International Journal of Linguistics and Computational Applications (IJLCA) ISSN 2394-6385 (Print)
Volume 4, Issue 2, April – June 2017 ISSN 2394-6393 (Online)

Dynamic Web Site Adaptation by Applying Web usage Mining Techniques

J.Umarani[#1], G.Thangaraju[*2], S.S.Prasanna Venkatesan[*3]

[1]*Research Scholar, Department of Computer Science, Barathiyar University, Coimbatore*
[2] *Research Scholar, Department of Computer Science, Karpagam University, Coimbatore*
[3] *Research Scholar, Department of Computer Science, Thanthai Hans Roever College, Perambalur*

Abstract— In the Cyberspace, most people are tailored to obtain intelligence from the World Wide Web. To endure and flourish, a Web site has to constantly encourage its overall layout and potentials while providing a variety of constructive information services to attract users. The Web Recommendation System prompts users to visit a Web site and browse at a deeper level. Using large browsing patterns don't satisfy user's dynamic need. The requirement for an adaptive recommendation system comes into the modern world to encourage the online users to get required details immediately. This paper proposes a novel Web recommendation framework, based on Page Classification Algorithm which can respond to new navigation trends and dynamically adapts recommendations for users with suitable suggestions through hyperlinks. This research enables Web sites with dynamic intelligence to effectively tailor users' needs by means of personalization. User behavior is identified for the improvement of website design

Keywords— *Web Usage Mining; Site adaptation; Classification algorithm*

1. Introduction

To acquire information online users has to utilize navigation and search. Most Web users normally make use of a browser to navigate online information. They initiate with the index page or a Web page found through a search engine or connected from another Web site. They also pursue the hyperlinks that they think relevant in the home page and the subsequent pages, as for as they have originated the desired information in one or more pages. They may also use search amenities afforded on the Web site to speed up information searching. For a Web site consisting of a very large number of Web pages and hyperlinks between them, these methods are not sufficient for users to find the desired information effectively and efficiently. A significant source of deprived website design is that the web developers' insightful of how a website should be structured and can be noticeably diverse from individuals of the users. Such discrepancy result in users cannot simply find the desired information in a website. This issue is tricky to handle because when establishing a website, web builders may not have a clear understanding of users' choice and can only systematize pages based on their own ideas. However, the assessment of website usefulness relies on the fulfillment of the users rather than that of the developers. Thus WebPages should be coordinated in a way that commonly suits the user's replica of how pages should be prearranged. The problem of improving user navigation on a website with minimal changes to the current structure is an important issue.

1.1 Transformation Vs Personalization

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. 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. Transformation approaches make use of aggregate usage data from web log 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

2. Literature Survey

The methods proposed by Mobasher et al. [1], [2] 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. These methods are web personalization based. Nakagawa and Mobasher [3] develop a hybrid personalization system that can dynamically switch between recommendation models based on degree of connectivity and the user's position in the site. For reviews on web personalization approaches, see [4]. Lin [5] 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. Perkowitz and Etzioni [6] 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.

In [7] an overview of the web mining concept has been presented and how it can be useful and beneficial to the business improvement by facilitating its applications in various areas over the internet. The contribution of this paper is towards the various areas containing web sites on internet, which can make best use of different web mining techniques. Santosh kumar et. al. [8] concentrates on web usage mining and in particular focuses on discovering the web usage patterns of websites from the server log files. The comparison of memory usage and time usage is compared using Apriori algorithm and Frequent Pattern Growth algorithm.

3. Proposed Approach

The pages on the Web site are classified into two categories: index pages and content pages [9]. 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.

- *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.
- *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.
- *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[10].
- *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.

3.1 Algorithm for Page Classification

An algorithm for page classification is introduced in this section is used to determine the category of a page; its file type is first checked. If it is not HTML, the page is certainly a content page and no other testing will be necessary. Otherwise, its end-of-session count, number of links, and reference length, are examined subsequently.

Two thresholds, *count_threshold* and *link_threshold* are introduced. If a page's end-of-session count is greater than *count_threshold*, it is classified as a content page. If a page's number of links is greater than *link_threshold*, it is tagged as an index page. These thresholds should be selected conservatively so that they positively identify content or index pages. Finally, if necessary, the page's reference length is checked against the cut-off point *t*. If its reference length is less than *t*, it is marked as an index page; otherwise it is marked as a content page.

The algorithm for page classification is outlined as follows.

