European Journal of Applied Sciences 8 (1): 47-50, 2016
ISSN 2079-2077

DOI: 10.5829/idosi.ejas.2016.8.1.22856

Better User Personalization: A Sequential Pattern Matching Approach

[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: The richness of web content has also made it progressively more difficult to leverage the value of information. Identifying users' topic of interest, recommending content to a user based on past behavior without major restructuring of the site is a major challenge. Mining knowledge about the usage of a website can be used effectively for user personalization and that can facilitate search information very fast and efficiently. This paper proposes a novel approach to facilitate user navigation without restructuring the site by mining knowledge and by a probabilistic classification. Based on the cluster information a new approach for on ranking the web pages resulting in users' possible link prediction is done. Segmentation of the log file of groups of users having similar navigation and similar pattern over time is studied. For pattern matching sequential patterns spanning over sessions is selected. In the Prefix Span algorithm which comes in sequential pattern matching a pattern growth method is employed. For better personalization in addition to prefix, a user based scan is performed in our new USP (user span pattern) algorithm. Results from extensive tests conducted on a real data set indicate that our model effectively improves the user navigation with minimal changes. The proposed model is more suitable for websites whose content remain stable overtime such as educational sites and is also suited for artistic, medical and military applications.

Key words: Classification · Segmentation · Web log · Navigation · User personalization · Link prediction · Pattern matching

INTRODUCTION

Web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services. This may be the data actually present in the web pages or data related to web activity. Web mining tasks can be divided mainly in to three classes: Web content mining, Web structure mining and Web usage mining. The first is traditional searching of the web pages via content, while in the second information is obtained from the actual organization of pages. The intra page structure includes links within the page, anchor texts as well as the code (HTML, XML and XHTML) for the page. Web usage mining looks at logs of web access.

Web is a collection of multiple billions of documents written in a way that enables a user to navigate using hyperlinks. Usage data of a web site is a rich source for mining knowledge about the web site and its users. Access pattern tracking is a specific type of usage mining that looks at the weblog data. Information needed by users from website may not be the same. Even the same user may require different pages at different times, or may have some regular browsing patterns. A personalized website according to individual user needs could be a fast effective solution for such browsing patterns. By classifying the access behavior of a particular user and by showing user traversals on hyperlinks websites can be personalized. A major challenge is to create a personalized site without much restructuring.

In this paper we propose a new probabilistic classification method based on Bayes classification for finding effectively to personalization of site based on the user log classifications. Ranking of pages is done based on the classified data. Group of users with similar navigation pattern is identifies from the clustered data. Ranking of pages and a particular user's future visit is also predicted.

Related Work: User traversals on hyperlinks between web pages reveal conceptual relationships between these pages. Comparing with the traditional Breadth First Search and Shortest Path method their method show more accurate results. But adding usage mining segmentations

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

and classification can better the results. B.Mobasher [1] presents a Web Personalizer system which provides dynamic recommendations, as a list of hypertext links, to users. Sessions can be another source for finding access patterns of user.Pattern discovery activities form major portion of the mining activities as look to find hidden patterns within log data. Haibin Liu and Vlado Keselj [2] proposed an automatic classification of web user navigation patterns. Sarukkai's method [Sarukkai 2000] for predicting most probably to-be-visited pages was improved by Jianhan Zhu [3] by the page cluster algorithm. Li, Yuxuan, *et al* [4, 5] proposed a method mining data and classifying the users probability of visiting particular page with sequential patterns from uncertain database. In the Prefix Span algorithm which comes in sequential pattern matching [6, 7] a pattern growth method is employed. For better personalization in addition to prefix, a user based scan is performed in our new USP (user span pattern) algorithm.

Proposed Scheme: This paper focuses on mainly three activities for user personalization of web pages.

Usage Classification: First is to select the top categories mostly visited by user. For this a new classification based on Bayesian Classifications is done. Bayesian Classifier is statistical classifier. It is based on Bayes Theorem. It has high speed, accuracy and has a comparable performance. Let A be a sample of data whose class label is unknown. Let H be a hypothesis that A belongs to class C. The classification is to determine P(H/A) ie, to prove A comes under or satisfies the hypothesis H.P(H) is the prior probability. P(A) is the probability that sample data is observed.

Bayes theorem states that
P(H/A)=[P(A/H) P(H)] / P(A)

ie, posterior= [likelihood*prior]/evidence. But here if we take the total login count of a user as the likelihood only log in and not visiting pages or crawler possibilities may result in not getting exact personalization. In order to avoid this category count is taken. Algorithm of the analysis done with EUN, our prototype.

Algorithm:
Step 1: Find the total category count for a particular category _id from the user category state table

Step 2: Find the count of each subcategory from the user sub category table.

Step 3: Joined the user category _id and category id from web category table.

Step 4: Probability of selecting a category will be count/total

Step 5: The probability and subcategoryid is stored in webcategory table.

Data analysis done for a sample educational website.It is observed that the category that is visited most will dynamically assigned as the first link to facilitate navigation.The category that is visited least will be the last link.When the user login not only the most visited category but also the subcategory and contents are loaded.

