in order to achieve the user navigation goal, the website structure must be altered in a way such that the number of paths needed to locate the targets in the improved structure is not larger than the path threshold.

In the example shown in Fig. 2, the user has traversed three paths before reaching the target. An intuitive solution to help this user reach the target faster is to introduce more links . There are many ways to add extra links. If a link is added from D to K, the user can directly reach K via D, and hence reach the target in the first path. Thus, adding this link "saves" the user two paths. Similarly, establishing a link from B to K enables the user to reach the target in the second path. Hence, this saves him one path. We could also insert a link from E to K, and this is considered the same as linking B to K. This is because both B and E are pages visited in the second path, so linking either one to K saves only one path. Yet, another possibility is to link C to F, a nontarget page. In this case, we assume that the user does not follow the new link, because it does not directly connect a page to the target. While many links can be added to improve navigability, our objective is to achieve the specified goal for user navigation with minimal changes to a website. We measure the changes by the number of new links added to the current site structure. There are several reasons that we should insert minimal links. First, minimizing changes to the current structure can avoid disorienting familiar users. Second, adding unnecessary links can lead to pages having too many links, which increases users' cognitive loads and makes it difficult for them to read and comprehend . Third, since our model improves site structures on a regular basis, the number of new links should be kept at minimum such that the links in a website in the whole course of maintenance do not expand in a chaotic manner.

V. PROPOSED SYSTEM

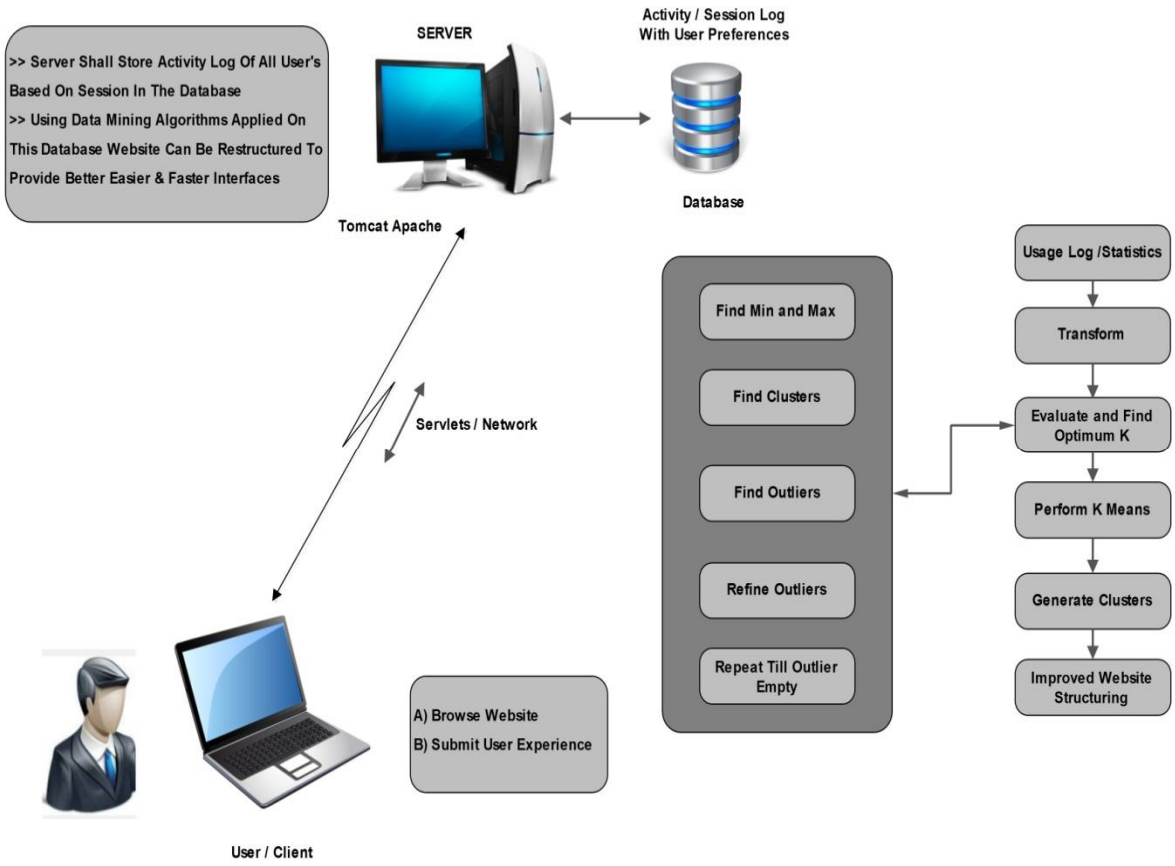
When User want Surfing on Internet that time user did not get actual information he want He/she has to spend a lot of time on that particular web site .In this Paper we propose a new Data mining algorithm k means i.e. improved k-means algorithm . This Improved K-means Work on at database of web server .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. 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 reduce time complexity over web.

VI. SYSTEM ARCHITECTURE

The diagram shown above shows the proposed system. The system consists of end users/client, a server for storing the data.

- 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.



ALGORITHM USED:

Smart K-mean:

(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 theinitial 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:

(7) If Distance of the tuple from the cluster mean<d then not an Outlier.

(8) Compute new centroids of the clusters.

(9) Calculate Euclidean distance of every outlier from the new cluster centroids and fmd the outliers not satisfying theobjective function in step 6.

(10) Let B={Yl,Y2,. ....Y p) be the set of outliers obtained in step 8 (value of k depends on number of outliers).

Repeat unti 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 objectivefunction as in step 6.

ii) Find the outliers if any

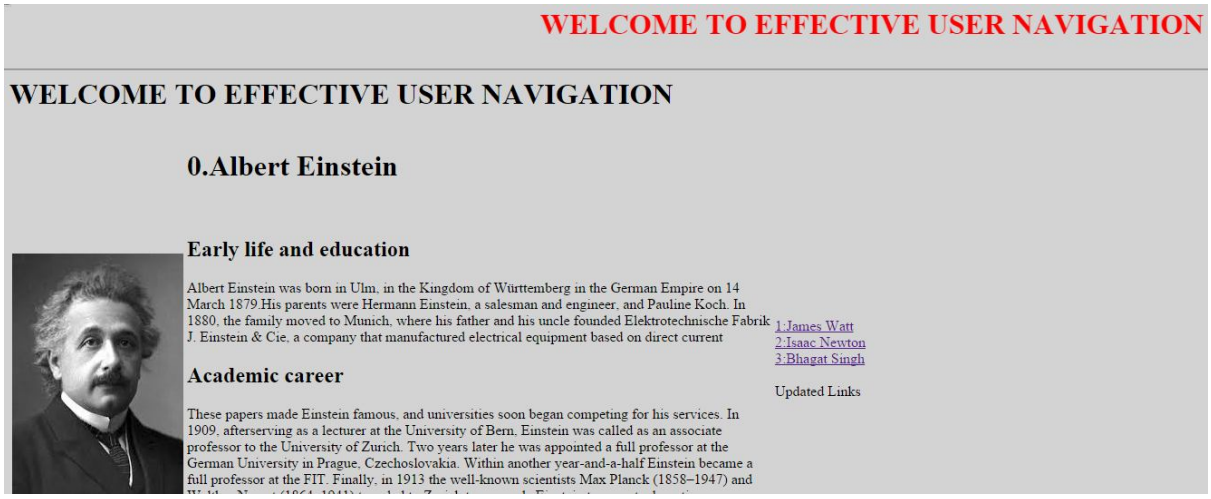
d) Calculate the distance of every outlier from the centroid of the existing clusters and adjust the outliers in theexisting which satisfy the objective function in step 6.

