IPHV7I2022X
International Journal Of Advanced Research and Innovation -Vol.7, Issue .II
ISSN Online: 2319 – 9253
Print: 2319 – 9245

MODELING THE INFORMATION NAVIGATION: IMPLICATIONS FOR INFORMATION ARCHITECTURE

[#1]S.Soumya- M.Tech Student,
[#2]G.Prasad-Assistant.Professor,
[#3]M.Sreelatha-Assoc.Professor, HOD,
Department of CSE,
SAHAJA INSTITUTE OF TECHNOLOGY AND SCIENCES FOR WOMEN, KARIMNAGAR.

Abstract: Website design is easy task but, to navigate user efficiently is big challenge, one of the reason is user behavior is keep changing and web developer or designer not think according to user's behavior. 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. 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. 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. In addition, we define two evaluation metrics and use them to assess the performance of the improved website using the real data set. Evaluation results confirm that the user navigation on the improved structure is indeed greatly enhanced. More interestingly, we find that heavily disoriented users are more likely to benefit from the improved structure than the less disoriented users.

Key Terms: ***Website design, user navigation, web mining, mathematical programming.***

I. INTRODUCTION

Internet are mostly used for information gathering, searching and commercial purpose.There is tremendous increase in number of internet users and it is increased day by day. Online shopping is an example where effective user navigation is very critical and important. Last few years commercial business has grown very much. People are doing online shopping more. So competition for companies doing ecommerce business has increased. The user may go to other website if navigation is not good. Hence companies are interested more to have better navigation of website. Despite that making efficient website is not trivial. The common approach to improve navigation is restructuring website but it is not good approach. User may loose interest because of familiarness caused by restructuring of a website. It is important to increase efficiency of a navigation keeping original structure intact.The reason of a poor navigation of a website is developed by developers understanding and views. Most of the times there is communication gap between developer and actual user. The information which is important for user that is may not be important in views of developer. User get frustrate if the information required is not found in minimum attempt. So website should be able to find required information with minimum navigation. Previous studies have found different issues on websites. In our work we are suggesting solution to improve website navigation.There are different studies have made to improve the efficiency of navigation. Those studies suggest transformation of website.

Data mining combines data analysis techniques with high-end technology for use within a process. The primary goal of data mining is to develop usable knowledge regarding future events.

The steps in the data mining process are:

- Problem definition
- Data collection and enhancement
- modeling strategies
- Training, validation, and testing of models
- Analyzing results
- Modeling iterations
- Implementing results.

There are lakhs of user for website since it is large source of information, web site also contain many links and pages every user require different pages at same time or same user may access different pages at different time. As user increases over www we need to make web intelligent

IPHV7I2022X
International Journal Of Advanced Research and Innovation -Vol.7, Issue .II
ISSN Online: 2319 – 9253
Print: 2319 – 9245

we concern here about intelligent website. To make web site intelligent we must know what is content of website, which are users and how website structured all this known as web mining. Web structure mining can be defined as mining of links between pages, which is also called as hyperlinks which enable user to access web sites in form of URL and navigate user. In web structure mining developer uses the data from web usage and change structure of web site, pages which is most visited and user spent more time is linked to the start page.

The goal of a Web site is to meet the needs of its users. As a result, as the interests of its users change over the time, a static Web site that does not change itself will soon become outdated and less useful. Accordingly, a Web site must constantly examine site use, and modify itself accordingly to best serve its users. In other words, Web sites should be adaptive. An adaptive Web site has been defined as a Web site that semi-automatically improves its organization and presentation by learning from visitor access patterns (Perkowitz and Etzioni, 1998). In this paper, an attempt is made to build adaptive Web sites, which improve their navigation based on access patterns of its users. An approach for reorganizing Web sites based on user access patterns is proposed. Our goal is to build adaptive Web sites by evolving site Structure to facilitate user access. To be more specific, we aim to build Web sites that provide users with the information they want with fewer clicks. This minimizes the effort on the user's side. By analyzing the usage of a Web site and the structure of the Web site, modifications to the Web site structure are found to accommodate changes in access patterns of its users. These modifications will be suggested to the Webmaster for consideration and implementation.

Motivation for choosing web structure mining is: since web site is big source of information, but users mostly browsing useless page which irritates user and user lost interest from searching data over website. 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; however, the measure of website effectiveness should be the satisfaction of the users rather than that of the developers. Thus, Web pages should be organized in a way that generally matches the user's model of how pages should be organized.

Despite the heavy and increasing investments in website design, it is still revealed, however, that finding desired information in a website is not easy [4] and designing effective websites is not a trivial task [5], [6]. Galletta et al. [7] 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 users encounter when browsing online stores. Palmer [8] highlights that poor website design has been a key element in a number of high profile site failures. McKinney et al. [9] 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.

