WEB RESTRUCTURING FOR EASY BROWSING

Kazade Sunita [1], Raut Dipali [2], Masal Archana [3], Palase Priyanka [4]
guide:Prof.Teke A.R.
Department Of Computer Engineering, SBPCOE, Indapur (Pune).

ABSTRACT- Website *design is easy task but, to navigate user efficiently is big challenge, one of the reason is user behaviour is keep changing and web developer or designer not think according to user's behaviour[4]. Designing well-structured websites to facilitate effective user navigation patterns has long been a challenge in web usage mining with various applications like navigation prediction and improvement of website management. By using the effective web structuring methods and web usage mining we will have a system that will ensure a completely different web navigation experience [11]. This paper addresses how to improve a website without introducing substantial changes. Specifically, we propose a mathematical programming model to improve the user navigation on a website while minimizing alterations to its current structure [12]. Results from extensive tests conducted on a publicly available real data set indicate that our model not only significantly improves the user navigation with very few changes, but also can be effectively solved. We have also tested the model on large synthetic data sets to demonstrate that it scales up very well. We proposed web transformation by using link mining*.

Keywords- Website design, User navigation, Web Mining, Mathematical Programming.

1. INTRODUCTION

Daily use of website is increased tremendously now a days. For each reason we are depending on websites. Also from company point of view main way to do advertise, May way to reach to customer with limited efforts is the website. As this growth in web usage current research in web era is in building the website which is very easy to get viewers inferring needs. So we may say that building a website is now not only the technical task but also an art to attract customers and give fast response for each intentional visit of customer.

Planning overall organized sites to encourage compelling client route has long been a test. An essential reason is that the web designers' understanding of how a site ought to be organized could be extensively unique in relation to that of the clients. While different strategies have been proposed to relink site pages to enhance safety utilizing client route information, the totally.

Despite the heavy and increasing investments in website design, it is still revealed, however, that finding desired information in a website is not easy and designing effective websites is not a trivial task. Galletta et al. indicate that online sales lag far behind those of brickand-mortar stores and at least part of the gap might be explained by a major difficulty user's encounter when browsing online stores. Palmer highlights that poor website design has been a key element in a number of high profile site failures. McKinney et al. also find that users having difficulty in locating the targets are very likely to leave a website even if its information is of high quality.

A primary cause of poor website design is that the web developers' understanding of how a website should be structured can be considerably different from those of the users. Such differences result in cases where users cannot easily locate the desired information in a website. This problem is difficult to avoid because when creating a website, web developers may not have a clear understanding of users' preferences and can only organize pages based on their own judgments. However, the measure of website effectiveness should be the satisfaction of the users rather than that of the developers. Thus, Webpages should be organized in a way that generally matches the user's model of how pages should be organized.

Previous studies on website has focused on a variety of issues, such as understanding web structures, finding relevant pages of a given page , mining informative structure of a news website, and extracting template from webpages .