e) B={ ZI,Z2 .... Zq}be the new set of outliers.(value of q depends on number of outliers)

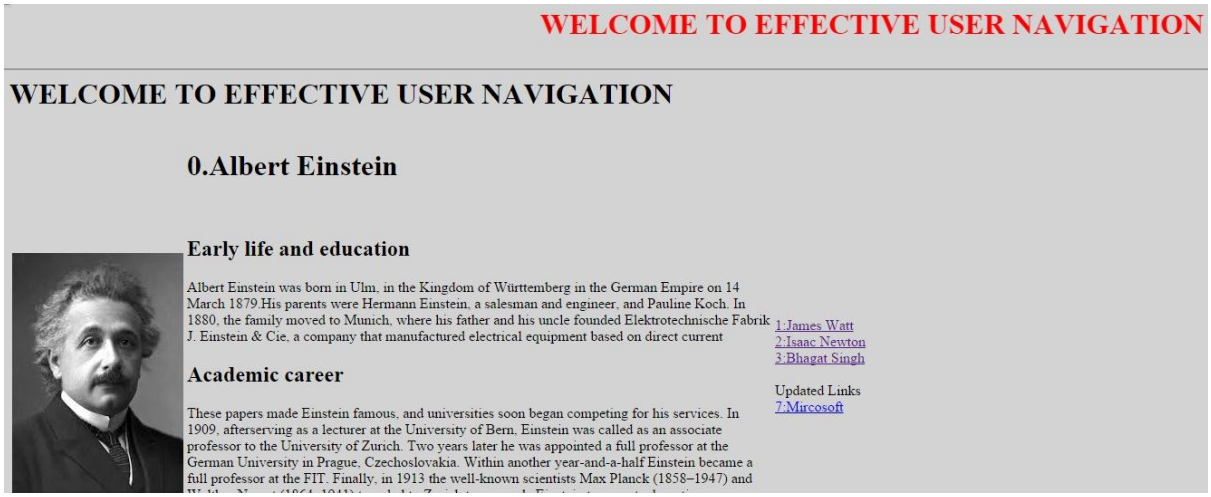
VII. RESULT COMPARISON OF EXISTING AND PROPOSED SYSTEM

EXISTING SYSTEM

Normal link on website:

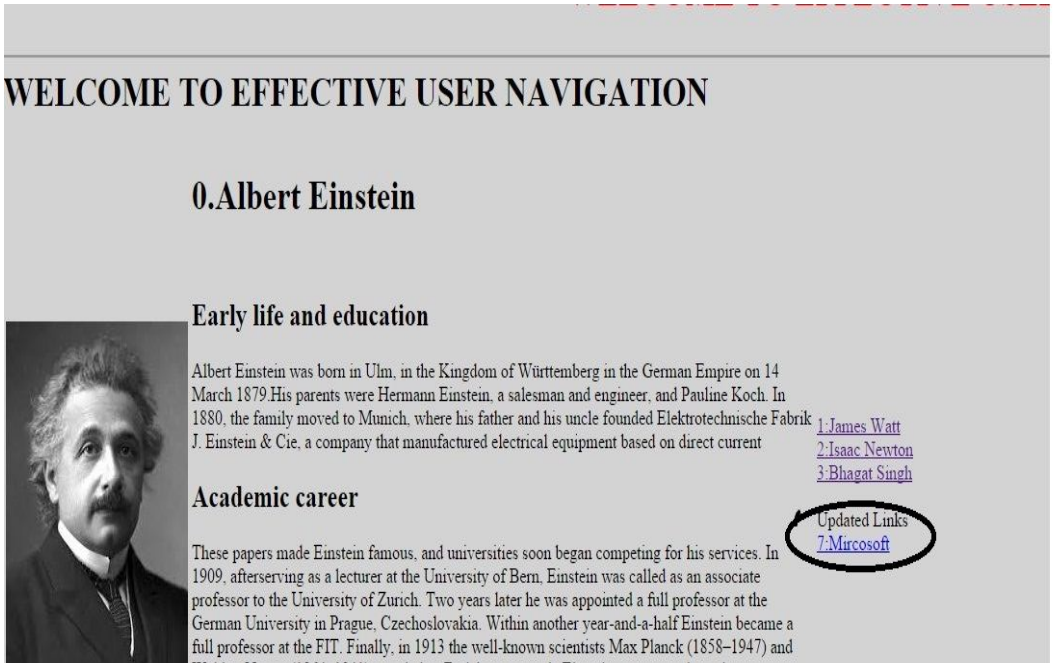


Normal link on website1:

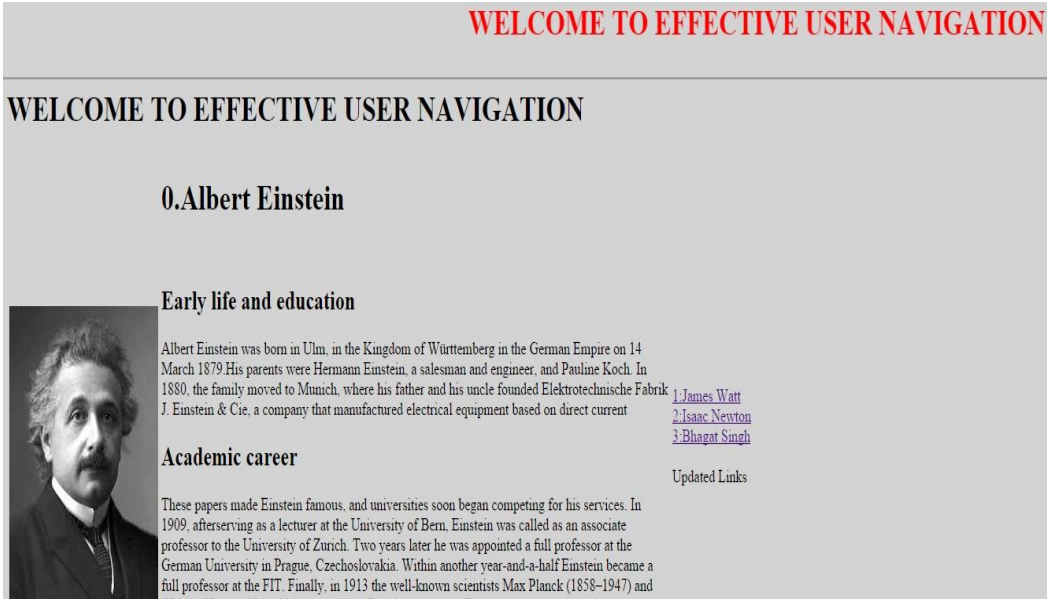


PROPOSED SYSTEM

Final updated link:



