International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 02 Issue: 09 | Dec-2015 www.irjet.net p-ISSN: 2395-0072

WEBSITE STRUCTURE IMPROVEMENT BY USING TAILORING METHOD

Milind Mahadeo Shinde

Student,
Department of computer engineering,
Imperial College of engineering & research,
Pune, India

Vinod S. Wadne

Prof.,
Department of computer engineering,
Imperial College of engineering & research,
Pune, India

Abstract - ***The www grows rapidly and increases the complexity of web applications and navigation of websites. This paper presents a widespread overview of web mining techniques and methods used for the estimation of reconciling systems to get better website navigation efficiency to enhance the efficiency of web site. The goal of this project is to make adaptive Web sites by developing the site structure to help user access. This technique consists of three steps: preprocessing, classification of page, and site reorganization. In particular, a mathematical programming model is used to enhance the navigation of user on a web while decreasing alterations to its present structure. Especially, two evaluation metrics are defined and used them to judge the performance of enhanced website using real data set. Specifically, the modules that include a Web personalization system is introduced, which gives focus on the Web usage mining. The proposed techniques are achieved superior web navigation very effectively and maintain efficiency and it is more efficient from existing system.***

Key Words: *Data and web mining, Web personalization, Mathematical programming model, Web reorganization…*

1. INTRODUCTION

The steady development in the range and use of the World Wide Web evolved new techniques of development and design of on-line Information Services. Generally lot of Web structures are complex and huge; so the users ignore the goal of its inquiry, or when they try to navigate through them, they can receive ambiguous results. On other side, the e-business region is quickly growing and the requirement for Web market places which expect the requirements of the customers is higher than ever evident.

That's why, the need for predicting requirements of user to enhance the retention of user and usability of a Web can be addressed by personalizing website. The definition of Web personalization is, every action which adapts the services or information supplied by a Web site for the needs of a specific customer, user or a set of customers and users, taking benefit of the knowledge obtained from the navigational behaviour of users and interests of individual, in grouping with the structure of the Web and multiple contents. The main aim of Web personalization is to "supply multiple users with necessary information, instead of expecting from them which ask for information explicitly".

Though finding required information on the website is difficult and creating efficient web is not a trivial job; there are lot of ever-increasing investments in web designing. If users found difficulty to reach the target pages are ready to abandonee a website; though it contains high quality information. The total reorganization of website could thoroughly alter the location of recognizable items and the creation of new website can disorient customers. So it cannot be repeatedly performed to enhance the user navigability. Different techniques are used to enhance the web structure like algorithms for Clustering, Ant colony system.

An algorithm Graph Partitioned Clustering used to assemble multiple users with same navigational pattern. The undirected graph is used; which is based on connectivity stuck between every pair of used web pages. To every edge in the graph is given a weight and it is based on frequency and connectivity time. The connectivity time calculates the amount of visit ordering to every two pages in the session. Thus navigation pattern for users enhance the website structure by reorganizing on it.

2. RELATED WORK

The Web Navigation is an essential part of a website. It is a way where a website is created in such a way that the surfer can view the information they wants. Lot of websites have a hierarchical association of content [1]. Particularly, it is often not clear that where a specific content or document is situated. Three techniques like First one, Optimize Time and Optimize Benefit are used for recommending extra links to Website Administrator. The algorithm is also used to find necessary pages or information in a website whose position is different from the position where visitors generally try to find them. The main focus is on the backtracking technique used by visitors when they do not find necessary information where users expect it. The region from where users backtrack is considered expected locations for these pages. When the web-site does not have an apparent separation in between content pages and index pages, then it can be difficult to differentiate

between other pages and target pages. So the algorithm can produce false expected locations for pages if it treats the target pages as backtracking points and can fail to spot desired locations if it treats the backtrack points as the target pages.