(1) l = 1/(mean reference length of all pages)
(2) t = -ln(1-_ ___l
(3) For each page p on the Web site
(4) If p's file type is not HTML 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

3.2 Site Reorganization

The goal of this phase is to reorganize the Web site such that its users will spend less time searching for the information they desire. The philosophy behind this is that a Web site will provide a better service to its users if it can cut down their navigation time by reorganizing the pa ges on the site. More specifically, we want to reorganize the pages so that users can access the information they desire with fewer clicks. Although other factors such as page layout affects navigation, the number of clicks a user has to

go through is the dominant factor for navigation since every click requires active rather passive effort from users and often involves a request to and a reply from the server.[11]

The general idea of reorganization is to cut down the number of intermediate index pages a user has to go through. To achieve this, we need to place the frequently accessed pages higher up in the Web site structure, i.e., closer to the home page, while pages that are accessed infrequently should be placed lower in the structure. In the meantime, we want to preserve the original site structure whenever possible, since it may bear business or organizational logics. Besides, dramatic changes of the site structure may confuse users. As a compromise between these two conflicting requirements, we introduce an evolutionary approach to Web site reorganization.

The basic idea is to locally adjust the site when a frequently accessed page should be promoted. In addition, two thresholds are introduced, that is, maximum number of links in an index page (*I*) and maximum number of links in a content page(*C*). An index/content page will not have more than *I/C*links after site reorganization, unless it has more links before reorganization, in which case its links will be intact. These two thresholds are introduced to achieve two objectives. First is to limit the number of links in a page so its layout will be reasonable. This will prevent extreme cases, for example, a flat site structure where all pages are linked from the home page. Second is to somehow contain the changes in the site structure. The selection of these thresholds can be done by the Webmaster or data analyst.

3.2.1 Formats in Site Reorganization

As mentioned earlier, in site reorganization, frequently accessed pages are put higher up in the site structure. On the contrary, infrequently accessed pages are placed lower in the site structure. In case such reorganization is not possible due to certain threshold such as maximum number of links in a page being exceeded, we will try to merge infrequent pages into a larger page.[12] The mergers will reduce the number of clicks by users due to fewer page requests, thus decrease navigation time. To prevent spurious results, the merging pages must be HTML files and at most one of them can be a content page.

To decide if a page is frequently accessed, a parameter, minimum frequency (*F*), is introduced. A page's frequency is defined as the number of sessions it is in divided by the total number of sessions. If a page's frequency is greater than *F*, it is called a frequent page otherwise, it is an infrequent page. In site reorganization, the pages are examined sequentially starting from the home page. For each page, we consider its immediate parents and children, where a parent is any page that has a link to it and a child is any page that it has a link to. Depending on the number of children it has, there are different formats and for each case, different actions may be taken according to the frequency and category of the pages involved. For each page, we consider three formats depending upon the number of children it has: 1, 2, and 3+. The three formats are illustrated as follows. For the sake of simplicity and also since the processing is done one parent at a time, only one parent is considered in the cases.

(I) Format I: The current page has one child. In this case, depending on the frequencies and categories of the pages, there are several possible outcomes, as shown in Figures 1, 2, and 3, where page 2 is the current page.

(a) Page 2 is an index page.

Apparently, Page B is redundant since it only serves as a link to page 3. Thus the most obvious solution will be to delete page2 and create a direct link from page1 to page3, as shown in Figure 1.

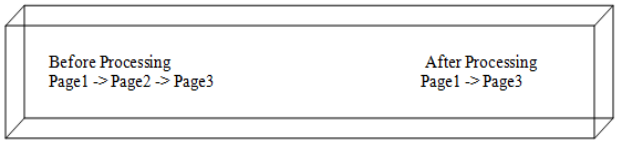

Fig.1: Format I, page 2 is an index page.

(b) Page B is a content page and page C is frequent.

Since page 3 is frequent, it should be promoted by adding a direct link from page 1 to it as shown in Figure 2. This assumes that page 1 has a free link, i.e., adding a link will not exceed its number of links limit. The maximum number of links in page 1 is determined by its category (*I* for index page and 3 for content page).

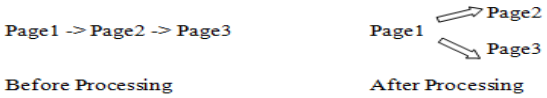

Fig.2: Format I, page 2 is a content page and page 1 has a free link

If page 1 has used its links to full capacity, but page 3 have a free link, it is sometimes worthwhile to demote page 2 to be a child of page 3 as shown in Figure 3. This is done if page 2 is used mostly to fetch page 3. This happens when the frequency of page 3 is more than half the frequency of page 2.