Improving site structure in which a well-constructed navigation schemes has an important impact to well rank in the search engines. Ensure that your web design has a proper navigation menu so that your visitors can use the website with ease. This way, they will be able to find out what they were looking for in the website in the first place. Google's search results are provided at a page level but it also likes to have a sense of what role a page plays in the site. **Best Practices for Site Navigation**

Create a naturally flowing hierarchy. As what I've stated earlier, your website should flow naturally where users can access first from general content to the more specific content they want on your site.

- Use mostly text navigation instead of images or animation.
- Put an HTML site map page on your site, and use an XML Sitemap file.
- Webmasters are also advised to have useful 404 page that guides the user back to a relevant section or page with a link back the home page in case site visitor encounters a broken link or types in an incorrect URL. If a search engine comes across such an error, it can have a negative impact on your search engine visibility. Google provides a 404 widget that you can embed in your 404 page to automatically populate it with many useful features.

II.RELATED WORK

Web usage mining, author in [1] explains about weblogs like who accessed order of page request, total time for page view. This paper includes several pre-processing like; **1: Data cleaning**-It is method of removing irrelevant items or logs like removing of file with .gif and .jpg extensions.

2: User identification-It involves USER ID for each user to provide uniqueness even different users are on same IP.

3: Session identification- This is defines according to time i.e. time between page request and page close or time out.

4: Path completion- It is defined as if some information or page is important and mostly accessed but not recorded in logs and not linked cause problem.

5: Formatting- It is method of converting transactions or logs it to a format of data mining like removal of numeric value for determining association rules.

Access Information Collection:

In this step, the access statistics for the pages

are collected from the sessions. The data obtained will later be used to classify the pages as well as to reorganize the site. The sessions obtained in server log preprocessing are scanned and the access statistics are computed. The statistics are stored with the graph that represents the site obtained in Web site preprocessing. The obvious problem is what should be done if a page happens to be the last page of a session. Since there is no page requested after that, we really couldn't tell the time spent on the page. Therefore, we keep a count for the number of times that the page was the last page in a session. The following statistics are computed for each page:

• Number of sessions in which the page was accessed;
• Total time spent on the page;
• Number of times the page is the last requested page of a session.

Page Classification:

In this phase, the pages on the Web site are classified into two categories: index pages and content pages (Scime and Kerschberg, 2000). An index page is a page used by the user for navigation of the Web site. It normally contains little information except links. A content page is a page containing information the user would be interested in. Its content offers something other than links. The classification provides clues for site reorganization. The page classification algorithm uses the following four heuristics.

(1) File type.

An index page must be an HTML file, while a content page may or may not be. If a page is not an HTML file, it must be a content page. Otherwise its category has to be decided by other heuristics.

(2) Number of links.

Generally, an index page has more links than a content page. A threshold is set such that the number of links in a page is compared with the threshold. A page with more links than the threshold is probably an index page. Otherwise, it is probably a content page.

(3) End-of-session count.

The end-of-session count of a page is the ratio of the number of time it is the last page of a session to the total number of sessions. Most Web users browse a Web site to look for information and leave when they find it. It can be assumed that users are interested in content pages. The last page of a session is usually the content page that the user is interested in. If a page is the last page in a lot of sessions, it is probably a content page; otherwise, it is probably an index page. It is possible that a specific index page is commonly used as the exit point of a Web site. This should not cause many errors at large.

(4) Reference length.

The reference length of a page is the average amount of time the users spent on the page. It is expected that the reference length of an index page is typically small while the reference length of a content page will be large. Based on this assumption, the reference length of a page can hint whether the page should be categorized as an index or content page. A more detailed explanation is given below, followed by a page classification algorithm based on these observations and heuristics.

Reference Length Method

The reference length method for page classification (Cooley, 2000) is based on the assumption that the amount of time a user spends on a page is a function of the page category. The basic idea is to approximate the distribution of reference lengths of all pages by an exponential distribution. A cut-off point, t, for reference length, can be defined as follows.

$$t = -\ln(1 - \gamma)/\lambda$$

where γ = percentage of index pages,
λ = reciprocal of observed mean reference length of all pages.

Algorithm for Site Reorganization

Based on the cases discussed in the previous Section, the algorithm for site reorganization is outline as follows.

(1) Initialize a queue Q
(2) Put children of the home page in Q (3) Mark the home page
(4) While Q not empty (5) *current page* = pop(Q) (6) Mark *current page*
(7) For each parent p of *current page*
(8) Local adjustment according to the cases in the previous section.
(9) Push children (maybe merged) of *current page* into Q if they are not marked

III.PROPOSED METHODS

In candidate link set first we find out the existent links. Sometimes it is possible that existent links are found in candidate link set. This is because of location of link or visibility of links is not properly placed. So, first we need to improve those links. Next, we need to find out relevant candidate link set. Relevant candidate links means links having larger path than path threshold. For filtering candidate links first use path threshold. Path threshold means maximum number of path allows reaching target page. Path threshold 1 means we consider only those link by which user reach target page in first path. Then we select those links. We cannot add all these links. We can add only those candidate links that are common for all users. For these we need to find out similarity between links. For finding similarity between links we use Dice's coefficient index. We apply index on relevant candidate link set. We obtain the dice's index between 0 and 1. 0 means dissimilar and 1 means exact similar. Then we use KNN classifier. KNN means k nearest neighbour classifier.