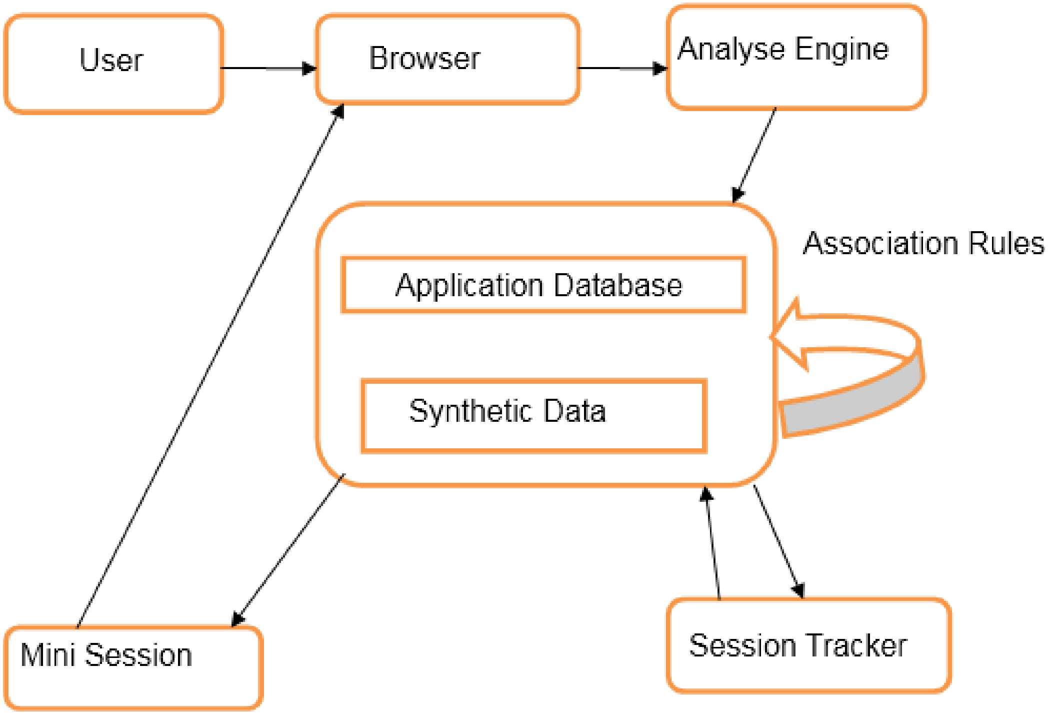

Fig 1. Architecture of user navigation system.

Our work, on the other hand, is closely related to the literature that examines how to improve website navigability through the use of user navigation data. Various works have made an effort to address this question and they can be generally classified into two categories to facilitate a particular user by dynamically reconstituting pages based on his profile and traversal paths, often referred as personalization, and to modify the site structure to ease the navigation for all users, often referred as transformation.

2. RELATED WORK

The growth of the Internet has led to numerous studies on improving user navigations with the knowledge mined from webserver logs and they can be generally categorized in to web personalization and web transformation approaches [11]. Web personalization is the process of "tailoring" webpages to the needs of specific users using the information of the users' navigational behavior and profile data. Perkowitz and Etzioni [11] 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 in user traversals, to facilitate user navigation. The methods proposed to create clusters of user's profiles from weblogs and then dynamically generate links for users who are classified into different categories based on their access patterns. Researchers also developed a hybrid personalization system that can dynamically switch between recommendation models based on degree of connectivity and the user's position in the site.

Web transformation, on the other hand, involves changing the structure of a website to facilitate the navigation for a large set of users instead of personalizing pages for individual users. An approach to reorganize webpages so as to provide users with their desired information in fewer clicks. However, this approach considers only local structures in a website rather than the site as a whole, so the new structure may not be necessarily optimal. Gupta et al.propose a heuristic method based on simulated annealing to relink webpages to improve navigability. This method makes use of the aggregate user preference data and can be used to improve the link structure in websites for both wired and wireless devices. However, this approach does not yield optimal solutions and takes relatively a long time (10 to 15 hours) to run even for a small website.

Lin develops integer programming models to reorganize a website based on the cohesion between pages to reduce information overload and search depth for users. In addition, a two-stage heuristic involving two integer-programming models is developed to reduce the computation time. However, this heuristic still requires very long computation times to solve for the optimal solution, especially when the website contains many links. Besides, the models were tested on randomly generated websites only, so its applicability on real websites remains questionable there are several remarkable differences between web transformation and personalization approaches. First, transformation approaches create or modify the structure of a website used for all users, while personalization approaches dynamically reconstitute pages for individual users. Hence, there is no predefined/built-in web structure for personalization approaches. Second, in order to understand the preference of individual users, personalization approaches need to collect information associated with these users (known as user profiles). This computationally intensive and time-consuming process is not required for transformation approaches. Third, transformation approaches make use of aggregate usage data from weblog files and do not require tracking the past usage for each user while dynamic pages are typically generated based on the users' traversal path. Thus, 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.