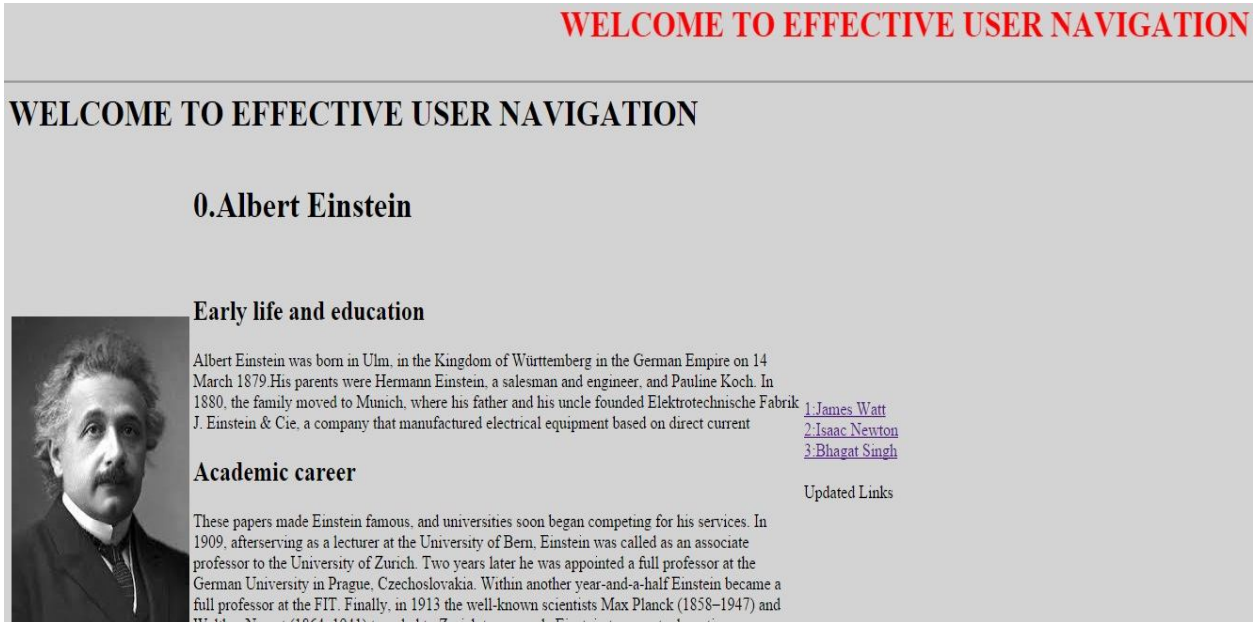
First page screenshot:



VIII. RESULT ANALYSIS

Screenshot 1:

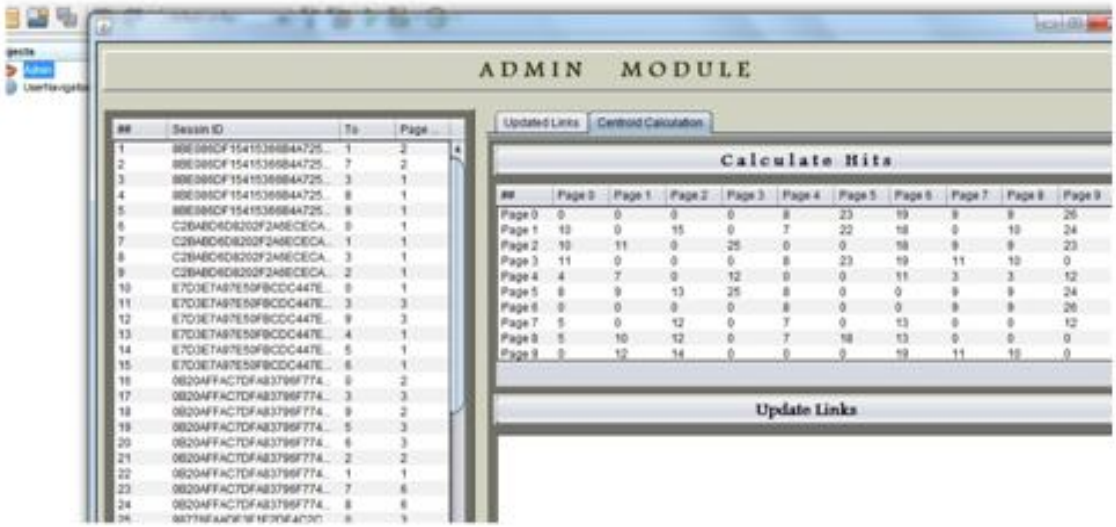
Here we can see the normal web page with some link and there one link i.e. Updated link
But it is initially no link present.



Screenshot 2:

Here in this window we can see number of hits. That mean when visit any link on website like we dick

On link that is the hit for that link.so that count of that hit store on this module.



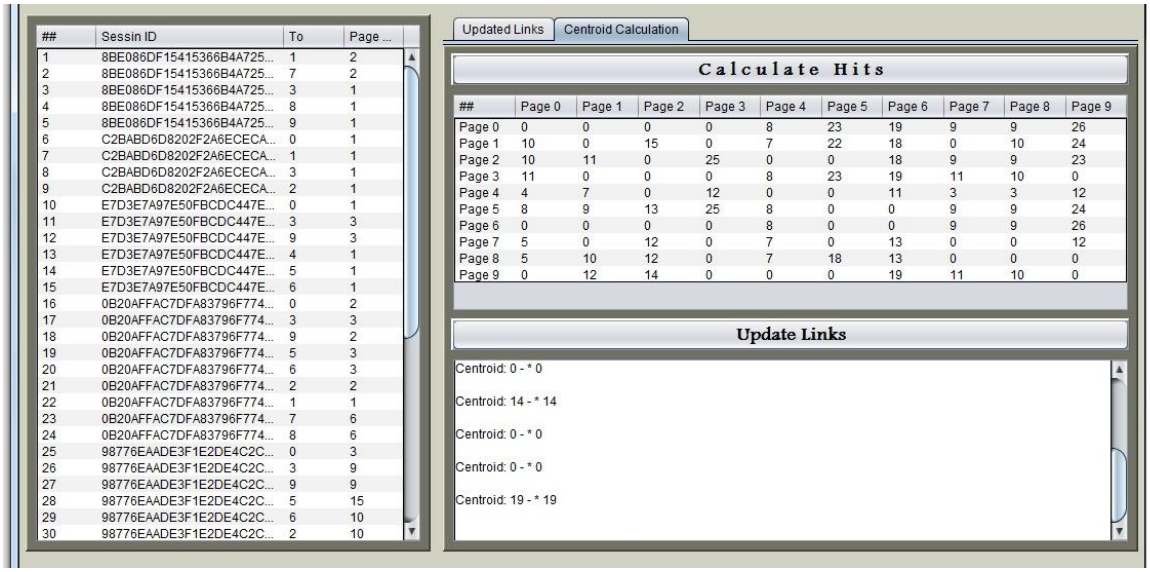
Screenshot 3:

This is the first window from the admin side only admin can see it .if admin use this with help of user name and password for security purpose.

Screenshot 4:

In previous screen we see the no of hit store in that module, here we used this no of hit as input to

The k-means algorithm. With help of that algorithm we find centroid of that data set.

