Enhancing User Navigation Using Web Transformation

[1]Madhuri B. Aswale, [2]Balasaheb B. Gite
[1]Student, [2]Assistant Professor
[1]Computer Engineering,
[1]SAE Kondhwa, Pune, India

Abstract **– There are billions of internet users worldwide. This increase in the use of internet leads the researcher, service provider to use this source of information. Discovery of required information in a website is a difficult task. Difficulty in user navigation has long been a challenge because while creating website web developers understanding regarding the structure of website can be different from that of user's preferences. Such circumstances results in difficulty to find desired information on website. However, the measure of website effectiveness should be the satisfaction of the users rather than that of the developers. Various methods have been proposed to improve navigation using users usage data to reorganise the structure but results are not effective. It is important to improve the navigation of website because they are mostly use for collecting information and for business purpose also, there is a need to improve the efficiency and the performance of a website. In this paper, we propose an optimal solution to improving the navigation of website also to find the user target accurately. Our approach presents web mining methods and techniques for the evaluation of systems in order to improve the efficiency of web site. For improvement here two algorithms are used k-means clustering algorithm to perform clustering Apriori algorithm to find frequent traversal path of user. Our aim is to enhance the functionality of website with small changes in website structure.**

Index Terms **- Data Mining; Web mining; Website Design; Web Personalization; Web Transformation.**

I. INTRODUCTION

Internet is widely used by people around the world. Large amount of information is available on internet. Lots of people share information using social site like Facebook which connect individuals to each other. This increase in the use of internet leads the researcher, service provider to use this source of information. Discovery of required information in a website are a difficult task [2]. If the user is not able to find their useful information, then the user exits the website even containing more quality information. The process of obtaining useful information from Web is known as web mining. As this web mining used to solve problems related to information retrieval and uses various data mining techniques to extract information from websites. Motivation to choose this idea is that website is a reach source of information and due to poor website design it is difficult to find desired information on website for users. There are two ways to improve website navigation: Web personalization and Web Transformation. There are some differences between web transformation [3] and personalization [4] approaches. Web personalization uses information of individual user and web logs for created by that particular user, it is more useful for dynamic website while Web transformation approach focuses on improving structure of website which can be used by all users and suitable for static website. Web personalization is a time consuming process than transformation.

In this paper web transformation approach is used to improve user navigation. Previous work on transformation done regarding to complete reorganization of website. But it has some disadvantages like new website may be disorient users [5] and it is highly unpredictable from economical point of view. So, here we proposed a method to improve website structure instead of reorganize it. So that web mining techniques can be used to solve the problem which describe above. Web mining can be defined as the automated discovery and analysis of useful information from the web documents. Data mining techniques are used to perform web usage mining i.e. to extract user web log data and to find links that need to improve for effective navigation [6]. This method is particularly useful for informational websites whose contents are static and relatively stable over time.

The rest of the paper is organized as follows: Section 2 Reviews related work. Section 3 presents existing system. Section 4 discusses issues related to existing system. Section 5 introduces proposed system. Section 6 presents mathematical representation. Section 7 discusses evaluation method and Section 8 concludes the paper.

II. LITERATURE SURVEY

Previous studies done to improve navigation on website are related to two different approaches Web personalization and web transformation. Web personalization is the process of finding webpages for particular user on basis of user's information. Perkowitz and Etzioni [3] describe an approach related to Web personalization in which index page contains links to particular pages based on the frequency of user traversal. The method describe by Mobasher et al. and Yan et al. [7] uses clustering to retrieve weblogs, these clusters of user profile dynamically generate links for users according to their access pattern.

Nakagawa and Mobasher [8] proposed a method in which user location is find out for that degree of connectivity is calculated. B. Mobasher, Honghua Dai, Tao Luo, Miki Nakagawa [9] develops techniques also based on clustering. This method generates collective profiles which can be used for recommendation system. It is used to improve web personalization by using user profile.