IPHV7I2022X
International Journal Of Advanced Research and Innovation -Vol.7, Issue .II
ISSN Online: 2319 – 9253
Print: 2319 – 9245

Evaluation of improved website:

Perform evaluation on improved website structure to assess whether its navigation effectiveness is indeed enhanced by approximating its real usage. Specifically we partition the real data ste into a training set and a testing set. We generate the improved structure using the training data and then evaluate it on the testing data using two metrics: the average number of paths per mini session and the percentage of mini sessions enhanced to a specified threshold. The first metric measured whether the improved website structure can facilitate users to reach their targets faster than the current one on average, and second metric measures how likely users suffering navigation difficulty can benefit from the improvements made to the site structure. The evaluation procedure using the first metric consists of three steps as follows:

1. Apply the MP model on the training data to obtain the site of new links and links to be improved.
2. Acquire from the testing data the mini sessions that can be improved, i.e., having two or more paths their length i.e., number of paths and the set of candidate links that can be used to improve them.

Path threshold:

It represents the goal for user navigation that the improved structure should meet and can be obtained in several ways. First it is possible to identify when visitors exit a website before reaching the targets from analysis of weblog files. It helps to make good estimation for the path thresholds.

Second surveying website structures visitors can help better understanding users expectations and make reasonable selections on the path threshold values. Third from firms collected large amounts of client-side web usage data over a wide range of websites. Analyzing each data sets can also provide good insights into selection of path threshold values for different types of websites.

Out-degree threshold:

Webpages can be generally classified into two categories [29]: 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 out-degree threshold for a page is highly dependent on the purpose of the page and the website. The out-degree threshold for index pages should be larger than content pages. In general out-degree threshold could be set at a small value when most webpages have relatively few links and new liks added the threshold can be gradually increased.

IV. IMPLEMENTATION METHODS & MODELS

4.1 Mathematical Model:- In this section we will discuss mathematical model of proposed system: Website is represented as a directed graph. Let nodes representing pages and arc representing links. Let n be the number of pages of the website. User may visit more than one page by accessing links during user session. Let Pi={P1,P2,…..,Pm} be the set of m link accessed by user Ui where i={1,2,….,m}. Let Pm be the page user Ui is looking for. Means Pm is the target page. Let U= {U1, U2 ,…,Un} be the set of N users. Let C= {C1, C2 ,…., Cp} be set of p candidate links that need to redesign and relink. We use Dice's coefficient index. Dice's coefficient index is used for checking similarity between two string set.

$$D(A, B) = \frac{2*|A \sqsubset B|}{|A|+|B|}$$

The aim of this paper is to identify links that are used to redesign. Usage pattern is used to analyze user's behavior on the Web. By analyzing user's behaviour we find out target page. And find out candidate links so that user can access target page faster.

4.2 Algorithm:-

Mining Candidate Link Algorithm Input: Pi – Users Profile data

Output: Links that can be use for redesign Steps-

1: We identify the usage pattern of users λ from Pi = {P1, P2 ,…., Pm} set for user Ui to get link Pm

2: For every access link set obtain the set of candidate links {C1, C2,….,Cp}

3: For all users and their all access link set obtain the set of candidate links.

4: Obtain the Dice's similarity coefficient for all candidate link set.

5: Apply KNN classifier.

6: Then the links having problem for maximum number of users are selected for redesign the website**.**

V. CONCLUSION

In this paper we propose a mathematical model to improve website structure for effective user navigation by minimum changes to its current structure. We propose an algorithm which gives set of links that can be re-link to improve website structure. This approach is a better solution than whole redesigning of a website. This algorithm is best suited for website with mostly static content such as educational website. This model can be extended to improve web structure of dynamic website. Our model is particularly appropriate for informational websites whose contents are relatively stable overtime. It improves website rather than reorganize it hence it is suitable for website maintenance on a

progressive basis. Our model has a constraint for out-degree threshold which is motivated by cognitive reasons. The model can be further improved by incorporating additional constraints that can be identified by incorporating additional constraints that can be identified using data mining methods.