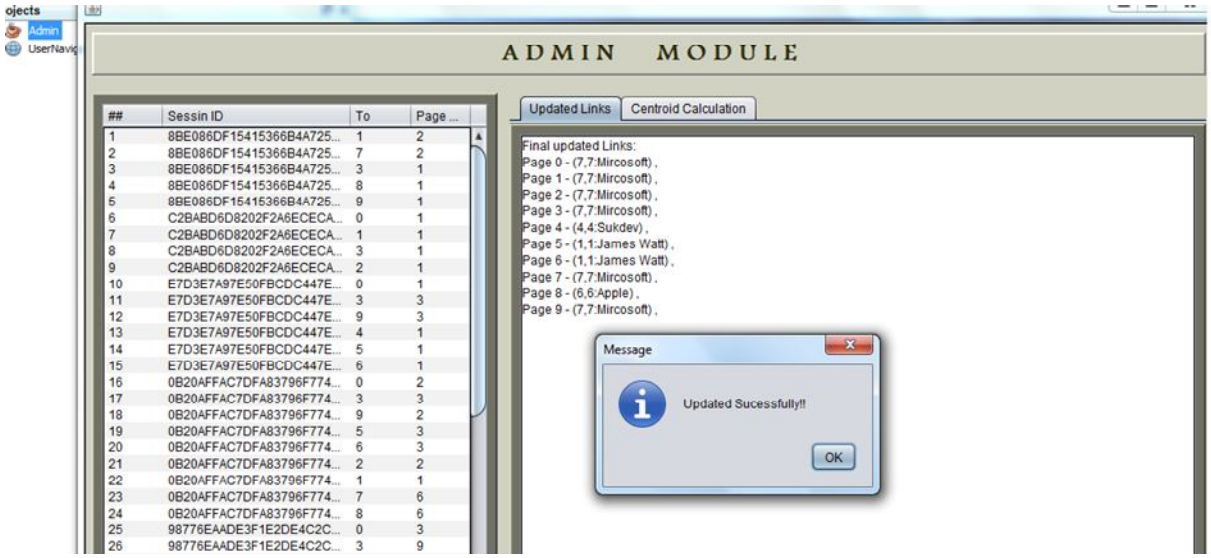


##	Sessin ID	To	Page ...
1	8BE086DF15415366B4A725...	1	2
2	8BE086DF15415366B4A725...	7	2
3	8BE086DF15415366B4A725...	3	1
4	8BE086DF15415366B4A725...	8	1
5	8BE086DF15415366B4A725...	9	1
6	C2BABD6D8202F2A6ECECA...	0	1
7	C2BABD6D8202F2A6ECECA...	1	1
8	C2BABD6D8202F2A6ECECA...	3	1
9	C2BABD6D8202F2A6ECECA...	2	1
10	E7D3E7A97E50FBCDC447E...	0	1
11	E7D3E7A97E50FBCDC447E...	3	3
12	E7D3E7A97E50FBCDC447E...	9	3
13	E7D3E7A97E50FBCDC447E...	4	1
14	E7D3E7A97E50FBCDC447E...	5	1
15	E7D3E7A97E50FBCDC447E...	6	1
16	0B20AFFAC7DFA83796F774...	0	2
17	0B20AFFAC7DFA83796F774...	3	3
18	0B20AFFAC7DFA83796F774...	9	2
19	0B20AFFAC7DFA83796F774...	5	3
20	0B20AFFAC7DFA83796F774...	6	3
21	0B20AFFAC7DFA83796F774...	2	2
22	0B20AFFAC7DFA83796F774...	1	1
23	0B20AFFAC7DFA83796F774...	7	6
24	0B20AFFAC7DFA83796F774...	8	6
25	98776EAADE3F1E2DE4C2C...	0	3
26	98776EAADE3F1E2DE4C2C...	3	9
27	98776EAADE3F1E2DE4C2C...	9	9
28	98776EAADE3F1E2DE4C2C...	5	15
29	98776EAADE3F1E2DE4C2C...	6	10
30	98776EAADE3F1E2DE4C2C...	2	10

##	Page 0	Page 1	Page 2	Page 3	Page 4	Page 5	Page 6	Page 7	Page 8	Page 9
Page 0	0	0	0	0	8	23	19	9	9	26
Page 1	10	0	15	0	7	22	18	0	10	24
Page 2	10	11	0	25	0	0	18	9	9	23
Page 3	11	0	0	0	8	23	19	11	10	0
Page 4	4	7	0	12	0	0	11	3	3	12
Page 5	8	9	13	25	8	0	0	9	9	24
Page 6	0	0	0	0	8	0	0	9	9	26
Page 7	5	0	12	0	7	0	13	0	0	12
Page 8	5	10	12	0	7	18	13	0	0	0
Page 9	0	12	14	0	0	0	19	11	10	0

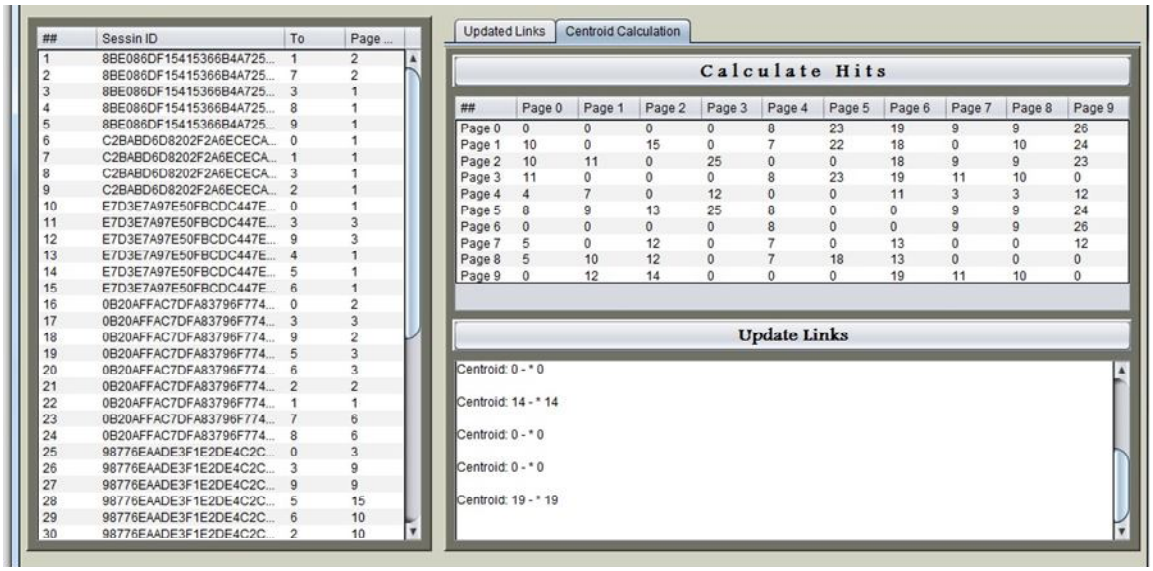
Screenshot 5:

With help of k-mean algorithm we find the cluster with centroid .According to cluster here reconstruct the link of web pages or automatically update the link.



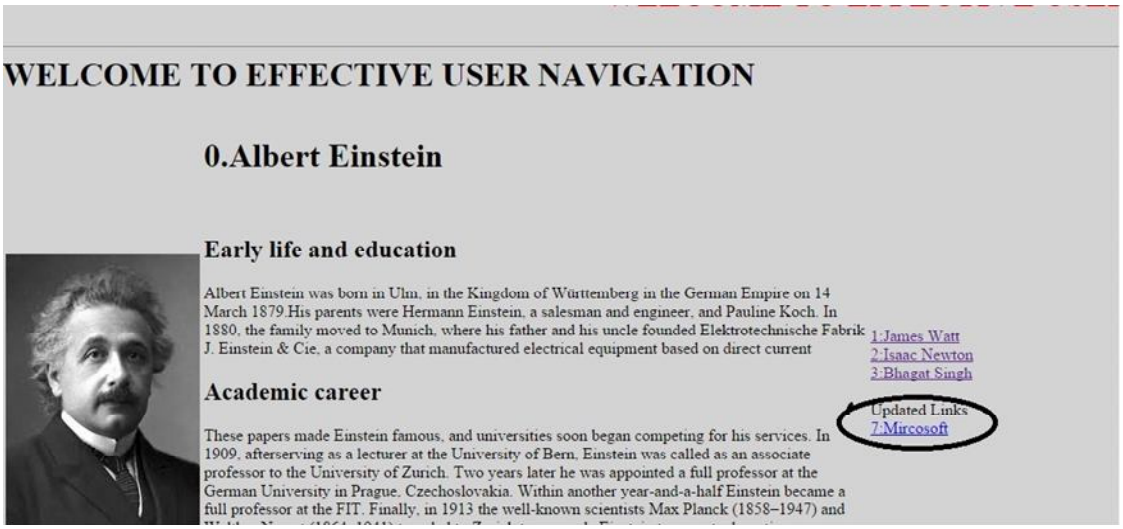
In previous screen we see the no of hit store in that module, here we used this no of hit as input to