Link Structure: Secondly to facilitate personalization and fast traversal a link structure is produced. Tree based link structure is displayed.BFS is adopted to display the hierarchy.

Future Prediction: Third to facilitate where to visit next, prediction based on past visits are done. For this users' traversal patterns without ordering in sessions can be selected. Number of times a user has visited the same category is predicted by segmentation of the log data first and then by clustering. For pattern matching sequential patterns spanning over sessions is selected. In the Prefix Span algorithm which comes in sequential pattern matching a pattern growth method is employed. For better personalization in addition to prefix, a user based scan is performed in our new USP(user span pattern) algorithm. We had first sorted the data and then a threshold based approach rather than time based one is adopted.

Algorithm:
A sort pattern is employed first to put the data in the sorted order.

Input S={u1,u2,u3,…un} //set of users

From category list {A,B,C..}, a,b,c…⊂ A ; a,b,c…⊂ B and so on.

Similarly aa,ab,ac……⊂ Aa ; ba,bb,bc ……⊂ Ab and so on
Output S= sort users based on category and then sub categories most visited.

A maximal reference sequence selected.

Set that as particular users future visit category.

Table 1: Sequence database for user1

User	Category	Sequence database	Projected count
u1	A	{Aab,Aababc,Abbccda,	{(A=4){{{(a=7),(b=5),(c=5),
	B	Aaaccda,Bac,Baabc,Babac,}	(d=1)}{(aa=2),(ab=4)…}}}};
			{(B=3){{(a=5),(b=2),(c=3)}
			{((aa=1),…}}}

From the projected count users' next visit can be set as {Aab}

Table 2: Efficiency of EUN based on time and number of mouse clicks

Task	With EUN system	Search/Success	Time Sec	No of mouse clicks
Page A	Yes	1/1	32	2
	No	1/1	35	2
Page Aa	Yes	3/3	42	3
	No	3/3	54	5
Page Ab	Yes	3/3	44	4
	No	3/3	54	6
Page Ac	Yes	3/3	44	4
	No	3/3	54	6
Page Aab	Yes	5/5	55	4
	No	5/4	65	9

Table 3: Future prediction of user visit

Prediction	Status	Attemps/Total	Rate
Page Aa	Success	5/7	0.71
	Failure	2/7	0.29
Page Ab	Success	5/7	0.71
	Failure	2/7	0.29
Page Aab	Success	5/8	0.625
	Failure	3/8	0.375
Page Ba	Success	6/7	0.86
	Failure	1/7	0.14
Page Baa	Success	5/7	0.71
	Failure	2/7	0.29

Steps:

- For each user find the data sequence and subsequence
- For each user for a particular Category from list {A,B,C..}, the subsequence with greater count is selected; then in that selected sequence iterations are done till the sequence with greater count is found; set as future visit
- If same category count category in order is set.

RESULTS AND DISCUSSION

We developed a prototype called EUN (Easy User Navigation system) and evaluated the results. The model is developed using Java and Apache Tomcat as server. An educational site is evaluated and task is taken as selection of a particular tutorial and its subpages. The below table shows how the user is facilitated easy navigation with the help of link prediction.

Using User Span Algorithm future prediction of user visit is done. Obtained results and Success rate evaluation is given below in Table 3.

Most probably to be visited page for the next attempt is predicted in table. It is observed that rate of success is found to be very high and failure rate is negligibly less. Time reduction with link view and without link view is studied and found that the system is efficient.

CONCLUSION

This work proposes a novel approach for user personalization. Number of times a user has visited the same category is predicted by segmentation of the log data first and then by clustering. Web structure mining is usually employed for personalization, but since user behavior may not be similar over time, here weblog data mining is done. Pattern analysis of subcategory also done with the help of a novel user pattern span algorithm. Success rate of user visit of the predicted page is also found to be high. Results with the EUN prototype show that user personalization is achieved. This system is most suitable for user personalization of websites whose content remain stable overtime such as educational sites and is also suited for artistic, medical and military applications. The prediction system can be extended for applications such as market watch, natural calamities prediction and speech recognition in future.

REFERENCES

1. Bamshad Mobasher, Robert Cooley and Jaideep Srivastava, 2000. Automatic personalization based on web usage mining, Communication of ACM, 43(8): 142-151.
2. Haibin Liu and Vlado Keselj, 2007. Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests, Data & Knowledge Engineering, Elsevier Publication, 61: 304-330.

3. Zhu Jianhan, 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.
4. Kumar, P.R. and A.K. Singh, 2010. Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval, American Journal of Applied Sciences, 7(6): 840-845.
5. Li, Yuxuan, 2013. Mining Probabilistic Frequent Spatio-Temporal Sequential Patterns with Gap Constraints from Uncertain Databases, Data Mining (ICDM), 2013. IEEE 13th International Conference on, IEEE.
6. Oren Etzioni, 1996. The world wide Web: Quagmire or gold mine, Communications of the ACM, 39(11): 65-68.
7. Rao V. Chandra Shekhar and P. Sammulal, 2013. Survey On Sequential Pattern Mining Algorithms, International Journal of Computer Application, (0975-8887): 76-12.