The main aim of a technique used to reorganize web sites residing on user access patterns is to make adaptive web sites by developing site structure to assist user access pattern [2]. Three steps are used in this approach: classification of page, pre-processing and website reorganization. In pre-processing stage, the pages in a web site are processed to form an internal demonstration of the site. User's page access information is collected from web server log. In next stage of page classification, web pages on the web site are divided into two categories like content pages and index pages based on the information of page access. The pages are classified and examined to organize them properly. The decision of reorganization algorithm bases on the user accesses information, not on the web content.

Hou et al [3] recommended an analysis of hyperlinks in the web page and new method to build the page source. Page similarity is used by two algorithms to discover relevant pages. The first one is, Extended Co-citation algorithm, it is a co-citation algorithm which extends the historical co-citation concepts. It is instinctive and short. The second algorithm is, Latent Linkage Information algorithm i.e. (LLI), which finds appropriate pages more efficiently and accurately by using theories in linear algebra, particularly singular value in matrix decomposition, to expose the relationships among pages. The recent page source minimizes the power of the pages similar web-site to a reasonable stage in page similarity dimension, avoids little helpful information being lost and prevents the results from being displaced by malicious hyperlinks. It can recognize the pages that are semantically related to the specified page and the pages which are related to the specified page in a vast sense. The In LLI algorithm, the page similarity concept could be modified for clustering of pages if the clustered pages are not huge.

A web personalization technique customizes a Web site according to the needs of particular users [4], which takes advantage of the information gathered from analysis of the navigational behaviour of user in association with another data collected in context of web, structure, user profile data and content. As the web grows tremendously, the web personalization domain has obtained big momentum in the commercial areas and research. It can be gained by getting advantage of the navigational behaviour of user, as exposed throughout the web usage logs processing, as well as interests and characteristics of users. The solution is not provided for techniques combination used in profiling of user, mining in web usage, acquisition of contents and website publishing and management task.

3. MOTIVATION

As per the brief literature survey, the main limitation of existing system is difficulty in navigation. This problem triggers most consumers to abandon a website and switch to a competitor. Generally, having traversed several paths to locate a target indicates that this user is likely to have experienced navigation difficulty. So due to this reason the main motivation is to improve the navigation effectiveness of a website with minimal changes. A mathematical programming model is used to improve the user navigation on a website while minimizing alterations to its current structure and show that this MP model not only successfully accomplishes the task but also generates the optimal solutions surprisingly fast and techniques that can accurately identify users' targets can be generated.

4. PROPOSED SYSTEM

4.1 Personalization Based On Web Usage Mining

The Web personalization process generally contains of three phases: preparation and transformation of data, discovery of patterns, and its recommendation. In the historical approaches of collaborative filtering, the phase of pattern discovery and the recommendation phase are performed in real time. The web personalization systems are based on website usage mining [5], which performs discovery of patterns offline. The data preparation phase converts raw web log data files in stream data which can be used by task like data mining. A wide range of different techniques of data mining can be applied on the web application data or click stream in the pattern discovery phase, like data clustering, mining of association rule [6, 7, 8], and in order pattern discovery. Generally, the active user session is considered for recommendation in conjunction with discovered patterns to be discovered to supply the personalized content.

4.2 Web Personalization

The process of web personalization is also called as "tailoring" web Pages according to the requirements of users using the information of navigational behaviour of user and user's profile data. An approach described by Perkowitz and Etzioni automatically going to synthesize index pages which include links to various pages pertaining to specific topics residing on the co-occurrence frequency user traversal pages, to enhance navigation of user. The methods which are proposed by Mobasher et al. generates clusters of the users profiles from website logs and then generate dynamic links for customers who are classified into various categories based on their different access patterns. The prior studies on website has mainly focused on different types of issues, like understanding of web structures, finding the appropriate pages from a given page, informative structure mining of a newly created website, and collecting webpage template.

4.3 Web Transformation

Specifically the web transformation, on another side, consists of change in the website structure to

enhance and facilitate the navigation for a huge set of users or group of users without personalizing of pages for individual separate users. An approach is described here to restructure web pages in such way that it can provide desired information to the users in minimum clicks. As this approach is beneficial to the group of users, it takes into account local structures in the website instead of a site as a whole; therefore the new architecture cannot be essentially optimal. A heuristic method proposed here which is based on simulated re-link web pages to enhance the website navigability. This technique makes use of collective data of user preference and can also be used to enhance the link structure in various websites for wired as well as wireless devices.