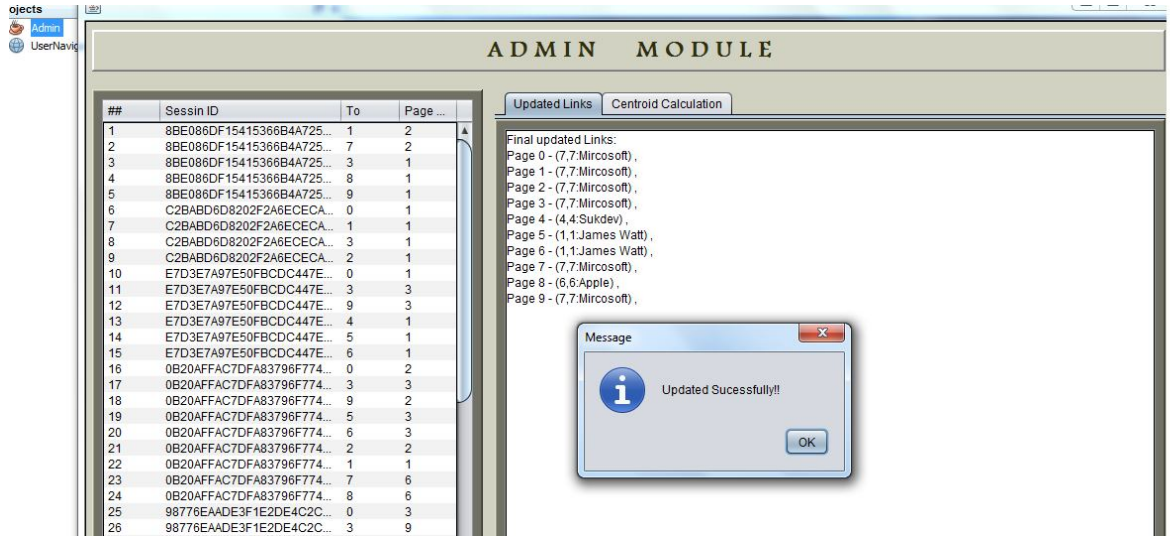
The k-means algorithm. With help of that algorithm we find centroid of that data set.



After the updating link these updated link which is present on first page of web site. This is automatically

Construction of link is done with help of k-means algorithm.