B. Mobasher, R.Cooley, J. Shrivastava presents [10] an approach where many of web mining techniques and activities are used for web personalization, they propose automatic and dynamic web personalization. B. Mobesher, R. Cooley and J. Shrivastava developed an approach where they collect information about users when users are offline as well and mine that data. In this approach association rule discovery and clustering is used to gather user profiles.
Web transformation is the process of changing structure of website to improve navigation. Fu et al. [11] develop a method to reorganize Webpages to give required information to users in few clicks. However, this method considers only local structures in a website instead of the whole site, so this one is May not an optimal solution. Gupta et al. [12] presents an approach based on finding maximum links form user preference data in websites. This approach takes a long time to run even for a small website. Lin developed an integer programming model that uses inter relationship between the web pages to reorganize a website. But it is applicable to small websites only, for large it takes long computation time.
Lin and Tseng [13] developed an ant colony system to improve user navigation. This approach provides solutions in a less time and is useful for small websites, not appropriate for large websites. W. Yan, M. Jacobsen, H. Garcia-Molina and U. Dayal [14] presents an approach in which users are classified in terms of visited pages. Using weblogs, clustering is done on the basis of users which visit similar pages. According to group they recommend links dynamically.

III. EXISTING SYSTEM

In existing system, the mathematical programming model [1] is used to enhance user navigation. User navigation data is used to reorganize the web pages. The out degree means number of Out links in a page and this out degree threshold is used to control the number of links in a page to minimize information Overload. Targeted pages are obtained with page-stay time information [15] and Mini sessions [16] are identified with path threshold [17] information. Backtracking algorithm is used to estimate backtracking in user access pattern. Average user navigation and benefited user count metrics are used to evaluate the navigation performance [18]. It improves the website with minimal changes to the existing structure and it is done by following three phases they are:

Site Structure Analysis

First is to identify how the web pages in a website are connected i.e. to analyze current structure of website. It can be easily done by drawing the graph where the nodes represent the web pages and edges represent the links.

Relevant Mini Sessions

Mini sessions are the group of pages visited by user for only one target in a particular session known as mini session. Here path threshold is set for each mini session. A mini session is relevant only if its length is larger than the corresponding path threshold otherwise it consider as an irrelevant mini session and only relevant mini session is considered for improvement.

Relevant Candidate Links

Candidate link are links that can be selected to improve the user navigation which are obtain from a relevant mini session. A link is said to be relevant if adding the link can improve the user navigation without increase in path threshold, otherwise it is considered as the irrelevant candidate link. Only the candidate link relevant to the mini session should be considered for improvement.

Issues In Existing System

Improvements should be done such that minimum changes should be performing in current site structure and user can reach to the target page in fewer clicks. In existing system a goal is set for user navigation for each target page. Target page is obtained with page stay time out. For a mini session having target page, determine whether user reach with minimum steps or not by comparing sessions length with path threshold. The following issues are identified from the existing website restructuring methods.

- User access pattern are not consider
- User target identification accuracy is low

IV. PROPOSED SYSTEM

In proposed system, we can generate a link based on data mining method Clustering and Association rule to find frequent links by using user access pattern. Here we proposed an optimal solution to find user target to improve user navigation. As our aim is to improve user navigation so efficiency of improved website is also find that how much user navigation is improved.
The proposed work consists of following modules:

1) Pre-processing
2) Cluster Formation
3) Pattern Discovery
4) Restructuring

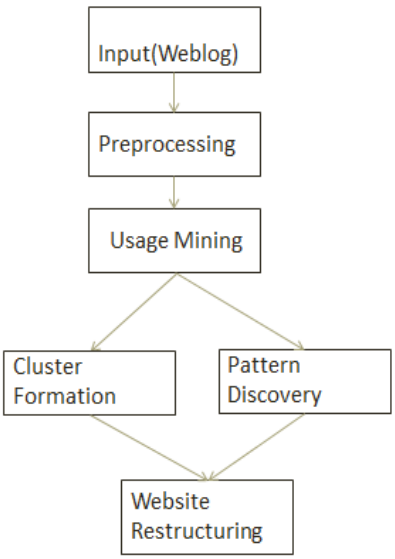


Fig.1 Proposed System

Pre-processing

In pre-processing first to understand the current site structure and collect weblogs in which user sessions are obtain. The Web site is represented as a graph in which a page is represented as a node and a link as an arc. Each page of the Web site is parsed sequentially and the links in the page are extracted to obtain the website structure. Preprocessing need to perform on weblog data to remove irrelevant data like incomplete request, sessions value less than defined threshold, repeated user, converting logs into data so that it can be mined. This pre-processing task is actually for removing irrelevant links so that time of building results and reorganizing of links will be reduced.

