Improving Navigability of user through Relink Webpages using Transformation Method

Savitri Sheshappanavar
PG student
Department of CSE, AMC Engineering,
Bangalore, India
cssavitri@gmail.com

N. Kayarvizhy,
Associate professor
Department of CSE, AMC Engineering,
Bangalore, India
kayarvizhy@gmail.com

***Abstract*—The World Wide Web is the most useful source of information. It increases the complexity of web applications and web navigation. Designing a good structured Website to provide navigability to a user is an important challenge now days. The main reason for poor website is that web developers understanding of how website should be structured will be different from those of the users. Most of the existing methods like relink web pages using user navigation data, the completely reorganize the link structure of a website can be highly unpredictable and cost of disoriented users remains unanalyzed. In this paper we propose a mathematical programming model to improve the navigation effectiveness of a website with only minimum changes to its current structure. And the proposed model is particularly appropriate for informational website whose contents are relatively stable over time. It improves a website rather than reorganizes it and hence is suitable for website maintenance and can be applied in a regular manner.**

***Keywords*—** *Web usage mining, website design, navigability, relink, mathematical programming.*

I. Introduction

The rapid growth of the web in the last decade makes it the largest publicly accessible data source in the world. Extraction of useful information from the Web data has become more popular and as a result of that web mining has attracted lots of attention .Web mining is an application of data mining to large web data repositories. Web Mining [1] is categorized into three areas of interest as web content mining, web usage mining and web structure mining. Web content mining is the process of extracting knowledge from the content of documents or their descriptions. Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. Web usage mining is the application that uses data mining to analyze and discover interesting patterns of user's usage data on the web.

The fast-growing number of Internet users also gives huge business opportunities to firms. To satisfy the demands of online customers, firms are heavily investing in the development and maintenance of their websites. Even though there is increasing investments in website design, it is still known that finding desired information in a website is not easy and designing effective websites is also not an easy task [2]. And even the poor website design has been a key element in many numbers of high profile site failures. Also, if a website has information of high quality, but users having difficulty in locating the targets in that are very likely to leave a website.

The main reason for poor website design is that the web developers understanding of how a website should be structured will be different from those of the users [3]. Because of that difference, users cannot easily locate the desired information in a website. And, this problem is very difficult to avoid because when creating a website, web developers may not have a clear understanding of user's preferences and can only organize pages based on their own judgments and keeping little bit idea of users. However, the measure of website effectiveness should be only the satisfaction of the users rather than that of the developers, because users will be using it. Thus, web pages should be organized in a way that generally matches the user's model of how pages should be organized.

Our work is closely related to improve website navigability with the use of user navigation data. To address this it can be classified into two categories [3] and first is to provide a particular user by dynamically reconstituting pages based on his preference which is referred as personalization approach and the other is to modify the structure in order to provide the navigation for all users which are referred as transformation method.

We consider mainly with transformation approach here. Previous work considering the transformation approaches basically focuses on methods to completely reorganize the link structure of a website [5]. But there are many drawbacks of reorganization. Mainly like it will change the location of familiar items and new website may disorient users because of new location. And, the reorganized website structure will be highly unpredictable too. Because the website structure will be designed by some experts with some organizational logic, but this logic may no longer exist in the newly reorganized website. And mainly reorganization approach will change the website structure dramatically, so it cannot be frequently performed to improve the navigability.

Considering these drawbacks of reorganizing the website, we address this question to improve the structure rather than reorganizing it. So, we develop a mathematical programming model that provides user navigation on a website with minimal changes to its existing structure to reach targets faster. And the model is appropriate for informational websites whose contents are static and stable over time. For example, universities, hospitals, tourist attractions and sports are some organizations that have informational websites. Our model may not be appropriate for dynamic websites [6], because study state cannot be reached in user access patterns in such websites and in that case log files cannot be used.

In the model, we consider the out-degree i.e., number of outward links [7] in a page as a cost term in the objective function instead of, as a hard constraint. Even if adding these links provide facilitate user navigation, however, model restricts on the new structure, as it avoids pages from having more links than a specified threshold value. This allows a page to have more links than the out-degree threshold if the cost is reasonable. And hence offers a good balance between minimizing changes to a website and reducing information overload to users.