REFERANCES:

[1] Pingdom, "Internet 2009 in Numbers," http://royal.pingdom. Com/2010/01/22/internet-2009-in-numbers/, 2010.
[2] J. Grau, "US Retail e-Commerce: Slower But Still Steady Growth,"http://www.emarketer.com/ Report.aspx?code=emarketer_2000492, 2008.
[3] Internetretailer, "Web Tech Spending Static-But High-for the Busiest E-Commerce Sites," http://www.internetretailer.com/ dailyNews.asp?id = 23440, 2007.
[4] D. Dhyani, W.K. Ng, and S.S. Bhowmick, "A Survey of Web Metrics," ACM Computing Surveys, vol. 34, no. 4, pp. 469-503, 2002.
[5] 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.
[6] J. Lazar, Web Usability: A User-Centered Design Approach. Addison Wesley, 2006.
[7] D.F. Galletta, R. Henry, S. McCoy, and P. Polak, "When the Wait Isn't So Bad: The Interacting Effects of Website Delay, Familiarity, and Breadth," Information Systems Research, vol. 17, no. 1, pp. 20- 37, 2006.
[8] J. Palmer, "Web Site Usability, Design, and Performance Metrics," Information Systems Research, vol. 13, no. 2, pp. 151-167, 2002.
[9] V. McKinney, K. Yoon, and F. Zahedi, "The Measurement of Web-Customer Satisfaction: An Expectation and Disconfirmation Approach," Information Systems Research, vol. 13, no. 3, pp. 296-315, 2002.
[10] 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.
[11] M. Perkowitz and O. Etzioni, "Towards Adaptive Web Sites: Conceptual Framework and Case Study," Artificial Intelligence, vol. 118, pp. 245-275, 2000.
[12] J. Lazar, User-Centered Web Development. Jones and Bartlett Publishers, 2001.
[13] Y. Yang, Y. Cao, Z. Nie, J. Zhou, and J. Wen, "Closing the Loop in Webpage Understanding," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 5, pp. 639-650, May 2010.
[14] J. Hou and Y. Zhang, "Effectively Finding Relevant Web Pages from Linkage Information," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 4, pp. 940-951, July/Aug. 2003.
[15] H. Kao, J. Ho, and M. Chen, "WISDOM: Web Intrapage Informative Structure Mining Based on Document Object Model," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 5, pp. 614-627,May 2005.
[16] H. Kao, S. Lin, J. Ho, and M. Chen, "Mining Web Informative Structures and Contents Based on Entropy Analysis," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 1, pp. 41-55, Jan. 2004.
[17] C. Kim and K. Shim, "TEXT: Automatic Template Extraction from Heterogeneous Web Pages," IEEE Trans. Knowledge and Data Eng., vol. 23, no. 4, pp. 612-626, Apr. 2011.
[18] 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.
[19] R. Gupta, A. Bagchi, and S. Sarkar, "Improving Linkage of Web Pages," INFORMS J. Computing, vol. 19, no. 1, pp. 127-136, 2007.
[20] 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.
[21] M. Eirinaki and M. Vazirgiannis, "Web Mining for Web Personalization," ACM Trans. Internet Technology, vol. 3, no. 1, pp. 1-27, 2003.
[22] 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.
[23] B. Mobasher, R. Cooley, and J. Srivastava, "Automatic Personalization Based on Web Usage Mining," Comm. ACM, vol. 43, no. 8, pp. 142-151, 2000.
[24] 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.
[25] 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.
[26] 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. CHEN AND RYU: facilitating effective user navigation through website structure improvement 587.
[27] 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.
[28] 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.
[29] Y. Fu, M.Y. Shih, M. Creado, and C. Ju, "Reorganizing Web Sites Based on User Access Patterns," Intelligent Systems in Accounting, Finance and Management, vol. 11, no. 1, pp. 39-53, 2002.
[30] M.D. Marsico and S. Levialdi, "Evaluating Web Sites: Exploiting User's Expectations," Int'l J. Human-Computer Studies, vol. 60, no. 3, pp. 381-416, 2004.
[31] J. Palmer, "Designing for Web Site Usability," Computer, vol. 35, no. 7, pp. 102-103, June 2002.
[32] J. Liu, S. Zhang, and J. Yang, "Characterizing Web Usage Regularities with Information Foraging Agents," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 5, pp. 566-584, May 2004. [33] P. Pirolli and S.K. Card, "Information Foraging," Psychological Rev., vol. 106, no. 4, pp. 643-675, 1999.
[34] E.H. Chi, P. Pirolli, and J. Pitkow, "The Scent of a Site: A System for Analyzing and Predicting Information Scent, Usage, and Usability of a Web Site," Proc. SIGCHI Conf. Human Factors in Computing Systems, pp. 161-168, 2000.
[35] C. Olston and E.H. Chi, "ScentTrails: Integrating Browsing and Searching on the Web," ACM Trans. Computer-Human Interaction, vol. 10, no. 3, pp. 177-197, 2003.