Middle-East Journal of Scientific Research 24 (3): 628-631, 2016
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2016.24.03.23020

Enhanced User Personalization by Web Log Mining and Link Structure Display

[1]*P.S. Ambili and* [2]*Varghese Paul*

[1]Saintgits College of Engineering, Kottayam, Kerala, India
[2]TocH Institute of Science and Technology, Cochin, Kerala, India

Abstract: Web is a large repository of data and people often rely on web content for knowledge enhancement. Typically the information present lack predetermined structure, users find it difficult to extract required data efficiently from web pages with large number of hyperlinks between them. If such webpages are personalized that can facilitate users search information very fast and effectively. Mostly applications follow page/link ranking based on prestige[a] that do not ensure personalization. This paper proposes a fast and efficient data extraction approach for user personalization without major restructuring the site. We personalized a virtual learning site by mining information segmenting it and then made more user friendly by displaying the link structure. Display of hierarchical tree structure for user search facilitation and extraction of sequence for future traversal prediction is adopted. A comparative study on the actual user traversals and predicted user traversal is done. Conducted extensive tests on a real data set of student users and results indicate that the proposed scheme efficiently improves the user navigation with minimal structural changes. The proposed design is more apt for websites whose content remain stable overtime such as virtual learning environments and is also suited for artistic, health care and military applications.

Key words: Hyperlinks · Web log · User personalization · Navigation · Link prediction · Traversal · Virtual learning

INTRODUCTION

People rely on web for information gathering. Since the web keeps growing, it needs to locate sites and pages from the broad umbrella of topics while minimizing time on fetching and inspecting irrelevant pages. Data mining techniques can help user to extract data from large databases as well as to discover patterns in data. This can be extended to web data which is termed as webmining. Mainly three divisions can be made for web mining tasks: Web content mining- a content wise search for web pages, Web structure mining- Mining the structure of pages for getting information and Web usage mining- which looks at web access log to give information.

Usage data of a web site is a rich source for mining knowledge about the web site and its users [1]. Patterns well represented can be examined, represented in a structured way, can be used effectively for easy user navigation and future prediction of user. Mostly applications follow

- prestige – high standing achieved through success or influence.

page/link ranking based on prestige[a] that do not ensure personalization. The browsing patterns of individuals are not uniform. A structured representation of the browsing pattern based on user log can help understand the relationships between the pages they have visited. The visualization of this structure enables user for easy navigation. The key challenge is to create a personalized site with user navigation link structure visualization and without much restructuring.

In this work we propose a novel solution for personalization from user log data. A probabilistic classification method is employed for classification. Based on the classified data ranking of pages is done. Well formed HTML pages present tree structure with nodes as attributes and text. Clustering of link hierarchies are done based on the similarities in

Corresponding Author: P.S. Ambili, Saintgits College of Engineering, Kottayam, Kerala, India

link/navigation. Group of users with similar navigation pattern is identified from the clustered data [2]. We made an attempt to visualize user based link organization structure which automatically change by learning from web usage data which helps much in personalization.

Related Work: Through web usage mining, the server log, registration information and other relative information left by user access can be mined [3]. This access patterns from log can be effectively used for personalization. Major challenge is to do personalization without much restructuring of the site. Haibin Liu and Vlado Keselj [4] proposed an approach based on the combined mining of web server logs and the contents of the retrieved web pages. Displaying the web site structure is a good idea for users to identify their current locations relative to the web site as a whole and this can facilitate faster and hassle free user navigation. Through the visualized structures users can identify relationships between the web pages they have visited.User traversals on hyperlinks between web pages reveal conceptual relationships between these pages [5].