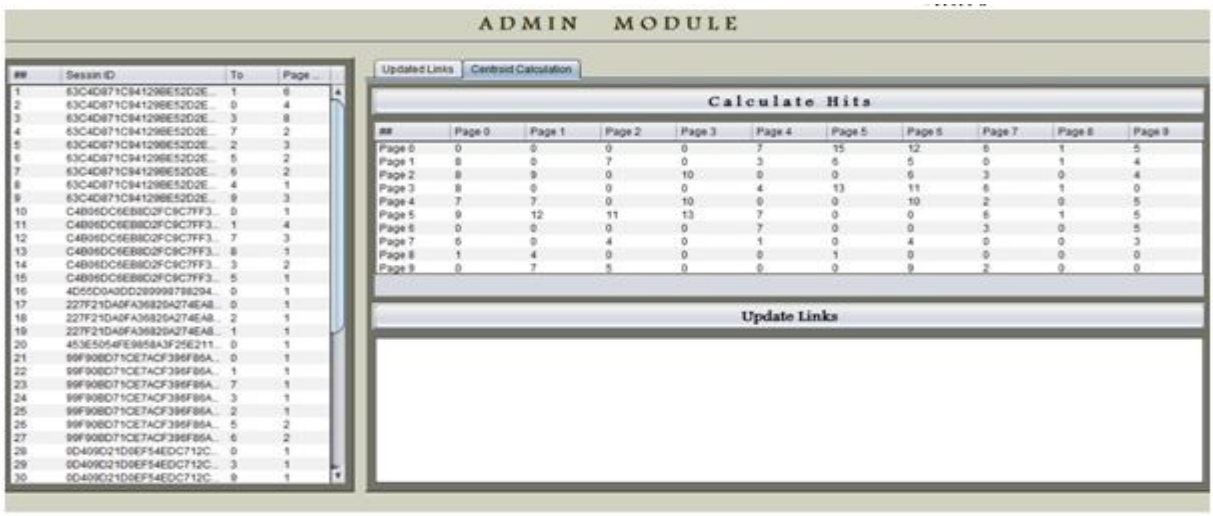


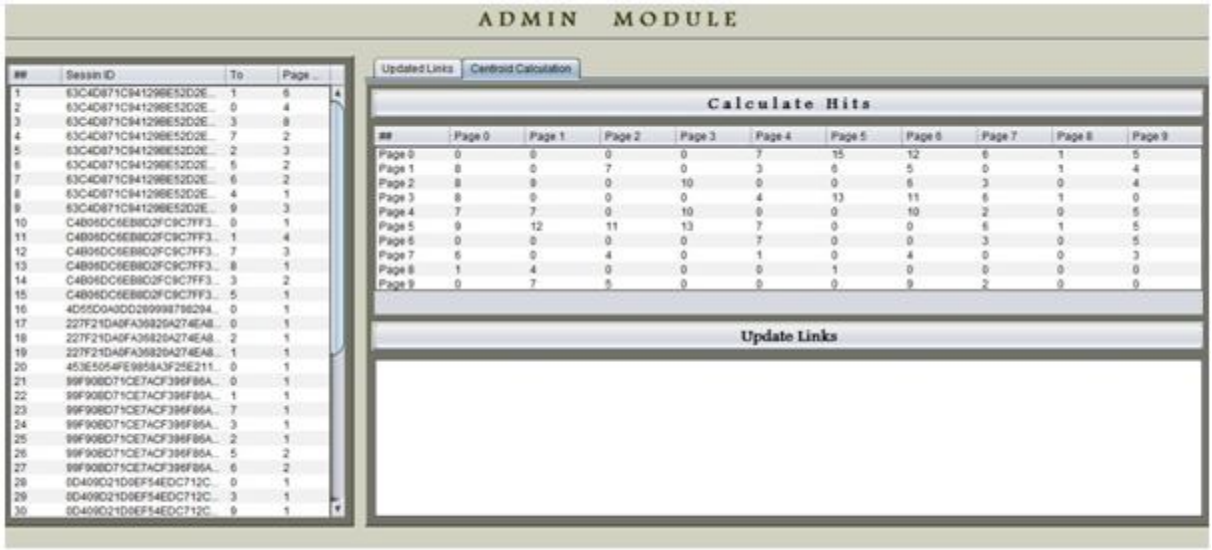


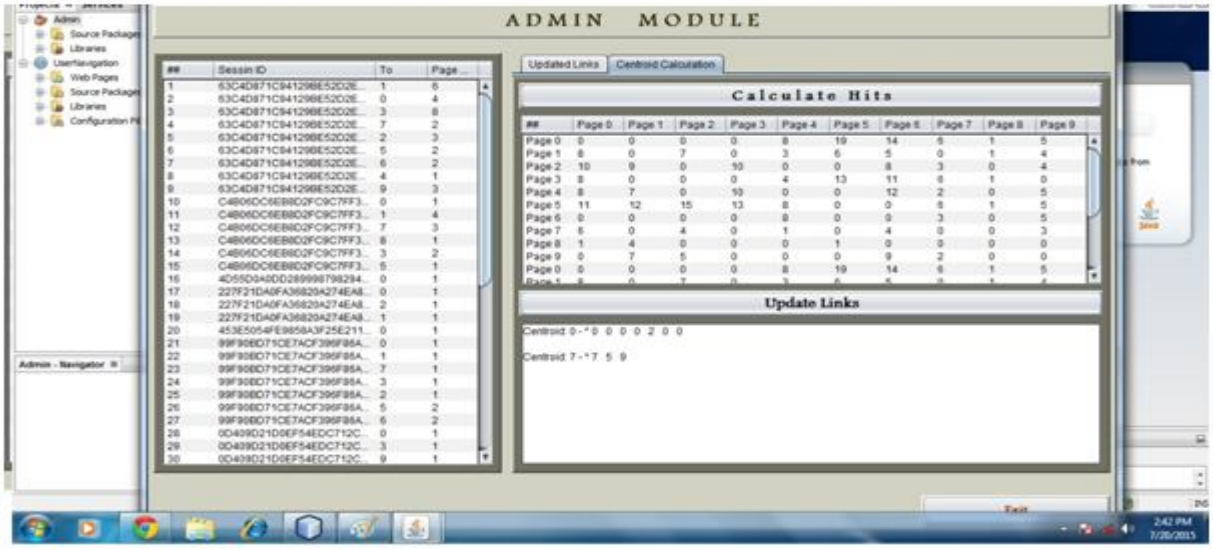
##	Sessin ID	To	Page ...
1	8BE086DF15415366B4A725...	1	2
2	8BE086DF15415366B4A725...	7	2
3	8BE086DF15415366B4A725...	3	1
4	8BE086DF15415366B4A725...	8	1
5	8BE086DF15415366B4A725...	9	1
6	C2BABD6D8202F2A6ECECA...	0	1
7	C2BABD6D8202F2A6ECECA...	1	1
8	C2BABD6D8202F2A6ECECA...	3	1
9	C2BABD6D8202F2A6ECECA...	2	1
10	E7D3E7A97E50FBCDC447E...	0	1
11	E7D3E7A97E50FBCDC447E...	3	3
12	E7D3E7A97E50FBCDC447E...	9	3
13	E7D3E7A97E50FBCDC447E...	4	1
14	E7D3E7A97E50FBCDC447E...	5	1
15	E7D3E7A97E50FBCDC447E...	6	1
16	0B20AFFAC7DFA83796F774...	0	2
17	0B20AFFAC7DFA83796F774...	3	3
18	0B20AFFAC7DFA83796F774...	9	2
19	0B20AFFAC7DFA83796F774...	5	3
20	0B20AFFAC7DFA83796F774...	6	3
21	0B20AFFAC7DFA83796F774...	2	2
22	0B20AFFAC7DFA83796F774...	1	1
23	0B20AFFAC7DFA83796F774...	7	6
24	0B20AFFAC7DFA83796F774...	8	6
25	98776EAADE3F1E2DE4C2C...	0	3
26	98776EAADE3F1E2DE4C2C...	3	9

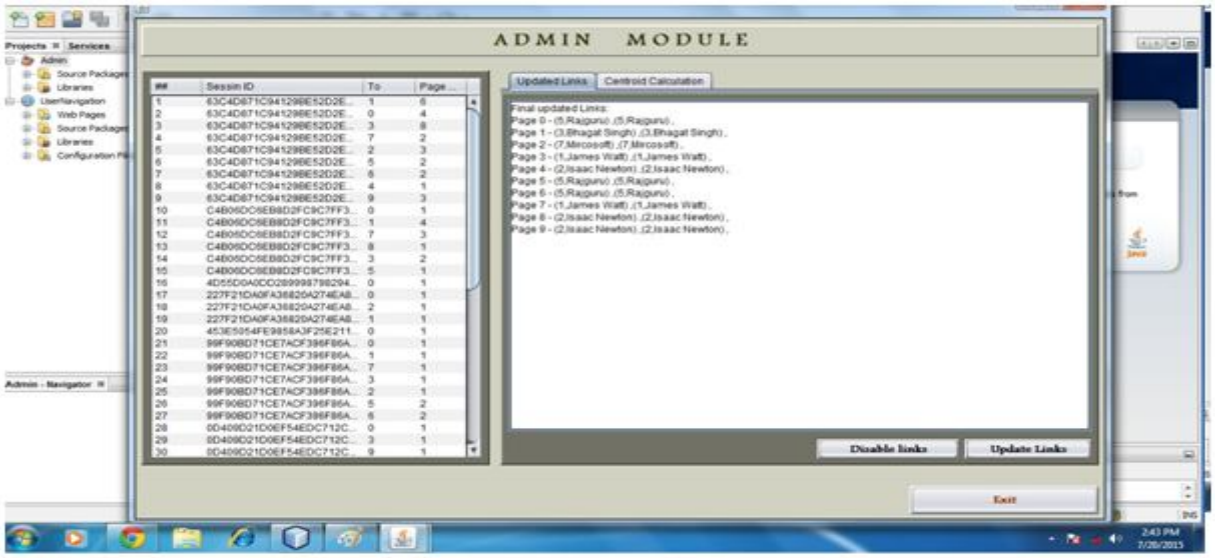
Steps to run software:

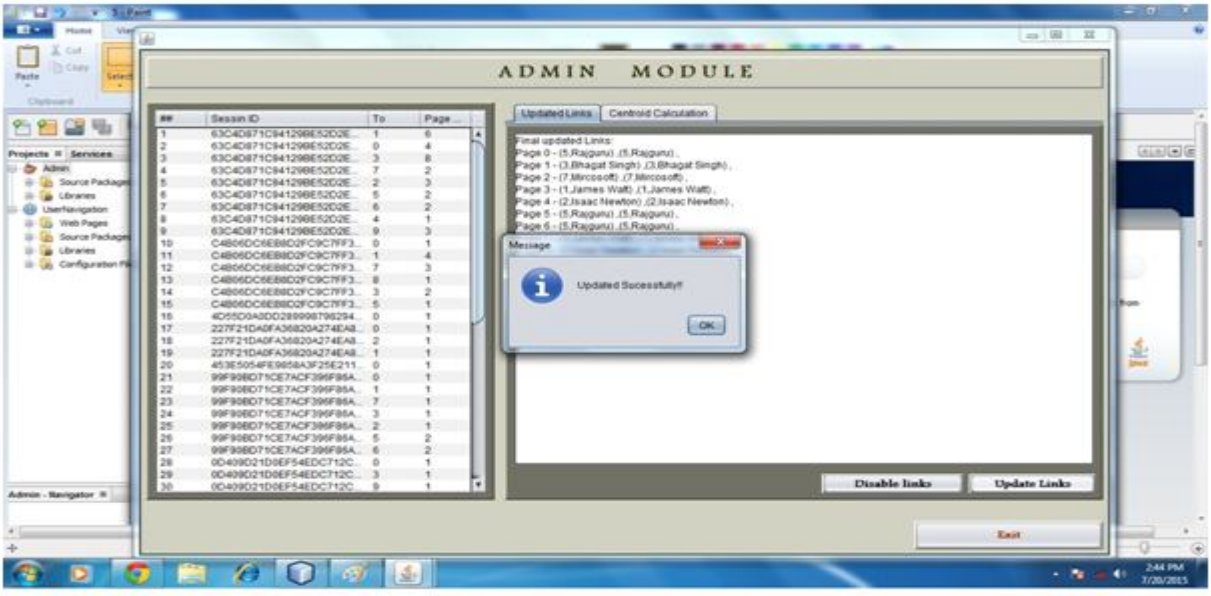
1. Login procedure of admin module :