3. PROBLEM DEFINATION

Existing System

Difficulty in navigation is reported as the problem that 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. Therefore, Webmasters can ensure effective user navigation by improving the site structure to help users reach targets faster. This is especially vital to commercial websites, because easy navigated websites can create a positive attitude toward the firm, and stimulate online purchases, whereas websites with low usability are unlikely to attract and retain customers.

Proposed System

In previous system when we want to visit different e-commerce websites, we have to goes through many links. Because of this user have to read many unwanted information's including webpages. In this searching process there will be large time consuming process. If any new user want to access such websites then he can't go on target webpage, because there is no any navigation system for user. So we propose our system to overcome these problems. For example, some visits are clearly purposeless and finish abruptly at pages that cannot be target pages, so these sessions are more likely to be browsing sessions. To the best of our knowledge, there is no algorithm developed to distinguish between the two types of sessions and further investigation on this question is needed. While we did not explicitly separate searching sessions from browsing sessions, the pre-processing steps can help eliminate many purposeless browsing sessions. As a result, the final improved website structure mainly shortens searching sessions and also reduces purposeful browsing sessions. We propose a mathematical programming model to improve the user navigation on a website while minimizing alterations to its current structure. In short we will restructure the website such that it will transform current website into fast, easy browsing website for end user.

4. EVALUATING NAVIGATION EFFECTIVENESS

Recognizing the drawbacks of website reorganization approaches, we develop a mathematical programming model that facilitates user navigation on a website with minimal changes to its current structure. This model is particularly appropriate for informational websites whose contents are static and relatively stable over time. The number of outward links in a page, i.e., the out-degree, is an important factor in modelling web structure. Prior studies typically model it as hard constraints so that pages in the new structure cannot have more links than a specified out-degree threshold, because having too many links in a page can cause information overload to users and is considered undesirable. Our model formulates the Out-degree as a cost term in the objective function to penalize pages that have more links than the threshold so page's out-degree may exceed the threshold if the cost of adding such links can be justified, To assess the user navigation on the improved website, we partition the entire real data set into training and testing sets. We use the training data to generate improved structures which are evaluated on the testing data using simulations to appropriate the real usage. We define two metrics and use them to assess whether user navigation is indeed enhanced on the improved structure. First metric measures whether the average user navigation is facilitated in that improved website and the second metric measures how many users can benefit from the improved structure. Evaluation results confirm that user navigation on the improved website is greatly enhanced. Users are faced with a decision point at each page; they use information scent to evaluate the likely effort and the probability of reaching their targets via each link and make navigation decisions accordingly [7].

Consequently, a user is assumed to follow the path that appears most likely to lead him to the target. This suggests that a user may backtrack to an already visited page to traverse a new path if he could not locate the target page in the current path. Therefore, we use the number of paths a user has traversed to reach the target as a proximate measure to the number of times the user has attempted to locate one target. We use backtracks to identify the paths that a user has traversed, where a back track is defined as a user's revisit to a previously browsed page. The intuition is that users will backtrack if they do not find the page where they expect it. Thus, a path is defined as a sequence of pages visited by a user without backtracking, a concept that is similar to the maximal forward reference defined in Chen et al. Essentially, each backtracking point is the end of a path. Hence, the more paths a user has traversed to reach the target, the more discrepant the site structure is from the user's expectation.

5. MATHEMATICAL MODEL

Our problem can be regarded as a special graph optimization problem. We model a website as a directed graph, with nodes representing pages and arcs representing links.

Let N be the set of all webpages and λ_{ij}, where i,j $\in$ N, denote page connectivity in the current structure, with

λ_{ij} = 1 indicating page i has a link to page j, and

λ_{ij} =0 otherwise.