Cluster Formation

After pre-processing clustering is done on the URL visited by all users. K-means clustering algorithm is used. To perform clustering for this we chooses two parameter, time for which user on website and number of clicks on URL of web pages, some threshold is defined to perform clustering on these parameter. The links obtain in clusters are those which having high stay time and large number of clicks. More clicks on URL and high stay time can give cluster with high frequency of page access and distance measure is Euclid.

Pattern Discovery

Then to extract useful pattern from dataset Apriori algorithm is used. It is the process of finding useful pattern or extract knowledge from weblog data by using various data mining algorithms. Then this extracted information can be represented in various forms such as graphs, tables, curves etc. To find links that need to be change also depends on user identification and traversing path of user. So to find out links that to change can use association rule algorithm. It is pattern extraction algorithm which identifies correlations between items in transactional databases. This algorithm searches all possible patterns for rules that meet the user-specified support and confidence thresholds. It refers to finding set of pages that are accessed together with a maximum support value.

Restructuring

Here the user behaviour and time which user spent on pages both are consider for finding the target page. Apriori algorithm is used to get frequent links and clustering is used to obtain clusters with high frequency of stay time and visited URL. If that obtained frequent links are available in cluster means user are on page for more time as well as it access frequently then, it will check it for out-degree threshold and existence i.e. link is already present in current structure or not. If both condition satisfies link will be restructured. If links are matches between frequent item set and the clusters then they are considered as candidate links. To achieve minimum changes in website structure one parameter is to consider i.e. an out-degree threshold which defines how many out links can be applicable to each link .For example out degree threshold is five so only five out- links are allowed on page. If at any page already five links are present then cannot make out-link from that page.

V. ALGORITHMS USED

1. K-means Clustering Algorithm

Clustering is the process of partitioning a group of data points into a small number of groups. In general, we have n data points $\mathbf{x}i$, i=1...n that have to be partitioned in k clusters. The goal is to assign a cluster to each data point. K-means is a clustering method that aims to find the positions μi, i=1...k of the clusters that minimize the *distance* from the data points to the cluster.

K-means clustering:

1. Select K points as initial centroid.
2. repeat
3. Form K clusters by assigning all points to the closest centroid.

4. Recomputed the centroid of each cluster.
5. Until the centroids don't change.

2. Association Rule Algorithm

Association rules can be divided into two steps:

- Find all item sets whose support is greater than the specified threshold. Item sets with minimum support are called frequent item sets.
- Generate association rules from the frequent item sets.

L1: = {frequent 1-itemsets};
k: = 2; // k represents the pass number
While (Lk-1)
Ck = New candidates of size k generated from Lk-1
For all transactions t Increment count of all candidates in Ck
That contained in t
Lk = All candidates in Ck with minimum support
k = k+1
Report Uk Lk as the discovered frequent item sets

VI. MATHEMATICAL MODEL

Problem Description

Website is represented as a graph in which pages are represented as nodes and links are represented as arcs. Let X be the number of pages in a website. Let S= {S1, S2… Sn} be the set of n users. Let L= {L1, L2, L3…… Ln} be the set of links visited by user Si where i= {1, 2, 3… n}. Let N= {N1, N2, N3…… Nn} be the set of stay time of user Si.

Activity 1

Here we will find clusters for which we have both set L and set N. Cluster of links Cij where i and j are nodes of link is obtained by distance measure Euclid is given by,

$$d(i, j) = \sqrt{(Ni - Nj)^2 + (Li - Lj)^2}$$

Here we also consider user behavior to find out target page, set L considered as candidate set i.e. set of all items.

Activity 2

Now set of frequently visited links Fij ={F1 ,F2, F3, …. , Fn} where F1={A,B,H,K} , F2={A, C, D, K} i.e. user navigated path is obtained using Apriori algorithm with minimum support and confidence.

Activity 3

After association similarity between cluster Cij and frequent set Fij is need to check. Links which are present in set Aij as well as in set Cij are Considered as relevant candidate links which need to restructure is given by,

$$Aij \rightarrow Cij$$

To obtain set of relevant candidate links R= {R1, R2, R3 ….. Rn} need to check out degree threshold i.e. how many out links are allowed for a page. Links are matches between frequent item set and the clusters are considered as links to be reorganized if it fulfills an out-degree threshold.

