Efficient Mining and Discovery of Users Interesting Patterns Using Path Traversal Graph
Priyanka M. Wani¹, Rahul Chinchore²
PG Student¹, Assistant Professor²
Department of CSE
G.H. Raisoni Institute of Engineering and Management, Jalgaon (Ms), India

Abstract:
For effective user navigation, designing of well-structured websites are a challenge over faculty of computer engineering. How websites having their ideal structure can be considered different from users is the main reason of web developers. For effective user navigation of websites, various methods that proposed to relinking web pages, the complete fulfillment of new website structure are may be unpredictable and the measurement analysis or cost of disorientation of users are still incomplete their analysis. Website design is an easy task but, to navigate user efficiently is a big challenge, one of the reasons is user behavior is keep changing and web developer or designer not think according to user’s behavior. The main basic reason behind that is understanding the concept of how developers developing a website structure according to users need rather than understandings of users. 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. The proposed Naive Bayes algorithm addresses how to improve website structure without modifying small changes that can occur during implementation. In addition the proposed system used data mining techniques to find out users target pages i.e. in which Pattern users want to access the web page and also facilitate effective user navigation by reorganizing the website structure.

Keywords: User Navigation, Data mining, Web mining

I. INTRODUCTION
There are millions of user for websites since it is a large source of information, web site also contain many links and pages every user require different pages at the same time or same user may access different pages at different time. As user increases over www we need to make web intelligent, 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 design encompasses many different skills and disciplines in the production and maintenance of websites. The different areas of web design include web graphic design; interface design; authoring, including standardized code and proprietary software; user experience design; and search engine optimization. Often many individuals will work in teams covering different aspects of the design process, although some designers will cover them all. The term web design is normally used to describe the design process relating to the front-end (client side) design of a website including writing markup. Web design partially overlaps web engineering in the broader scope of web development. Web designers are expected to have an awareness of usability and if their role involves creating markup then they are also expected to be up to date with web accessibility guidelines. Previous studies on website has discussed on a variety of issues, such as, extracting template from WebPages, mining informative structure of a news website, finding relevant pages of a given page, and understanding web structures. On the other hand, our work is closely related examines how to improve website navigability through the use of user navigation data. Various works have made an effort to address this question and generally it can be classified into two categories: first is to facilitate a particular user by dynamically reconstituting pages based on his profile and traversal paths, often referred as personalization, and second is to modify the site structure to ease the navigation for all users, often referred as transformation. We perform experiments on a data set which is collected from a real websites. The results of these experiments indicate that our model can significantly improve the size structure with only few changes. Besides all this, the optimal solutions of the mathematical model are effectively obtained, suggesting that our model impractical to real-world websites. We also tested our model with synthetic data sets that are larger than the real data set. The solution times are remarkably low for all cases tested, ranging from fraction of second to up to 34 seconds. The solution times are shown to increase reasonably with the size of the website, indicating that the proposed MP model can be easily scaled to a large extent. Motivation for choosing web structure effective user navigation through website structure improvement 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.

II. LITERATURE SURVEY
The purpose of this review is to report, evaluate, and discuss the findings from research. A particular focus of this review is to facilitating effective user navigation through website structure improvement.
• May Wang, Benjamin Yen, The study aims to improve Web navigation efficiency by reorganizing Web structure. Navigation efficiency is defined mathematically for both navigation with / without target destination pages, e.g. for experienced and new users. To help experienced users not to lose their orientation, structure stability is taken into consideration. Stability constraint can also help website designers control the maintaining effort of Web. This study proposes a mathematical programming method to reorganize Web structure in order to achieve better navigation efficiency. Designer can specify the user requirements and how stable the website structure should be. An e-banking example is given to illustrate how the method works in scenarios where user surfs with target destination. This study has the advantage of assessing and improving navigation efficiency and of relieving the designer of tedious chore to modify the structure in transformation.

• Devenish Dyane, Ng EeKeong and Sourav S Bhowmick, The unabated growth and increasing significance of the World Wide Web has resulted in a flurry of research activity to improve its capacity for serving information more effectively. But at the Heart of these efforts lie implicit assumptions about “quality” and “usefulness” of Web resources and services. This observation points towards measurements and models that quantify various attributes of web sites. The science of measuring all aspects of information, especially its storage and retrieval or Informetrics has interested information scientists for decades before the existence of the Web. Is Web Informetrics any different, or is it just an application of classical Informetrics to a new medium? In this paper, we examine this issue by classifying and discussing a wide ranging set of Web metrics. We present the origins, measurement functions, formulations and comparisons of well-known Web metrics for quantifying Web graph properties, web page significance, web page similarity, search and retrieval, usage characterization and information theoretic properties. We also discuss how these metrics can be applied for improving Web information access and use.