4.4 Knowledge discovery from web logs

For predicting the user behavior and preparing the web structure calculation of web access log is important task. The applications point of view is considered here and the information collected from various patterns of web usage can be useful to perform various tasks related to e-services, e-business, e-education, and on-line communities and so on. On another view side, there is fast growth in the data density and data size. Therefore information given by the running web log file in analysis tools can enhance inadequate information and so vast intelligent data mining techniques are required.

In website usage mining, the web log files have an important role. All the necessary knowledge and information gathered from web log files is used in navigation process of user. Different types of users have also assigned varied navigational patterns with it. Generally, users continuously alter their focus, so it is difficult to gain such navigational behavior related knowledge. The knowledge in navigation pattern is used by users for two reasons: For website personalization system and to to help users by predicting the user's future request.

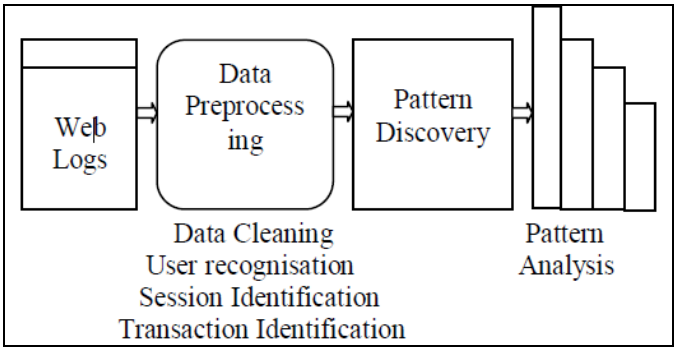

Fig -1: Knowledge Discovery

4.5 Mathematical Model

The mathematical programming model is used here for enhancing the navigational behavior of user on website at the same time reducing the contents for alterations to its present structure. The Results are conducted on a in public available real data set from wide-ranging tests to show that this model not only enhances the navigation of user with less changes, but it can be efficiently solved.

To notify the path has been traversed by the user backtracking is used, and a backtracking technique is defined as a revisit of user to a earlier browsed page. The main fact is that users can perform backtrack if they don't get d location. Hence, a path can be defined as the series of pages which are visited by a user or customer without backtracking; and this concept is analogous to maximal forward reference. Basically, every backtracking point is supposed as ending of a path.

A mathematical programming model is shown below. It is used to decrease the alteration to the existing structure of web-site by enhancing the navigation of user as follows [9]:

$$\text{Minimize} \sum_{(i,j)\in E} x_{ij}[1-\lambda_{ij}(1-\varepsilon)] + m\sum_{i\in N_E} p_i$$

subject to

$$c_{kr}^S = \sum_{(i,j)\in E} a_{ijkr}^S x_{ij}; r = 1,2,\ldots,L_p(k,S), \quad k = 1,2,\ldots,L_m(S), \forall S \in T^R \tag{1}$$

$$\sum_{k=1}^{b_j}\sum_{r=1}^{L_p(k,S)} c_{kr}^S \geq 1; \forall S \in T^R, j = tgt(S) \tag{2}$$

$$\sum_{j:(i,j)\in E} x_{ij}(1-\lambda_{ij}) + W_i - p_i \leq C_i; \forall i \in N_E \tag{3}$$

$$x_{ij} \in \{0,1\}, p_i \in \{0\} \cup Z^+, \forall (i,j) \in E, i \in N_E. \tag{4}$$

5 HIGH LEVEL DESIGNS

5.1 Architecture of Proposed system

The proposed system is used for user navigation by making use of relevant mini session and its relevant candidate links. In transformation techniques, the mathematical programming model is used to improve the navigation of users instead of personalization of individual user or single user. The process web pages tailoring as per need of users by using information regarding to the navigational behavior of user and its profile information is called as personalization of web. Therefore, the technique of personalization is a tailoring of web pages. A backtracking can be defined as a revisit of user to earlier browsed page. If user do not find page at expected

location, then user will backtrack. So, a path is defined as a no of pages and its sequence visited by a user without backtracking. This concept is same as maximal forward reference. Basically, each and every backtracking point is end of path. The experiments were conducted, on a data set collected from a real website and on synthetic data sets. The model is first tested with changing parameters values on all data sets. The real data is then partitioned into training and testing data.