The current out-degree for page i is denoted by

$$\mathrm{Wi} = \sum_{j \in N} \lambda_{ij}$$

From the log files, we obtain the set T of all mini sessions. For a mini session S 2 T, we denote tgt(S) the target page of S. Let Lm(S) be the length of S, i.e., the number of paths in S, and Lp(k; S), for 1 k≤Lm(S), be the length of the kth path in S, i.e., the number of pages in the kth path of S. We further define docno(r; k; S), for 1≤k≤Lm(S) and 1≤r≤Lp(k; S), as the rth page visited in the kth path in S. Take the mini session S in Fig. 2 for example, it follows that Lm(S) ¼ 3; Lp(1; S) ¼ 3, and docno(1; 1; S) ¼ A, as this mini session has three paths and the first path has three pages (A, D, and H) in which page A is the first page. We define E ¼ f(i; j) : i; j 2 N and 9S 2 T such that i 2 S and j ¼ tgt (S)g and NE ¼ fi : (i; j) 2 Eg. In essence, E is the set of candidate links that can be selected to improve the site structure to help users reach their targets faster. Our problem is to determine whether to establish a link from i to j for (i; j) 2 E. Let xij 2 f0; 1g denote the decision variable such that xij ¼ 1 indicates establishing the link.

The problem of improving the user navigation on a website while minimizing the changes to its current structure can then be formulated as the mathematical programming model below:

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

subject to

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

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

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

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

The objective function minimizes the cost needed to improve the website structure, where the cost consists of two components:

1) the number of new links to be established (the first summation), and

2) the penalties on pages containing excessive links, i.e., more links than the out-degree threshold (Ci), in the improved structure (the second summation).

6. TECHNIQUES

Real Data Set

The real data set was collected from the Music Machines website and contained about four million requests that were recorded in a span of four months. This data set is publicly available and has been widely used in the literature. Table 6 shows the number of pages in the website that had out-degrees within a specified range. This website has in total 916 pages, of which 716 have an out-degree of 20 or less, with the majority (83 present) of the remaining pages having 40 links or less. Before analysis, we followed the log processing steps described in to filter irrelevant information from raw log files. These steps include: 1) filter out requests to 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, 2) ignore unsuccessful requests (returned HTTP status code not 200), and 3) remove requests to image files (.gif, .jpg, etc.), as images are in general automatically downloaded due to the HTML tags rather than explicitly requested by users. We utilized the page-stay time to identify target pages and to demarcate mini sessions from the processed log files. Three time thresholds (i.e., 1, 2, and 5 minutes) were used in the tests to examine how results changes with respect to different parameter values. Furthermore, we adapted the Algorithm proposed in to identify the backtracking pages in mini sessions, which are then used to demarcate the paths traversed to reach the targets.

Synthetic Data Set

In addition to the real data set, synthetic/artificial data sets were generated and considered for computational experiments to evaluate the scalability of our model with respect to the size of the website and the number of mini sessions. For this reason, the artificial website structures and mini sessions were generated to have similar statistical characteristics as the real data set. For instance, the average out-degree for pages in the real website is 15, so the link structure for the artificial website was generated in a way such that each page contained 15 links on average. Three websites consisting of 1,000, 2,000, and 5,000 webpages were constructed. Our approach for generating the link structure is similar to that described in. Particularly, to generate the link structure that contains 1,000 pages, we first constructed a complete graph of

1,000 nodes (pages) and each directed edge was assigned a random value between 0 and 1. Then, we selected the edges with the smallest 15,000 values to form the link structure, resulting in an average out-degree of 15 for this website. In a similar manner, we generated the link structures for other artificial websites. The mini sessions were generated in a slightly different manner. Specifically, we directly generated the set of relevant candidate links for each mini session instead of creating the user's traversal path. As a result, this allows us to directly apply the model on synthetic data sets. The sets of relevant candidate links in synthetic data sets has similar characteristics with those from the real one, comprising one to five relevant candidate links per each relevant mini session, with each link being randomly selected.

7. MODULES DESCRIPTION

Web Personalization

Web personalization is the process of "tailoring" webpages to the needs of specific users using the information of the users' navigational behavior and profile data. Jerkewitz and E-zine 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 in user traversals, to facilitate user navigation. The methods proposed by MO basher et al. and Yan et al. create clusters of users profiles from weblogs and then dynamically generate links for users who are classified into different categories based on their access patterns.