II. Related work

Website navigation has been assumed as the most important design features across many domains, including finance, e-commerce, entertainment, education, government, and medical. This led to many studies on improving user navigations with the knowledge mined from web server logs. And, here the web server is the logs [8] which maintain a history of page requests, information about the request, including client IP address, request date/time, page requested, HTTP code, bytes served, user agent, and referrer are typically added. Many commercial web log analyzer tools are available in the market that analyzes the web server log data to produce different kinds of statistics.

Past research mainly centered on the analysis of server log files to identify user access behavior, suggest opportunities to redesign the structure of a website, and develop adaptive websites. And, they categorized into web personalization and transformation approaches [4].

Web personalization approaches dynamically reconstitute pages for individual users based on their preferences. As in [4], describe an approach that automatically synthesizes index pages which contain links to pages pertaining to particular topics based on the co-occurrence frequency of pages. The methods proposed by Mo basher et al. [9] [10] Create clusters of user profiles from weblogs and then dynamically generate links for users who are classified into different categories based on their access patterns.

Web transformation involves changing the structure of a website to facilitate the navigation for a large set of users. Fu et al. [11] describes an approach to reorganize webpages so as to provide users with their desired information in fewer clicks. However, this approach considers only local structures. Gupta [6] propose a heuristic method based on simulated annealing to relink webpages, but it's not an optimal solution. Lin [7] develops integer programming models to reorganize a website based on the cohesion between pages to reduce information overload and search depth for users.

When we consider these two methods there are lots of difference between them, where the personalization approaches are more suitable for dynamic websites whose contents are more volatile and transformation approaches are more appropriate for websites that have a built-in structure and store relatively static and stable contents. And here in our work, we consider mainly the transformation approach.

III. Proposed work

The Weblog files [8] must be collected and broken down into user mini sessions, which can be used to analyze the website and user interaction. Before, analysis log preprocessing steps [8] are followed to filter irrelevant information from raw log files. A session is considered as a group of activities performed by a user during his visit to a site and it may include one or more target pages. Since, in our analysis we use number of paths traversed to find only one target, so here we use a different term that is mini session to refer to a group of pages visited by a user to only one target. Thus, a session may contain one or more mini sessions. We use the page-stay timeout heuristic described in [12], to demarcate each mini session.

We use backtracking to know the path that a user has traversed and backtrack is defined as a user's revisit to a previously browsed page. And usually users will backtrack if they do not find the page where they expect it [12]. A path is defined as a sequence of pages visited by a user without backtracking, and is similar to the maximal forward reference [41].

A. An Example

As in Fig. 1, illustrates a mini session, where a user starts from *A*, browses *D* and *H*, and backtracks to *D*, from where he visits *C*, *B*, *E*, *J*, and backtracks to *B*. Then, this user goes from *B* to *F* and finally reaches the target *K*. Now we denote the mini session by $S = \{\{A, D, H\}, \{C, B, E, J\}, \{F, K\}\}$, where an element in *S* represents a path traversed by the user. In this example, mini session *S* has three paths as the user backtracks at *H* and *J* before reaching the target *K*.

In the example shown in Fig. 1, the user has traversed three paths before reaching the target. But, the solution for user to reach the target faster is to introduce more links [10]. And in many ways extra links can be added. Consider, 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 save the user other two paths. Similarly, adding a link from *B* to *K* enables the user to reach the target [14] in the second path. Hence, this saves one path. We can also insert a link from *E* to *K*, and this is 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. And another possibility is to link *C* to *F*, which is a non target page. In this case, we assume that the user does not follow this new link, because it does not directly connect a page to the target *K*.