In the clustering algorithm called PageCluster, Jianhan Zhu *et al.* [1] proposed a method which clusters conceptually-related pages on each conceptual level of the link hierarchy based on their in-link and out-link similarities. Jalali *et al*. [6] presented a system for extracting user's navigational behavior using a graph partitioning model. Li, Yuxuan, *et al.* [7] proposed a method mining data and classifying the users probability of visiting particular page with sequential patterns from uncertain database. B. Mobasher *et al*. [8] pointed out that the discovery of patterns from usage data by itself is not sufficient for performing the personalization tasks.. In the Prefix Span algorithm which comes in sequential pattern matching [9] a pattern growth method is employed. For better personalization in addition to prefix, a user based scan is performed in the user span pattern algorithm proposed by Ambili P.S *et al.* [2]. In the current work we propose a novel procedure of traversing the link structure represented as graph/tree structure for better personalization. Similar traversals of different users are clustered to analyze user behavior. For classification probability based approach similar to Bayes classification is followed.

Proposed Scheme: The key aspects for user personalization this work proposes are:

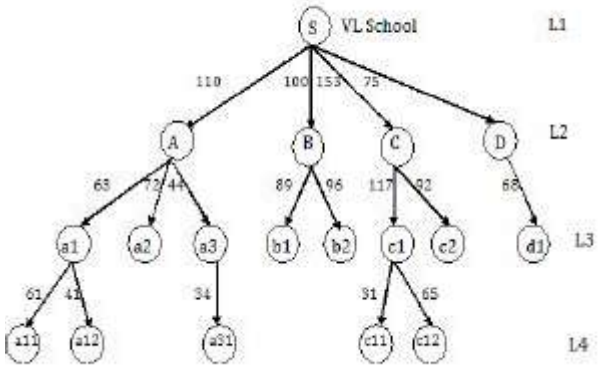

Fig. 1: Link hierarchy representation of VL School

Usage Classification: First classification of the data to select the top priority pages visited by user. For that first segmentation of the weblog data for user is done. From the selected segmented part a probability based classification is done and sorted based on total category count. The total subcategory count is also found and probability of user visit in the order of priority is set based on the obtained value of total/count.

Tree Structure: The link structure visualization of a website can facilitate fast traversal. Link hierarchy of navigation can be visualized through link structure. Here a link structure display is formed to facilitate personalization and fast traversal for our EUN system [2].

Since the webpages can be represented as a hierarchical structure a tree based representation is selected in the current work. Fig1 shows a tree based link structure representation to depict link hierarchy of a website, with our Virtual Learning System for user1. Started with home page at the first conceptual level L1, different categories are given as the Level 2 and subpages/next link hierarchies are given in the next level L3 and so on.

In the traditional BFS method for traversal, an order is maintained with initial start at left child then traverse to right and then to grand children. In our approach

Step1: At level 2 a sort procedure is employed and categories are put in sorted order. The category with highest count is dynamically assigned as the first tab which personalizes the system.

Step 2: Let S1 be the first item at level2. Select the left child of S1 and traverse all the grandchildren.

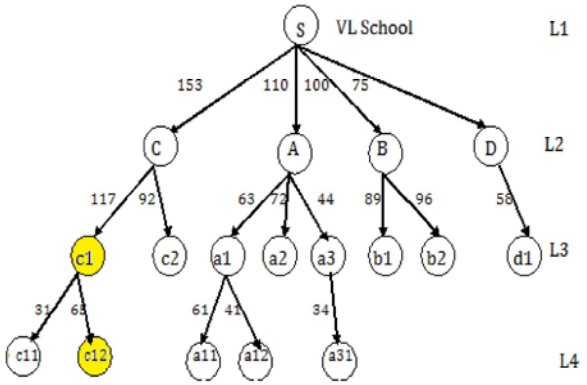

Fig. 2: Link hierarchy representation of VL School

Find the total count for each child.

Step 3: Repeat the step for right other grandchildren of S1

Step 4: The category with highest count is highlighted and this can be used for future prediction.

To facilitate user personalization link structure display and highlighting of previously visited page is done at login. This algorithm follows a sequential pattern matching approach with similar pages.

Fig2 shows the personalized structure of user1.Cc_1c_{12} is predicted as the future visit.