• Ramakrishnan Srikant and Yinghui Yang, Many websites have a hierarchical organization of content. This organization may be quite different from the organization expected by visitors to the website. In particular, it is often unclear where a specific document is located. In this paper, we propose an algorithm to automatically find pages in a website whose location is different from where visitors expect to find them. The key insight is that visitors will backtrack if they do not find the information where they expect it: the point from where they backtrack is the expected location for the page. We present an algorithm for discovering such expected locations that can handle page caching by the browser. Expected locations with a significant number of hits are then presented to the website administrator. We also present algorithms for selecting expected locations (for adding navigation links) to optimize the benefit to the website or the visitor. We ran our algorithm on the Wharton business school website and found that even on this small website, there were many pages with expected locations different from their actual location.

• Mingjun Li, Mingxin Zhang, Jinlong Zheng and Ying Lu, The improved algorithm selects expected locations (for adding navigation links) in backtracks set at the point of the earlier and the less backtracks, which avoids effectively negative impact to the accuracy of the overall analysis by the long access sequence. The experimental results show that the improved algorithm can find expected pages effectively, thus can achieve the target of adjustment and reorganization of website.

• Ms. Jissin Mary Kunjukutty and Ms. A. Priya, Web mining techniques are used to analyze web resource details. Content mining, structure mining and usage mining are the main types of web mining. Web page contents are analyzed in the content mining process. Structure mining technique is used to analyze the web site and page layouts. User access details are analyzed using usage mining methods. Web site structures are altered to improve the user navigations. Web personalization method reconstructs the page links with reference to the traversal path and profile of a particular user. Transformation mechanism is applied to modify the site structure for all users. User navigation data is used to re-link web pages to improve navigability. The out degree refers the number of outward links in a page. Out degree threshold is used to control the number of links in a page to minimize information overload in a page. Targeted pages are identified with page-stay time information. Mini sessions are identified with processed logs and path threshold information. Mathematical programming model is used to improve the user navigation on a website with minimum alteration in the current structure. Backtracking algorithm is used to estimate backtracking pages from mini sessions. Average user navigation and benefited user count metrics are used to evaluate the navigation performance. The web site restructuring scheme is enhanced with frequent pattern mining mechanism. Dynamic out degree threshold estimation model is adapted for the system. Target page identification process is performed with sequential patterns. Relative link information is used for the navigation pattern analysis.

III. PROBLEM DEFINITION

An essential driver of poor site outline is that the web engineers’ understanding of how a site ought to be organized can be extensively not the same as those of the clients. Such problems bring about situations where clients can’t undoubtedly spot the sought data in a site. This issue is hard to dodge on the grounds that when making a site, web engineers might not have an agreeable understanding of clients’ inclination and can just arrange pages focused around their judgments. Notwithstanding, the measure of site adequacy ought to be the fulfillment of the clients instead of that of the designers. Therefore, Web pages ought to be sorted out in a manner that for the most part matches the client’s model of how pages ought to be composed. Results from broad tests directed on a freely accessible true information set demonstrate that our model not just altogether enhances the client route with not very many progressions, additionally can be adequately illuminated. What’s more, we characterize two assessment measurements and use them to evaluate the execution of the enhanced site utilizing the true information set. Assessment results affirm that the client route on the enhanced structure is for sure enormously improved. All the more interestingly, we find that intensely perplexed clients are more inclined to profit from the enhanced structure than the less confused clients. The World Wide Web acts as an interactive and popular way to transfer information. Due to the enormous and diverse information on the web, the users cannot make use of the information very effectively and easily. People get more and more dependent on internet for getting any kind of information. Record of different web user’s web using pattern are get stored in web log repository.
which are great source of knowledge about user’s navigation. With increasing the use of internet, number of web sites and web pages are increasing rapidly. So analyse and discovering user’s interesting patterns are necessary for web administrator and recommendation system. Web usage mining is a part of data mining. In this, mining techniques are applied to web data for finding user’s interesting patterns. That means in which patterns user want to access web pages and web-sites. While website visitor visited web pages from one web page to another web page, certainly it creates a path. That path shows how the user traverses from one web page to another web page directly. There is also one condition is that visitors wish to one web page but there may be no direct web page access so visitor has to follow via-links which shows indirect access of web page. Hence for whole scenario of web accessing Web Navigation Pattern is basic factor to identify intention of user.