VII. RESULTS

The below screen shots experimentally shows the proposed work:

Login Page:

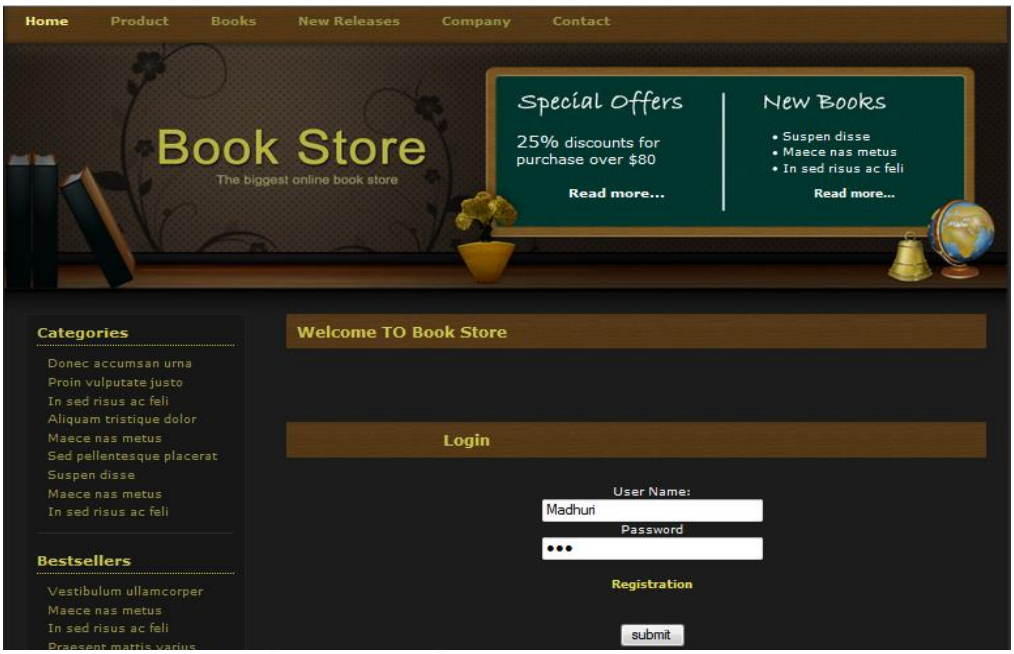


Fig 2: Login Page

Search Page:

Fig 3: Search Page

Figure 1 and 2 shows the login page and successful login status of website. After that arrow shows the navigation on website by user.

Once user logged in session get started. The tag shows links to search relevant data.

Server Log:

id	username	ip_add	mac_add	browser_info	date	url
547	chetan	192.168.1.107	(NULL)	Mozilla/5.0 (Windows NT 6.1; rv:30.0) Gecko/20100101 Firefox/30.0	05/20/2014 16:01:19	http://localhost:8080/Book
548	chetan	192.168.1.107	(NULL)	Mozilla/5.0 (Windows NT 6.1; rv:30.0) Gecko/20100101 Firefox/30.0	05/20/2014 16:01:44	http://localhost:8080/Book
549	chetan	192.168.1.107	(NULL)	Mozilla/5.0 (Windows NT 6.1; rv:30.0) Gecko/20100101 Firefox/30.0	05/20/2014 16:01:50	http://localhost:8080/Book
550	chetan	192.168.1.107	(NULL)	Mozilla/5.0 (Windows NT 6.1; rv:30.0) Gecko/20100101 Firefox/30.0	05/20/2014 16:02:00	http://localhost:8080/Book
551	chetan	192.168.1.107	(NULL)	Mozilla/5.0 (Windows NT 6.1; rv:30.0) Gecko/20100101 Firefox/30.0	05/20/2014 16:02:05	http://localhost:8080/Book
552	chetan	192.168.1.107	(NULL)	Mozilla/5.0 (Windows NT 6.1; rv:30.0) Gecko/20100101 Firefox/30.0	05/20/2014 16:02:11	http://localhost:8080/Book
553	chetan	192.168.1.107	(NULL)	Mozilla/5.0 (Windows NT 6.1; rv:30.0) Gecko/20100101 Firefox/30.0	05/20/2014 16:02:19	http://localhost:8080/Book
554	chetan	192.168.1.107	(NULL)	Mozilla/5.0 (Windows NT 6.1; rv:30.0) Gecko/20100101 Firefox/30.0	05/20/2014 16:02:22	http://localhost:8080/Book
555	chetan	192.168.1.107	(NULL)	Mozilla/5.0 (Windows NT 6.1; rv:30.0) Gecko/20100101 Firefox/30.0	05/20/2014 16:02:25	http://localhost:8080/Book
556	chetan	192.168.1.107	(NULL)	Mozilla/5.0 (Windows NT 6.1; rv:30.0) Gecko/20100101 Firefox/30.0	05/20/2014 16:02:36	http://localhost:8080/Book
557	chetan	192.168.1.107	(NULL)	Mozilla/5.0 (Windows NT 6.1; rv:30.0) Gecko/20100101 Firefox/30.0	05/20/2014 16:11:54	http://localhost:8084/Book
558	chetan	192.168.1.107	(NULL)	Mozilla/5.0 (Windows NT 6.1; rv:30.0) Gecko/20100101 Firefox/30.0	05/20/2014 16:12:09	http://localhost:8084/Book
559	chetan	192.168.1.107	(NULL)	Mozilla/5.0 (Windows NT 6.1; rv:30.0) Gecko/20100101 Firefox/30.0	05/20/2014 16:12:13	http://localhost:8084/Book
560	chetan	192.168.1.107	(NULL)	Mozilla/5.0 (Windows NT 6.1; rv:30.0) Gecko/20100101 Firefox/30.0	05/20/2014 16:12:22	http://localhost:8084/Book
561	chetan	192.168.1.107	(NULL)	Mozilla/5.0 (Windows NT 6.1; rv:30.0) Gecko/20100101 Firefox/30.0	05/20/2014 16:12:26	http://localhost:8084/Book
562	chetan	192.168.1.107	(NULL)	Mozilla/5.0 (Windows NT 6.1; rv:30.0) Gecko/20100101 Firefox/30.0	05/20/2014 16:12:31	http://localhost:8084/Book
563	chetan	192.168.1.107	(NULL)	Mozilla/5.0 (Windows NT 6.1; rv:30.0) Gecko/20100101 Firefox/30.0	05/20/2014 16:12:33	http://localhost:8084/Book
564	chetan	192.168.1.107	(NULL)	Mozilla/5.0 (Windows NT 6.1; rv:30.0) Gecko/20100101 Firefox/30.0	05/20/2014 16:12:34	http://localhost:8084/Book

Fig 3: Log Maintenance On Server

Fig. 3 Shows all the logs of user navigation on server side including following attributes like user IP address, session Duration, browser information, url visited etc.

Admin Page:

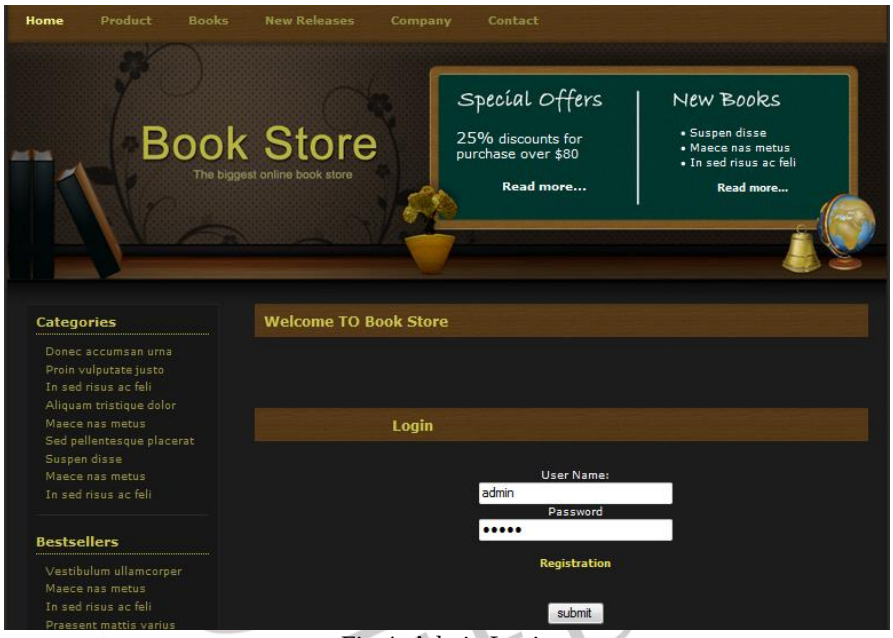
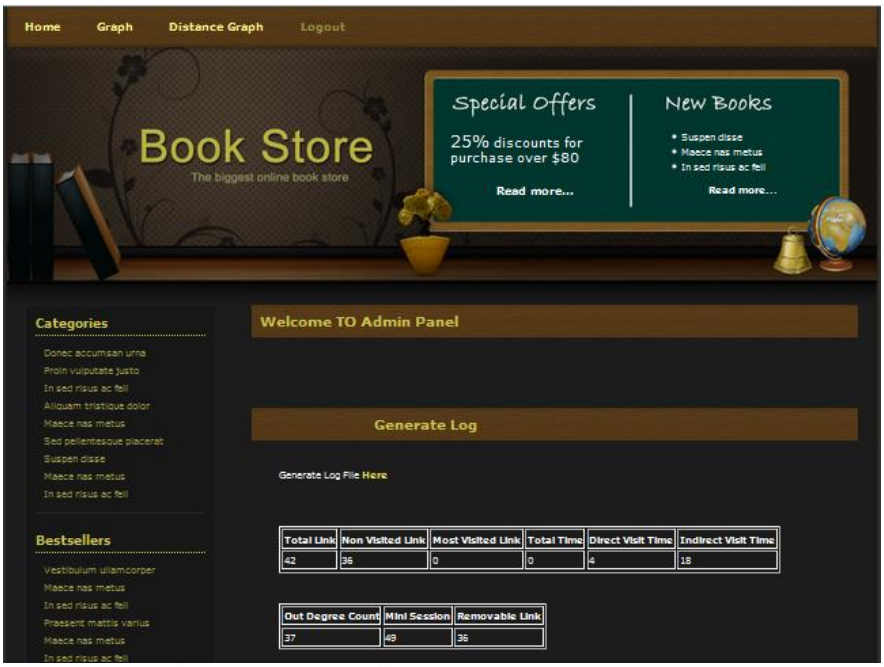