RESULTS AND DISCUSSION

Experimental evaluation is done with our virtual learning environment VLSchool. The model is developed using Java and Apache Tomcat as server. 230 genuine student user logins are tested with four courses. It is observed that the system provided perfect personalization and thus enhances easy navigation. The probably to be visited pages is listed in table1 with total number of attempts.

Predictions were found to be working satisfactorily with high success rate.

Comparison of Our Method and BFS Method: In BFS method all the nodes/links are assigned same priority. They are traversed level by level. Our method is a weight based method. From the list the node/link with highest weight is prioritized. Then traversal is continued from the highest weighed node. A comparison of traversal time for the mostly visited/predicted pages with BFS method and our method is shown in Table2. Table data confirms that faster traversal to the mostly visited page is achieved through our method.

Table 1: Future prediction of user visit

Prediction	Status	Attempts/Total	Rate
Page Aa_1a_{11}	Success	97/120	.81
	Failure	23/120	.19
Page Aa_1a_{12}	Success	56/100	.56
	Failure	44/100	.44
Page Cc_1c_{11}	Success	115/140	.82
	Failure	25/140	.18
Page Cc_1c_{12}	Success	137/170	.81
	Failure	33/170	.19

Table 2: Comparison of our method and BFS

No of Nodes/level	45/7	58/7	90/11
Traversal time in seconds for mostly visited page with BFS	54	60	125
Traversal time in seconds for mostly visited page with our system	38	40	71

CONCLUSION

This work proposes a novel approach for better user personalization through link structure representation. For ranking of pages, the count of number of visited pages are mined from web log data rather than taking the number of logins. This avoids the possibility error due false logins or crawlers. The factor of user behavior changes over time is also a reason for selecting weblog mining rather than web structure mining. In addition to personalization a future prediction is done in our system and the success rate of predicted page visits are found to be high. For clustering and grouping of similar category visited users a pattern analysis based on user span pattern is adopted. Results obtained with 230 student user logins show that better user personalization is achieved and traversal time is lesser than compared to BFS method. The proposed system is better suited for virtual learning environments whose content has a consistency over time.This system is also suitable for artistic, health care and military applications pages. The prediction system has future scope in the areas of natural calamities prediction, E-commerce and market watch.

REFERENCES

1. Jianhan Zhu, Jun Hong and John G. Hughes, 2004 “Page Cluster: Mining Conceptual Link Heirarchies From Weblog Files for Adaptive Website Navigation” ACM Transactions on Internet Technology special issue on “Machine Learning for the Internet” 4(2): 185-208.

2. Ambili, P.S. and Dr. Varghese paul, 2016. “User Span Pattern: A Sequential Pattern Mining Approach for Personalization”, International Journal of Applied Engineering Research ISSN 0973-4562, 11(1): 621-624.
3. Dong, D., 2009. “Exploration on Web Usage Mining and its Application,” in IEEE.
4. Haibin Liu and Vlado Kes˘elj, 2007. “Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users’ future requests”, Data and Knowledge Engineering 61, Elsevier publication, pp: 304-330.
5. Kumar, P.R. and A.K. Singh, 2019. “Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval”, American Journal of Applied Sciences, 7(6): 840-845.
6. Jalali, M., *et al.,* 2008. “A new clustering approach based on graph partitioning for navigation patterns mining,” in International Conference on Pattern Recognition, pp: 1-4.
7. Li, Yuxuan, *et al.,* 2013. "Mining Probabilistic Frequent Spatio-Temporal Sequential Patterns with Gap Constraints from Uncertain Databases." Data Mining (ICDM), 2013 IEEE 13th International Conference on. IEEE.
8. Bamshad Mobasher, Honghua Dai, Tao Luo and Miki Nakagawa, 2002. “Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization” 2002-Springer Data Mining and Knowledge Discovery ISSN: 1384-5810 (Print) 1573-756X (Online) January, 6(1): 61-82.
9. Chandra Shekhar Rao, V. and P. Sammulal, 2013. "Survey On Sequential Pattern Mining Algorithms". International Journal of computer application (0975-8887), 76: 12.