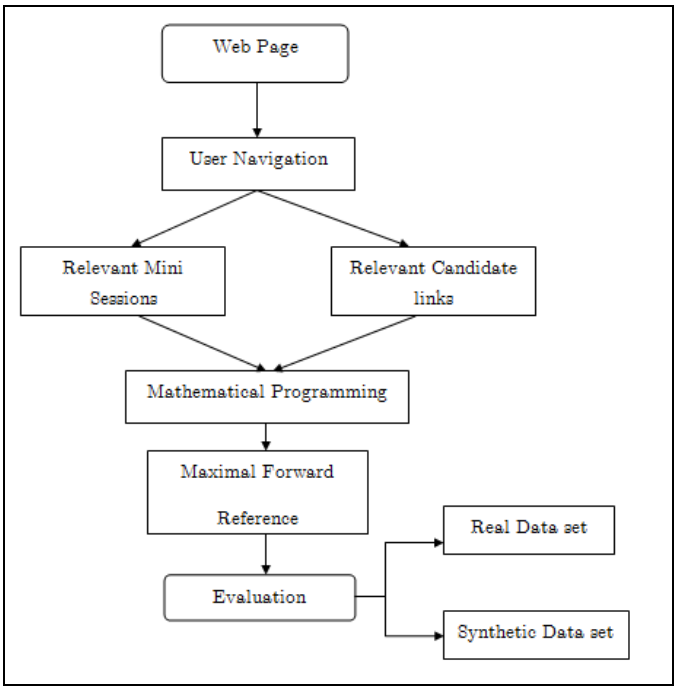

Fig -2: System Architecture

5.2 Mathematical model using Deterministic Finite Automata

The mathematical module for proposed system is represented by following figure 3. It consists of five tuple DFA which is represented as M.

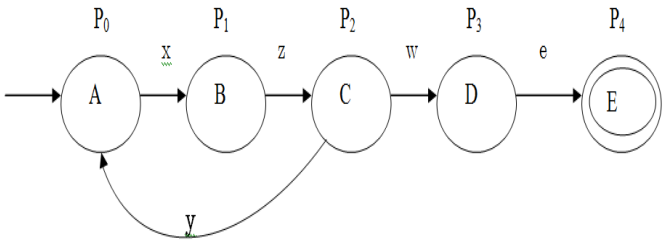

Fig- 3: Deterministic Finite Automata (DFA)

M= {Q, Є, ∂, Qo, F}
Where,

Q= {Po, P1, P2, P3, P4}
Є = {x, z, w, e, y}
∂= {A, B, C, D, E}
Qo= {Po}
F= {P4}
Where,

P0 = process of personalization,

P1 = process of transformation

P2 = Backtracking

P3 = Reduction of problem size

P4= Evaluation

x = url for single user

z = url for target page

y = Back to previous page

w = insert minimum links

e = target page ID

5.3 Algorithm to find Expected Location

1) In website; build the hash table of links
2) Then partition web log by visitor:
By visitor ID as a Primary Key sort the web log file and Time as Secondary key. Partition web log file by hashing on visitor ID Partition each log separately
Scan web log and take out the sequence of pages for every visitor ID
3) For each visitor, partition web log
4) Expected location for visitor and target page:
Let (P1, P2, . . . Pn) be set of visited pages
B = Φ denotes list of backtrack pages
a) for i = 2 to n-2 begin
b) if ((Pi-1 = Pi+1) or (no link from Pi to Pi+1))
c) Add Pi to B
end
If (B not empty) Add (Pn, B, Pn-1) to table

The hash table is built in first step. Then the web log is partitioned by visitor in Step 2. In next Step 3, we divide the sequence of accesses for every visitor by the target pages which they visit. So it is assumed that the website administrator either specifies set of feasible target pages, or it specifies a time threshold value to differentiate between target pages and the other pages. In final step 4, all predictable locations (if any) are found for that target page, and added it to a table and it is used by the next step of algorithm.