Fig 4: Admin Login

Above figure shows the admin login page where mostly all the data regarding to user navigation is proceeds and only authorized person allow accessing this data.

Testing Results:



(a)

id	non_visited	visited
1	20	10
2	10	4
3	13	4
4	23	5
5	23	2
6	16	3
7	20	4
8	11	2
9	22	5
10	13	1
11	17	2
12	14	3
13	22	5
14	11	1
15	22	2

(b)

id	mst	username
494	new_release.jsp	priya
495	booked.jsp	priya
496	airtaxi.jsp	priya
497	train.jsp	priya
498	bus.jsp	priya
499	logout.jsp	priya
839	home.jsp	admin
840	logout.jsp	admin
898	home.jsp	suraj
899	company.jsp	suraj
900	service.jsp	suraj
901	service.jsp	suraj
902	service.jsp	suraj
903	service.jsp	suraj
904	service.jsp	suraj
905	service.jsp	suraj
906	service.jsp	suraj
907	service.jsp	suraj
908	logout.jsp	suraj

(c)

Fig 5: Testing result (a) Link Structure To Be Improved (b) Visited and Non-visited Links (c)Most Visited Links By User

Execution Time:

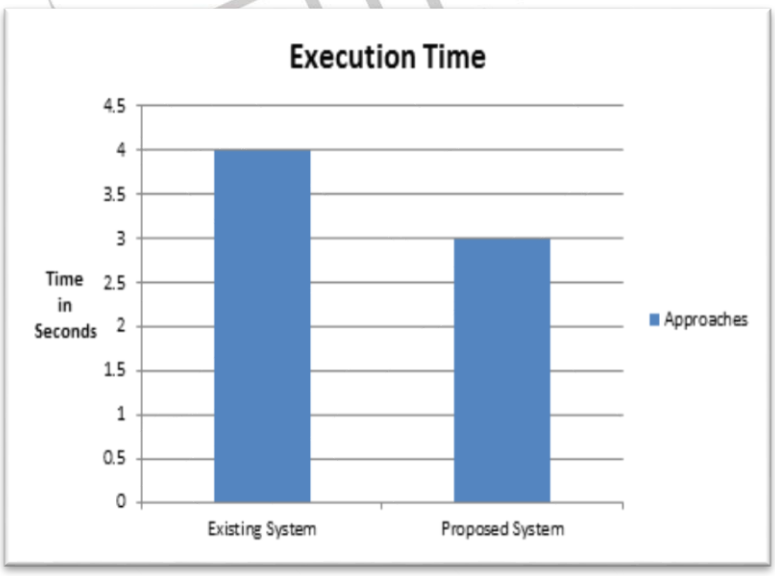


Fig 6: Execution Time

Fig.6 shows the difference between execution time taken by existing system and proposed system during execution.