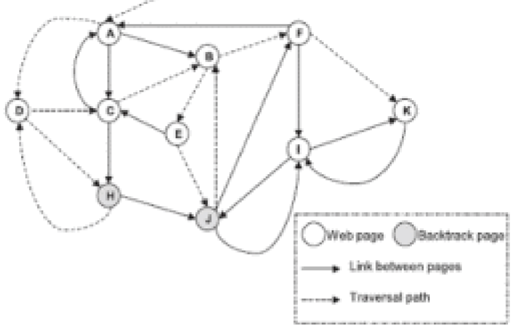

Fig. 1. Example of a mini session

B. *Proposed Transformation Approach*

In the Fig. 2, shows the proposed transformation approach for restructuring the website. This approach includes five steps, and beginning is basically to collect raw data of weblog file. And then preprocessing operation [11] is applied on it, to filter irrelevant information from raw log files. These steps include: i) filter out requests for pages generated by Common Gateway Interface (CGI) or other server-side scripts as we only consider static pages that are designed as part of a website structure, ii) ignore unsuccessful requests, and iii) remove requests to image files, as images will be automatically downloaded due to the HTML tags [15]. Processed information will be stored in a database. And then it is broken up into user sessions which contain one or more mini sessions. And mathematical programming model (MP) is applied. Lastly, links can be added based on the mini session as in MP model and restructuring of website is done with minimal changes.

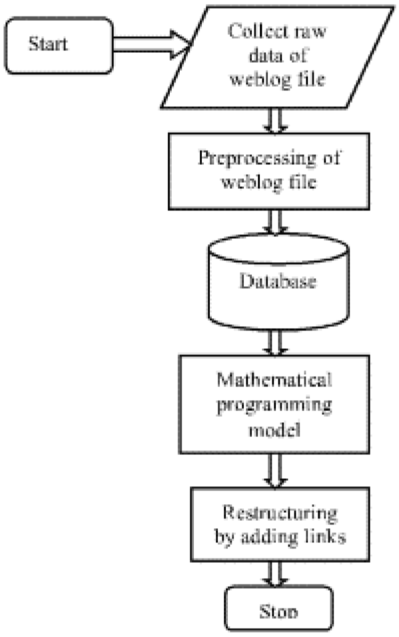
Start
Collect raw data of weblog file
Preprocessing of weblog file
Database
Mathematical programming model
Restructuring by adding links
Stop

Fig. 2. Proposed transformation approach for restructuring

IV. Methodology

In our model we will consider a website as a directed graph, with nodes as pages and arcs as links. Let N be the set of all webpages and λ_{ij}, where i, $j\epsilon$ N, denote page connectivity in the current structure, with $\lambda_{ij=1}$ which indicate page i has a link to page j, and λ_{ij}=0 otherwise. The current out-degree for page i is denoted by $W_i = \Sigma_{j\epsilon N}\lambda_{ij}$.

As the log files is broken up into the set of mini sessions, then we get the set T all of mini session. As in the previous example mini session S ϵ T, and we denote $tgt(S)$ the target page of S. Let $L_m(S)$ is number of paths in S and $L_p(K,S)$ be the length of the k^{th} path in S. Consider the example mini session S in Fig.1, then $L_m(S)$=3,$L_p(1,S)$=3, and $docno(1,1,S)=A$. We define $E= \{(i, j): i, j\epsilon N$ and $i \epsilon S$ and $j=tgt(S)\}$.Here E is the set of candidate links that can be selected to improve site structure which helps users to reach their targets faster.

Our problem is to determine whether to establish a link from i to j for (i, j) ϵ E. Let x_{ij} ϵ {0, 1} be decision variable such that x_{ij}=1, which indicates establishes the link.

Webmasters can set a goal for user navigation for each target page and is denoted by b_j, which is path threshold for j. We can determine that the user navigation goal is satisfied or not in S by comparing. If length of S is larger than b_j, then user navigation is below the goal and thus we need to alter the site structure. Otherwise, improvement is not needed in S.

The solution to a problem of improving the user navigation in a website with minimal changes to its current structure and can be formulated as an MP model below:

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

Subject to

$c^s_{kr}=\sum_{(i,j)\epsilon E} a^s_{ijkr}\ x_{i,j}$;$r=1,2, \ldots ,L_p(k\ ,S)$,

k=1, 2,…, $L_m(S)$, $\forall S \epsilon T^R$

The above function minimizes the cost to improve the website structure and costs consists of two components and are: a) the number of new links to be added as in first summation, and b) the penalties on pages containing more links than the threshold in the improved structure as in the second summation.