Fig.3: Format I page A has no free links, but 3 does

3.2.2 Algorithm for Site Reorganization

Based on the cases discussed in the previous section, the algorithm for site reorganization isoutline as follows.

- Initialize a queue *Q*
- Put children of the home page in *Q*
- Mark the home page
- While *Q* not empty

- *current_page* = pop(*Q*)
- Mark *current_page*
- For each parent *p* of *current_page*
- local adjustment according to the cases in Section 5.1
- Push children (maybe merged) of *current_page* into if they are not marked

4. Experiments

To evaluate the effectiveness of the site reorganization algorithm, we examine the number of pages, the number of links, as well as the average number of clicks in a session, before and after their organization. The parameters used in the site reorganization algorithm are listed in Table 1. The effects of various values of *F* are reported in this paper. The values of *I* and *C* are fixed to 30 and 10, respectively. Results from most other values of *I* and *C* show only minor differences. Since the total number of sessions is fixed at 77629, we use absolute values instead of percentages for *F* for illustrative purpose.

Parameter	Definition
F	minimum frequency.
I	maximum number of links in an index page.
C	maximum number of links in a content page.

Table 1: Parameters used in Site Organization

Figure 4 shows the total number of pages on the reorganized Web site for various values of *F*. The original Web site is shown as *F* = 0. It turned out that for the Web site we tested, pages were not deleted, only merged. When *F* = 0, pages cannot merge because every page is a frequent page. However, this is apparently not always true for other Web sites.

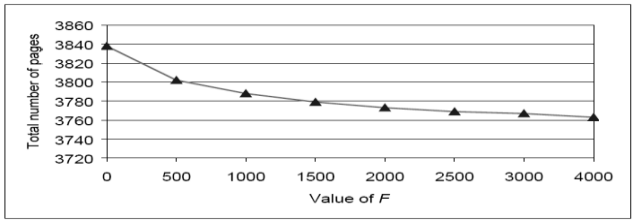

Fig.4: Effects of *F* on total number of pages

As shown in Figure 4, when *F* increases, the total number of pages on the reorganized site declines. This is because when *F* increases, more and more pages will be counted as infrequent and more pages will be merged into a single page. It should be pointed out that only 1,242 out of 3,838 pages on the Web site are accessed in the log files. The majority of the pages on the site are untouched. This probably explains why the reorganization does not cause drastic changes in the total number of pages.

5. Conclusion

A study on building adaptive Web sites is reported. An approach to reorganize Web sites based on user access patterns has been proposed. This approach aims to build Web sites that provide its users the information they want with less clicks. By analyzing the usage of a Web site and the structure of the Web site, modifications to the Web site structure are found to improve the structure of the Web site. In this approach, the Web site and its server log are first processed to acquire its structure and access information. The pages on the site are then classified into index or content pages based on access information. The Web site is finally examined to find better ways to organize the pages. Two algorithms, one for page classification and the other for site reorganization, have been developed. The proposed approach has been implemented and tested on a real data set from an actual Web site. The results demonstrate a high accuracy in page classification and a decrease in the number of clicks the user must perform to get interested information. Judging from the results obtained so far, it can be concluded that the approach is promising for adaptive Web sites.

References

[1] 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.

[2] 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.

[3] 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.

[4] M. Eirinaki and M. Vazirgiannis, "Web Mining for Web Personalization," ACM Trans. Internet Technology, vol. 3,no. 1,pp. 1-27, 2003.

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

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

[7] Pradnya Purandare, WEB MINING: A KEY TO IMPROVE BUSINESS ON WEB IADIS European Conf. Data Mining 2008.

[8] B.Santhosh Kumar,K.V.Rukmani, Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms Int. J. of Advanced Networking and Applications Volume:01, Issue:06, Pages: 400-404 (2010).

[9] A. Scime and L. Kerschberg, *Websifter: an Ontology-based Personalizable Search Agent for the Web*, Proc. Int' l Conf. on Digital Libraries, 439-336, Kyoto, Japan, 2000.

[10] Min Chen and young U.RYU "Facilitating effective user navigation through website structure improvement "IEEE,Volume 25,pp 571-588,2013.

[11] Mahajan, R., Sodhi, J., & Mahajan, V. (2014). Usage pattern discovery from a web log in an Indian elearning site. *Education and Information Technologies*.

[12] Min Chen and young U. Ryu "Facilitating Effective User Navigation through Website Structure Improvement" IEEE KDD Vol. 25. 2013.