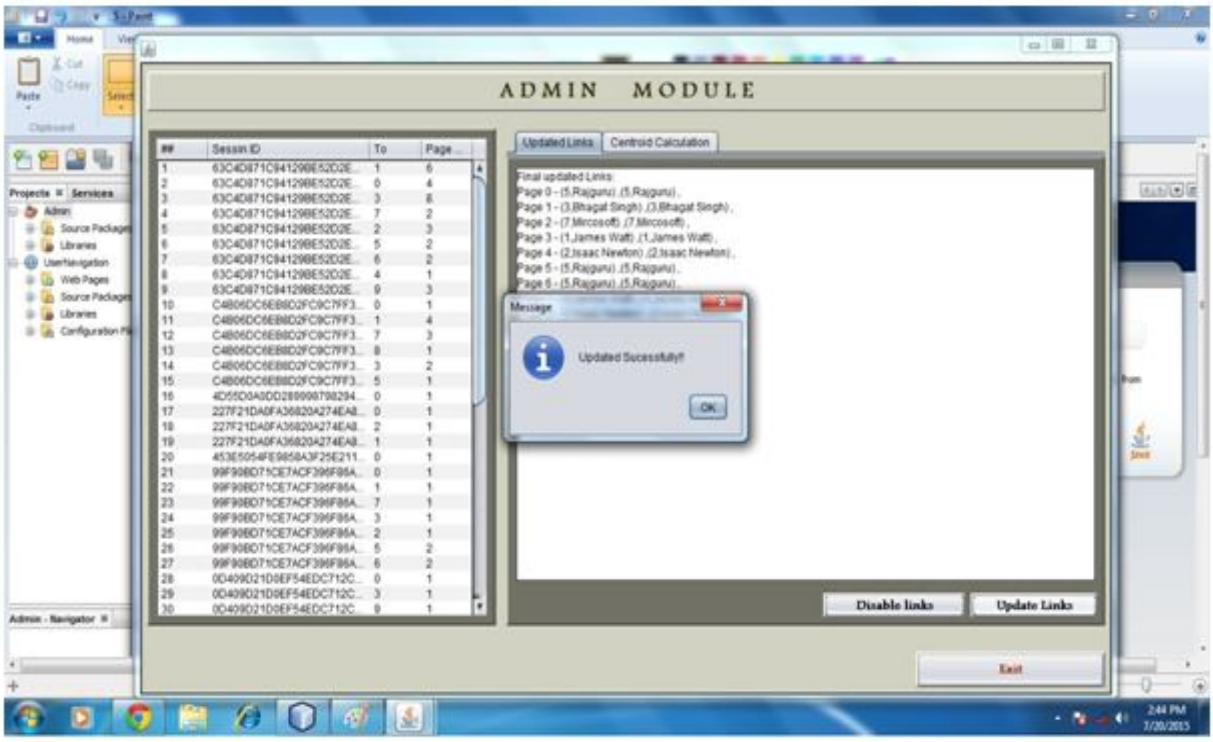




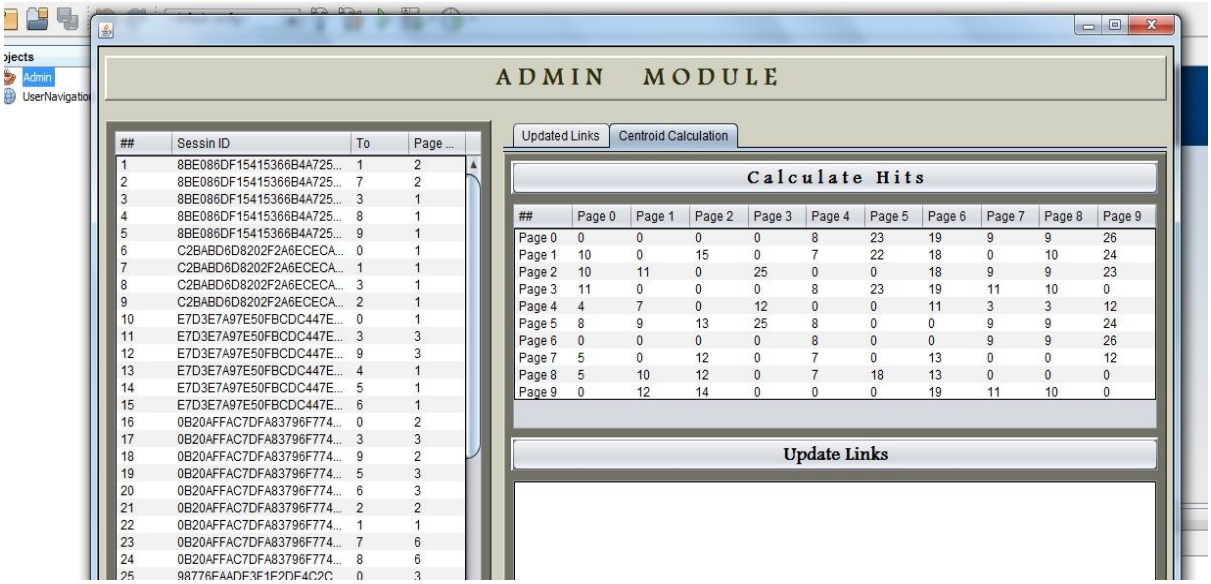

ADMIN MODULE
Updated Links
Centroid Calculation
Calculate Hits
Update Links
Exit


ADMIN MODULE
Updated Links
Centroid Calculation
Disable links
Update Links
Exit


ADMIN MODULE
Updated Links
Centroid Calculation
Message
Updated Sucessfully!!
OK
Disable links
Update Links
Exit