We have observed that some links may be left by users because of poor design or there may be some ambiguous terms used with links. So such links must be improved first before adding new links. Thus, the first summation that is $[1-\lambda_{ij}(1-\varepsilon)]$, where ε is a very small number, in the function so which allow the model to select an existing link whenever possible.

Consider, if (1-ε) is not present, then there will be no cost in selecting an existing link. For example, if (1-ε) is removed and the penalty term is not included, the costs of establishing new links, i.e., $\sum_{(i,j)\epsilon E} x_{ij}(1-\lambda_{ij})$ when selecting all existing links are same as the cost when nothing is selected. And this is because there is no cost in selecting an existing link, i.e., (1-λ_{ij})=0, when λ_{ij} =1.Thus, we add (1-ε) to impose a very small cost on improving an existing link such that the model will select the minimal number of existing links for improvement.

When there is no penalty term (m=0), the number of new links and the number of existing links to be improved are the same across different out-degree thresholds(C). This is because the out-degree threshold plays no role in the MP model if the penalty term is removed from the objective function. When the penalty term is imposed, i.e., m≠0, we find that while a larger multiplier for the penalty term (m) leads to more new links, it also adds fewer links to nodes having excessive links. This is anticipated because as m increases, the MP model would prefer to establish more links to pages with small out-degrees in order to prevent large penalties.

V. CONCLSION AND FUTURE ENHANCEMENT

Our proposed mathematical programming model is used to improve the web user navigation by minimal changes to the current web site structure. The model is particularly appropriate for informational websites whose contents are relatively stable over time. It improves a website rather than reorganizes it and hence is used regularly for website maintenance. We model the out-degree as a penalty term in the objective function and this provides flexible website structures and also offers a good balance between minimal changes to web site and reducing overload to users.

The paper can be further extended in several directions and one can be like the technique to identify the exact and accurate user's targets, which is critical to our model.

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,March 2013.

[2] Internet retailer, "Web Tech Spending Static-But High-for the Busiest E-commerce Sites", http://www.internetretailer.com/dailynews.asp?id=23440,2007

[3] Shaily G. Langhnoja , Mehul P, "Web usage mining using association rule mining", International conference on data mining, 2013.

[4] D. Dhyani, W.K and S.S Bhowmick, "A survey of web metrics", ACM Computing Surveys, vol 34, no.4, pp.469-503, 2002.

[5] M. Perkowitz and O .Etzioni, "Towards Adaptive Websites" Conceptual Framework and case study", Artificial Intelligence, vol. 118, pp.245-275,2000.

[6] M. kilfoil et al., "Towards an adaptive web: The state of the Art and science", proc. comm Network and services Research conf., pp. 119-130, 2003.

[7] R .Gupta, A. Bagchi and S. Sarkar, "Improving Linkage of Web pages", Informs, j. Computing, vol.19, no.1, pp.127-136, 2007.

[8] C.C. Lin, "optimal Web Site Reorganization Considering Information Overload and Search Depth," European j, Operational Research , vol 173,no.3,pp.839-848,2006.

[9] Navin Kumar Tyagi, A. K. Solanki,"Analysis of server log by web usage mining for website improvement" IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 8, July 2010.

[10] 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.

[11] B. Mobasher , r. Cooley, and j. Srivastava,"Automatic Adaptive Websites through Usage-Based Clustering of URLs," Proc. Worshop Knowledge and data Eng. Exchange ,1999.

[12] Y. Fu, M.Y. Shih, M. Creado, and C. Ju, "Reorganizing Using an Ant Colony System," Expert Systems with Applications,vol.37,no,no.12,pp.7598-7605,2010.

[13] R. Srikant and Y. Yang , "mining Web logs to improve Web site Organization," proc.10th Int'l Conf. World Wide Web,pp.430-437,2001.

[14] 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.

[15] M. Morita and Y. Shinoda, "Information Filtering Based on user Behavior analysis and best match text retrieval," Proc.17th Ann. Research and Development in information Retrieval, pp .272-281, 1994.

[16] P. Pirolli and S.K. Card, "Information foraging," Psychological Rev., vol.106, no.4, pp.643-675, 1999.