Web transformation

Web transformation, on the other hand, involves changing the structure of a website to facilitate the navigation for a large set of users instead of personalizing pages for individual users. Fu et al. describe an approach to reorganize web pages so as to provide users with their desired information in fewer clicks. However, this approach considers only local structures in a website rather than the site as a whole, so the new structure may not be necessarily optimal. Gupta et al. propose a heuristic method based on simulated annealing to relink web pages to improve navigability. This method makes use of the aggregate user preference data and can be used to improve the link structure in websites for both wired and wireless devices.

Maximal Forward Reference

We use backtracks to identify the paths that a user has traversed, where a back track is defined as a user's revisit to a previously browsed page. The intuition is that users will backtrack if they do not find the page where they expect it. Thus, a path is defined as a sequence of pages visited by a user without backtracking, a concept that is similar to the maximal forward reference defined in Chen et al. Essentially, each backtracking point is the end of a path. Hence, the more paths a user has traversed to reach the target, the more discrepant the site structure is from the user's expectation.

Mini Sessions

Recall that a mini session is relevant only if its length is larger than the corresponding path threshold. Consequently, only relevant mini sessions need to be considered for improvement and this leads to a large number of irrelevant mini sessions (denoted as TI) being eliminated from consideration in our MP model.

Out-Degree Threshold

Web pages can be generally classified into two categories- index pages and content pages. An index page is designed to help users better navigate and could include many links, while a content page contains information users are interested in and should not have many links. Thus, the out-degree threshold for a page is highly dependent on the purpose of the page and the website. Typically, the out degree threshold for index pages should be larger than that for content pages.

8. CONCLUSION

Main goal of this paper is to improve the navigation effectiveness of a website while minimizing changes to its current structure by using mathematical programming model. A critical issue that has not been examined in the literature. We will structure the web-site such that it will transform current website into fast, easy browsing website for end user. It improves a website rather than reorganizes it and hence is suitable for website maintenance on a progressive basis. The tests on a real website showed that our model could provide significant improvements to user navigation by adding only few new links. Optimal solutions were quickly obtained, suggesting that the model is very effective to real-world websites.

REFERENCES

[1] M. Perkowitz and O. Etzioni, "Towards Adaptive Web Sites: Conceptual Framework and Case Study,"Arti_cial Intelligence,vol. 118, pp.245-275, 2000.

[2] T. Nakayama, H. Kato, and Y. Yamane, "Discovering the Gap between Web Site Designers' Expectations and Users' Behavior," Computer Networks, vol. 33, pp. 811-822,2000.

[3] V. McKinney, K. Yoon, and F. Zahedi, "The Measurement of Web- Customer Satisfaction: An Expectation and Confirmation Approach," Information Systems Research,vol. 13, no. 3, pp. 296- 315, 2002.

[4] R. Gupta, A. Bagchi, and S. Sarkar, "Improving Linkage of Web Pages," INFORMS J. Computing, vol. 19, no. 1, pp. 127-136, 2007.

[5] C.C. Lin, "Optimal Web Site Reorganization Considering Information Overload and Search Depth," European J. Operational Research.

[6] X. Fang and C. Holsapple, "An Empirical Study of Web Site Navigation Structures' Impacts on Web Site Usability," Decision Support Systems, vol.43, no. 2, pp. 476-491,2007.

[7] Pingdom, "Internet 2009 in Numbers," http://royal.pingdom. com/2010/01/22/internet 2009-in-numbers/, 2010

[8] Joy Shalom Sona, Asha Ambhaikar "Reconciling the Website Structure to Improve the Web Navigation Effciency" June 2012.

[9] 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.

[10] C.S. Miller and R.W. Remington, Modeling Information Navigation:Implications for Information Architecture, Human Computer Interaction, vol. 19, pp. 225-271, 2004.

[11] 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.

[12] Min Chen and young U. Ryu "Facilitating Effective User Navigation through Web-site Structure Improvement" IEEE KDD vol no. 25.2013