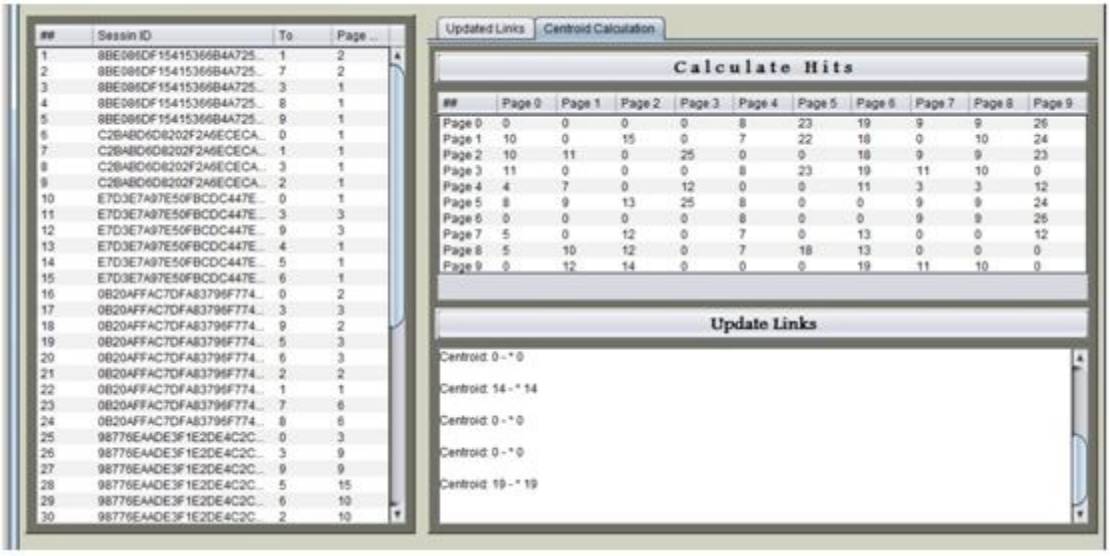


2. Calculate no. of hit:

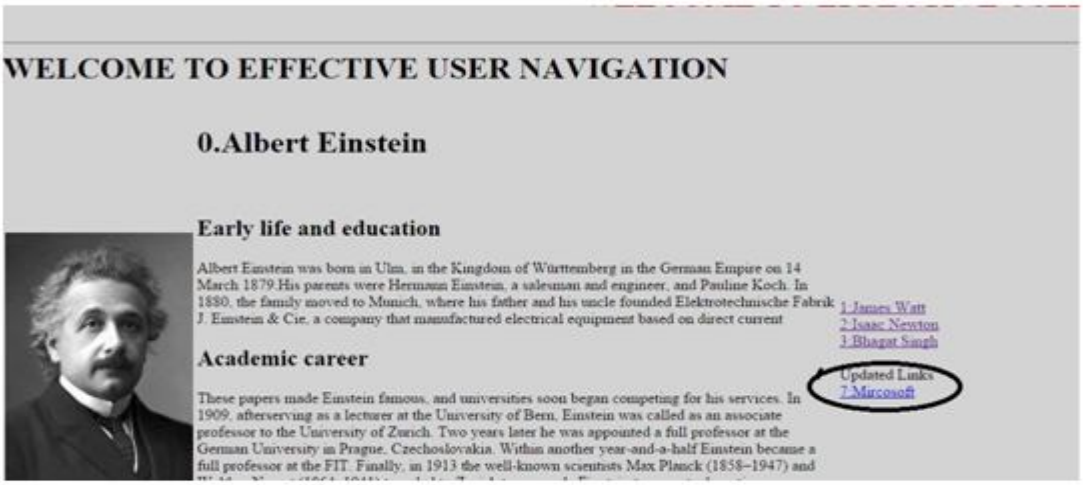


##	Page 0	Page 1	Page 2	Page 3	Page 4	Page 5	Page 6	Page 7	Page 8	Page 9
Page 0	0	0	0	0	8	23	19	9	9	26
Page 1	10	0	15	0	7	22	18	0	10	24
Page 2	10	11	0	25	0	0	18	9	9	23
Page 3	11	0	0	0	8	23	19	11	10	0
Page 4	4	7	0	12	0	0	11	3	3	12
Page 5	8	9	13	25	8	0	0	9	9	24
Page 6	0	0	0	0	8	0	0	9	9	26
Page 7	5	0	12	0	7	0	13	0	0	12
Page 8	5	10	12	0	7	18	13	0	0	0
Page 9	0	12	14	0	0	0	19	11	10	0

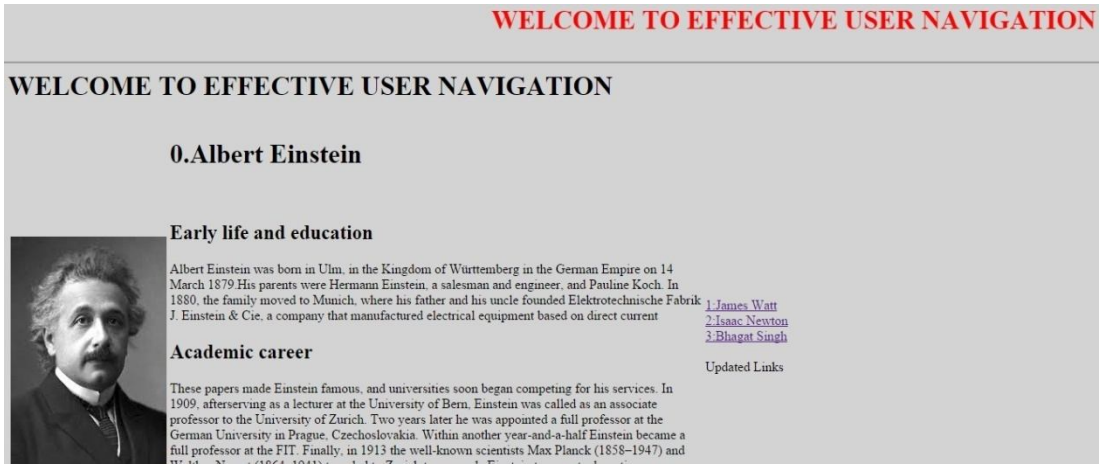
3. Calculate centroid:



4. Final updated link:



5. First page screenshot;



IX. 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 information in minimum span of time so that we can say it is time efficient.

X. FURTHER WORK

The limitation with proposed system if we select the number containing zero then result produced is not proper . As a future scope we will work to remove this limitations.

ACKNOWLEDGEMENTS

We are presenting this paper with great pride. There are too much efforts ofgardener to yield the beautiful owners . So we should not forget him while praising owners.It is a matter of gratification for we to pay our respects and acknowledgments to allthose who have imparted knowledge and helped we to complete my paper. Wewould first like to acknowledge the great contribution and support we have received inthis endeavor from my Guide Prof. DeeptiVarshney.We are very grateful and indebted to the project guide for providing her enduring patience, guidance and invaluable suggestions. she was the one who neverlet our moral down and always supported we through us thick and thin.

We would also like to thank all the Staff members for their invaluable co-operation and permitting we to work in the Computer labs of SRCOE.

REFERENCES

[1] T.Nakayama,KatoAndY. Yamane,Discovering the Gap between WebSite Designers Expectations and Users Behavior, Computer Networks,vol.33,pp.811-822,2000.

[2] M.Perkowitz and O.Etzioni,Towards Adaptive WebSites:Conceptual Framework and CaseStudy, ArtificialIntelligence,vol.118,pp.245- 275,2000.

[3] Y.Fu,M.Y.Shih,M.Credo, andC.Ju, ”Reorganizing WebSites Based on UserAccess Patterns,” Intelligent Systemsin Accounting, Financeand Management, vol.11,no.1, pp.39-53,2002.

[4] 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.

[5] M.Eirinaki and M.Vazirgiannis, Web Mining for Web Personalization,ACMTrans.Internet Technology,vol.3, no.1, pp.1-27,2003.

[6] Shehroz, Ahmad, ”Cluster center initiation algorithm for K-means clustering”, Pattern Recognition Letters 25,pages 1293-1302,2004.

[7] 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.

[8] Facilitating Effective User Navigation through Website Structure Improvement, MinChenand YoungU.Ryu,Knowledge an dData Engineering, Vol.25,No.3, March2013

[9] Min Chen and Young U. Ryu” Facilitating Effective User Navigation through Website Structure Improvement”
IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 3, MARCH 2013

[10] Anupama Chadha, Suresh Kumar” An Improved K-Means Clustering Algorithm: A Step Forward for Removalof Dependency on K” 2014 International Conference on Reliability, Optimization and Information Technology