IV. IMPLEMENTATION DETAILS

4.1 Existing Work
A most important cause of poor website should be structured and can be considerably different from those of the users. Such differences result in cases where users cannot easily find the desired information in a website. This issue is difficult to handle because when creating a website, web developers don’t have clear understanding of users’ preferences and can only organize pages based on their own ideas.

Existing System Algorithm:
In an existing system k-means algorithm is use structure improvement.

Input: set of k means ml
(1), Mk (1)

Assignment step: Assign each observation to the cluster whose mean yields the least within sum of squares (WCSS). Since the sum of squares is the squared Euclidean, this is intuitively the “nearest” mean. (Mathematically, this means partitioning the observations diagram generated by the means). Where each is assigned to exactly

Update step: Calculate the new means to be the arithmetic mean is a sum of squares (WCSS) objective.

4.2 Propose Work
In this project we are presenting and extend navigation through website structure improvement through website structure. This approach delivers the efficiency as well as effectiveness of proposed methods for improvement of website creating website web developers don’t have clear understanding of project our main aim is to present approaches to overcome the limitations. In this will add new algorithm which will efficiently do the structure. For this purpose we are using

Algorithm
Input: Real websites dataset
• Random sampling: To handle large data sets, we do generally the random sample fits in a tradeoff between accuracy and efficiency. Design is that the web developers perceptive of how a developer may not have a clear used for effective user navigation through website: according to the one, even if it could be is assigned to two or more of them. Centroids of the observations in the new clusters least-squares estimator, this also minimizes the within extending the Naive Bayes clustering algorithm improvement. The current method is dealing with website. However this method is suffered from limitations like clients’ require existing method we improved user navigation through website NAIVE BAYES clustering algorithm. Random sampling and draw a sample main memory. Also because of the random sampling there is users effective within-cluster, within-cluster for user. User navigation. While requirement. Thus in this data set.

• Partitioning for speed up: The basic idea is to partition the sample space into p partitions. Each partition contains n/p elements. Then in the first pass partially cluster each partition until the final number of clusters reduces to n/pq for some constant q = 1. Then run a second clustering pass on n/pq partial clusters for all the partitions. For the second pass we only store the representative points since the merge procedure only requires representative points of previous clusters before computing the new representative points for the merged cluster. The advantage of partitioning the input is that we can reduce the execution times.

• Labeling data on disk: Since we only have representative points for k clusters, the remaining data points should also be assigned to the clusters. For this a fraction of randomly selected representative points for each of the k clusters is chosen and data point is assigned to the cluster containing the representative point closest to it.

Output: Set of web links that needs to be redesign and relink. The algorithm Steps for page classification is outlined as follows.

1) l = 1/(mean reference length of all pages)
2) t = -ln(1-)
3) For each page p on the Web site
4) If p’s file type is not static page or
5) P’s end-of-session count > count_threshold
6) Mark p as a content page
7) Else If p’s number of links > link_threshold
8) Mark p as an index page
9) Else If p’s reference length < t
10) Mark p as an index page
11) Else
12) Mark p as a content page

Algorithm for Site Reorganization
Based on the previous steps, the algorithm for site reorganization is outline as follows.

1) Initialize a queue Q
2) Put children of the page in Q
3) Mark the 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
9) Push children (maybe merged) of current_page into Q if they are not marked.
V. RESULTS AND ANALYSIS