Detection of backtracks finds in Step 4(b). Additionally to check for the nonappearance of a link from the current page to the next page, we also see it if the previous page and next pages are same or not. The last check takes precaution of the case that where visitors use a navigation link to go to the previous page visited without using the "back" button present in the browser.

5.4 Algorithm to optimize time

The algorithm to optimize time Recommend the set of pages that minimize the number of times the visitor has to backtrack, i.e., the number of times the visitor does not find the page in an expected location. The goal of this algorithm is to minimize the number of backtracks the visitor has to make. We use the number of backtracks as a proxy for the search time.

Algorithm is as follows:
Repeat
For each record begin
Let m be the number of expected locations in this record
For j = 1 to m
Increment support of value (Ej) by m+1-j;
end
Sort pages by support and P = Page with highest support
If support(P) ≥ St begin
Add (P, support(P)) to list of recommended pages.
For each record begin
For k = 1 to n begin
If value (Ek) = P
Set Ek, Ek+1, . . . , En to null
end
end
end
until (support (P) < St);

6 PERFORMANCE EVALUATION

In Web Organization; to locate particular destination; no of redirection links has been traversed and 3 mini sessions are used to check weight on no of path traversed. The tracking result shows values for objective function, path threshold, current Outdegree and Outdegree threshold. The links whose value does not exceed the value of Outdegree threshold are considered for web organization. Following table shows the links used in mini-session 1.

Objective Function	Path Threshold	Current Outdegree	Outdegree Threshold
0.2758	0.5045	0.8046	1.3091
0.6719	0.2582	0.3810	0.6392
0.7477	0.8359	0.3130	1.1489
0.9648	0.4220	0.0195	0.4414
0.3026	0.2329	0.8037	1.0366
0.1944	0.6156	0.5873	1.2029

Table -1: Tracking Result

6.1 Graphical Result

The time efficiency comparison between existing system and proposed system is given in milliseconds. The proposed system makes use of mathematical model and Optimize time algorithm to perform faster and exact search as compared to existing system.

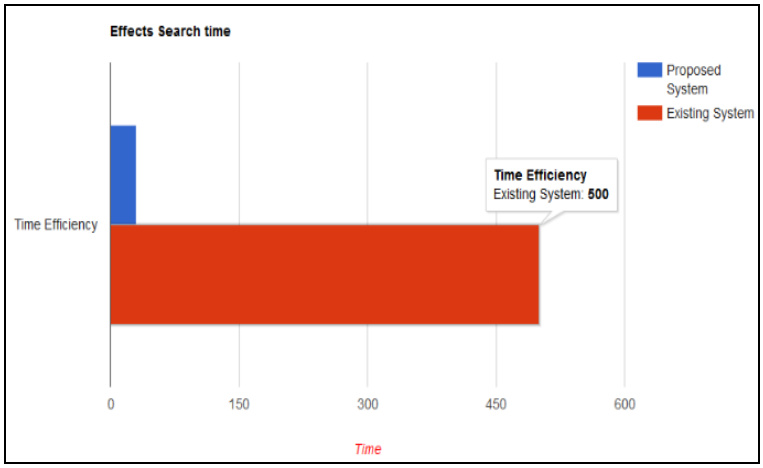

Chart -1: Time efficiency of existing system

If the existing system reaches to the required target in 500 milliseconds then proposed system reaches the goal in just 30 milliseconds as shown in chart 2. The time efficiency value for proposed system can be varied according to no. of traversed paths.