VIII. CONCLUSION

In this paper, an optimal solution is given to improve website user navigation. Web usage mining is used to analyze the user behaviors on website. Web page links are used to manage the user navigations in the web site. Web page links are restructured to improve the user navigations by using Pattern based model. The system requires minimum site structure changes as user access pattern is considered for improvement.

IX. FUTURE WORK

This paper can be extended with the following features. The system can be integrated with web personalization problem and work will be leads to develop a method to improve structure of dynamic website.

ACKNOWLEDGMENT

I'm thankful to many persons who contributed to the completion of this research. Particularly I like to thank our PG Coordinator Prof. S. N. Shelke to help in research and Head of department and Guide Prof. B.B. Gite to allow me to continue this topic. Lastly I would like to Department of computer engineering, Sinhgad College of engg. , Kondhwa to share their knowledge with me during my research work.

REFERENCES

[1] Min Chen and Young U. Ryu "Facilitating Effective User Navigation Through Website Structure Improvement" IEEE Transaction on knowledge and Data Engineering, Vol. 25, no. 3, March2013
[2] T. Nakayama, H. Kato, And Y. Yamane, "Discovering the gap between Web Site Designers Expectations and Users Behavior," Computer Network, vol. 33, pp. 811-812,2000.
[3] M. Perkowitz and O. Etzioni, "Towards Adaptive Web Sites: Conceptual Framework and Case Study," Artificial Intelligence, vol. 118, pp. 245-275, 2000.
[4] M. Eirinaki and M. Vazirgiannis, "Web Mining for Web Personalization," ACM Trans. Internet Technology, vol. 3, no. 1,pp. 1-27, 2003.
[5] M. Kilfoil et al., "Toward an Adaptive Web: The State of the Art and Science," Proc. Comm. Network and Services Research Conf.,pp. 119-130, 2003.
[6] B.Padmanabhan and A. Tuzhilin, "On the use of optimization for data mining: Theoretical interactions and ecrm oppotunities," in Management Science.
[7] B. Mobasher, R. Cooley, and J. Srivastava, "Creating Adaptive Web Sites through Usage-Based Clustering of URLs," Proc. Workshop Knowledge and Data Eng. Exchange, 1999.
[8] M. Nakagawa and B. Mobasher, "A Hybrid Web Personalization Model Based on Site Connectivity," Proc. Web Knowledge Discovery Data Mining Workshop, pp. 59-70, 2003.
[9] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa, "Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization, "Data Mining and Knowledge Discovery, vol. 6, no. 1, pp. 61-82, 2002.
[10] B. Mobasher, R. Cooley, and J. Srivastava, "Automatic Personalization Based on Web Usage Mining," Comm. ACM, vol. 43, no. 8,pp. 142-151, 2000.
[11] Y.Fu,M.Y.Shih, M.Creado,and C. Ju, "Reorganizing Web Sites Based on User Access Patterns", vol. 11, no. 1, pp. 39-53,2002.
[12] R. Gupta, A. Bagchi, and S. Sarkar, "Improving Linkage of Web Pages," INFORMS J. Computing, vol. 19, no. 1, pp. 127-136, 2007.
[13] C. Lin and L. Tseng, "Website Reorganization Using an Ant Colony System," Expert Systems with Applications, vol. 37, no. 12,pp. 7598-7605, 2010.
[14] W. Yan, M. Jacobsen, H. Garcia-Molina, and U. Dayal, "From User Access Patterns to Dynamic Hypertext Linking," Computer Networks and ISDN Systems, vol. 28, nos. 7-11, pp. 1007-1014, May 1996.
[15] R.Srikant and Y. Yang, "Mining web logs to improve web site organization," in Conf. World Wide Web.
[16] B. R. Cooly and J. Srivastav, "Data preparation for mining world wide web browsing patterns," in Knowledge and Information System.
[17] R. Bucklin and C. Sismeir, "A model of website browsing behavior estimated on clickstream data," in J. Marketing Research.
[18] P. Pirolli and S. K. Card, "Information foraging," in Psychological Rev.