5.1 EVALUATION PROCEDURE

We used simulations to approximate the real usage and to evaluate how the user navigation could be enhanced in the improved website structure. The use of simulation for website usability evaluation is very popular and has been widely used in modeling users' choices in web navigation and usability test. However, simulation studies often have to make simplifying assumptions in order to simulate real-life scenarios, posing questions on the generalizability of the results. In the context of our simulation approach, we assume that users would find their target pages effectively through a new/improved link if it exists. In practice, certain criteria related to the visual design of web interfaces need to be followed in order to effectively apply the suggested changes to a website. We note that there exist an abundant literature on both webpage design and hyperlink design. We do assume that Webmasters follow the guidelines and suggestions from such studies when creating and editing links and designing webpages. Consequently, in the simulation approach used for user navigation evaluation, we assume that new links are carefully designed and existing links are appropriately edited. In addition, they should also be placed in proper places for users to easily locate. Thus, these links should provide users with accurate knowledge on the contents on the other end of a link and help them make correct selections. Because of the assumption made for the new and improved links, the claimed benefit can be interpreted as the upper bound and optimal benefit of our model. However, we would like to claim that improved and newly added links could guide users to find their target pages more efficiently to some extent. This is because: 1) our method establishes efficient paths to target pages that were not available in the website structure before optimization, and 2) our method suggests improving links that would lead to users' target pages efficiently but missed by users (since they did not know what these links would lead to), so that more efficient navigation can be facilitated. Since our evaluation is simulation based, a usability study involving real users may help strengthen the results of our study and deserves further investigation. However, we note that such usability studies are generally more expensive and time consuming in the context of website evaluation and hence are usually conducted on small sized websites. In contrast, simulation can be easily implemented, quickly performed for various parameter settings, and tested on a large scale.

5.2 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 set 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.

Figure 5.1 Browse dataset which is consisting of page numbers that requested by users to find relevant information.

Figure 5.2 Filtering all request from dataset which are more than threshold value

Figure 5.3 Output contains total number of mini session, Traversal path of users, users target pages.
VI. CONCLUSION AND FUTURE SCOPE

Website rearranges provides user to improve navigability, this analyzes the wide areas of website rearrangement and link examination on the basis of web logs and user session and data mining techniques applied on web data, which provides user to reach target. The Propose methodology conclude that NAIVE BAYES algorithm is use to improve the navigation effectiveness of a website while minimizing changes to its current structure. It improves a website rather than reorganizes it and hence it is suitable for website maintenance on a progressive basis. The tests on a real websites dataset showed that NAIVE BAYES algorithm could provide significant improvements to user navigation by adding only few new links. In future if data mining methods find that most users access the finance and sports pages together, then this information can be used to construct an additional constraint.

VII. REFERENCES


[7]. A. J. L. Yao-Te Wang, \Mining web navigation patterns with a path traversal graph,” vol. 38, June 2011.


[9]. w. s. ma shu yue, liu wen cai, \The study on the preprocessing in web log mining.”


Table. 5.6 Existing System Results

<table>
<thead>
<tr>
<th>Multiplier</th>
<th>Penalty</th>
<th>Path Threshold</th>
<th>New Links</th>
<th>Improved Links</th>
<th>Excessive Links</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>9180</td>
<td>1806</td>
<td>7354</td>
<td>0.109</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>1643</td>
<td>648</td>
<td>1256</td>
<td>0.484</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>590</td>
<td>375</td>
<td>441</td>
<td>0.391</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>9180</td>
<td>1798</td>
<td>7910</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1659</td>
<td>662</td>
<td>940</td>
<td>0.625</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>613</td>
<td>384</td>
<td>192</td>
<td>0.422</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>9180</td>
<td>1798</td>
<td>7310</td>
<td>0.140</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1681</td>
<td>671</td>
<td>926</td>
<td>0.641</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>647</td>
<td>383</td>
<td>170</td>
<td>0.563</td>
<td></td>
</tr>
</tbody>
</table>

Table. 5.7 Proposed System Results

<table>
<thead>
<tr>
<th>Multiplier</th>
<th>Penalty</th>
<th>Path Threshold</th>
<th>New Links</th>
<th>Improved Links</th>
<th>Excessive Links</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>9180</td>
<td>1806</td>
<td>7354</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>1643</td>
<td>648</td>
<td>1256</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>590</td>
<td>375</td>
<td>441</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>9180</td>
<td>1798</td>
<td>7910</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1659</td>
<td>662</td>
<td>940</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>613</td>
<td>384</td>
<td>192</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>9180</td>
<td>1798</td>
<td>7310</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1681</td>
<td>671</td>
<td>926</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>647</td>
<td>383</td>
<td>170</td>
<td>0.5</td>
<td></td>
</tr>
</tbody>
</table>