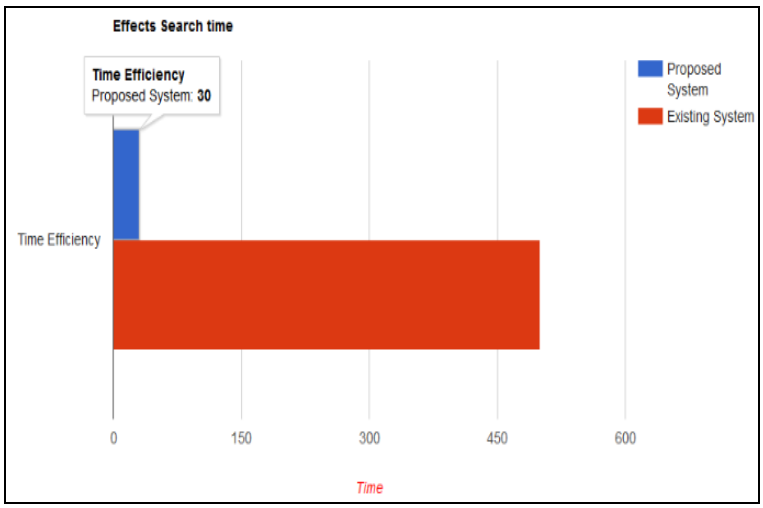

Chart -2: Time efficiency of proposed system

7 CONCLUSIONS

The tailoring method and K- means clustering techniques are used for proper navigation purpose. To enhance the navigation effectiveness; the algorithms like finding expected location of requested information and optimization of time are used. The transformation approach is generally suitable for informational websites, whose contents don't change along with time. The web personalization module works well with the sites which contents dynamic information by considering information related to each visitor. The proposed system reaches the target very quickly as compared to existing system.

REFERENCES

[1] Srikant R. and Yang Y. (2001) 'Mining Web Logs to Improve Web Site Organization', Proc. 10thInt'I Conf. World Wide Web, pp.430-437.

[2] Fu Y., Shih M.Y., Creado M. and Ju C. (2002) 'Reorganizing Web Sites Based on User Access Patterns', Intelligent Systems in Accounting, Finance and Management Vol. 11 No. 1 pp.39-53.

[3] Hou J. and Zhang Y. (2003) 'Effectively Finding Relevant Web Pages From Linkage Information', IEEE Trans. Knowledge and Data Eng., Vol. 15 No.4 pp.940-951.

[4] Eirinaki M. and Vazirgiannis M. (2003) 'Web Mining for Web Personalization', ACM Trans. Internet Technology Vol. 3 No. 1 pp.1-27.

[5] B. Mobasher, R. Cooley, and J. Srivastava. Automatic personalization based on web usage mining Communications of the ACM, 43(8):142–151, 2000.

[6] R. Agarwal, C. Aggarwal, and V. Prasad. A tree projection algorithm for generation of frequent itemsets. In Proceedings of the High Performance Data Mining Workshop, Puerto Rico, April 1999.

[7] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94), Santiago, Chile, Sept 1994.

[8] R. Srikant and R. Agrawal. Mining generalized association rules. In Proceedings of the 21st International Conference on Very Large (VLDB'95), Zurich, Switzerland, September 1995.

[9] Min Chen and Young U. Ryu, "Facilitating Effective User Navigation through Website Structure Improvement", VOL. 25, NO. 3, MARCH 2013

[10] M.S. Chen, J.S. Park, and P.S. Yu, "Efficient Data Mining for Path Traversal Patterns," IEEE Trans. Knowledge and Data Eng., vol. 10, no. 2, pp. 209-221, Mar./Apr. 1998.

[11] R. Cooley, B. Mobasher, and J. Srivastava, "Data Preparation for Mining World Wide Web Browsing Patterns," Knowledge and Information Systems, vol. 1, pp. 1-27, 1999.

[12] C.C. Lin and L. Tseng, "Website Reorganization Using an Ant Colony System," Expert Systems with Applications, vol. 37, no. 12, pp. 7598-7605, 2010.

[13] B. Mobasher, "Data Mining for Personalization," The Adaptive Web: Methods and Strategies of Web Personalization, A. Kobsa, W. Nejdl, P. Brusilovsky, eds., vol. 4321, pp. 90-135, Springer-Verlag, 2007.

[14] 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.