Journal of Theoretical and Applied Information Technology
20th September 2014. Vol. 67 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PERSONALIZING USER DIRECTORIES THROUGH NAVIGATIONAL BEHAVIOR OF INTERESTING GROUPS AND ACHIEVING MINING TASKS

[1]MS.R.KOUSALYA [2] DR. V. SARAVANAN
[1] Ph.D, Research Scholar, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India
Head/Assistant Professor, Department of Computer Applications,
Dr. N.G.P Arts and Science College, Coimbatore -641 048, Tamil Nadu, India
[2] Professor & Director, Department of Computer Applications,
Sri Venkateswara College of Computer Applications and Management, Coimbatore-641 105, Tamil Nadu, India

ABSTRACT

Web directory is a collection of web pages with links. A grouping or organizing web content into thematic hierarchies are known as Web Directories which are correspond to topics listed are easily visualize by the people. The key focus on our methodology is that is personalizing web directories using navigation patterns prepared for user profiles. A user profiles containing the information about a user or group of user's previous visits categories. Patterns are generated based on navigational behavior and interest of the individual user or a group of user. Our work provides a set of mining tasks and personalization techniques to customize the organization of user directory based on corresponding pattern behavior.

Keywords: *Personalization, User Directory, User Profile, Navigation Patterns, Clustering*

1. INTRODUCTION

Personalization is a technique mining navigational patterns using similarity measures. Personalization based on the interest in an individual, group or organization. In context in previous work the Web directory is considered a thematic hierarchy and personalization result based on total data used. In the most of the work on Web usage mining, the usage data that is analyzed as a user navigation pattern, to a certain extent than a particular Web site, a high amount of thematic diversity is a result. A user can browse a particular directory to their web search that relevant to different topics. A directory known as user directories is a structure of thematic categories in which web resources are classified [2]. A user directory can help users to understand how topics are related and assist key word search by narrowing it within the particular category. User directories can be categories the topics of a general interest of the user. For example, the Open Directory Project (ODP) is a popular directory of general interest used by Google web Directory, for e.g. (http://directory.google.com).

The tremendous growth of web data sources, different user communities accessing them, cause the need for personalization the web content and web services. To face this issue, there is a need for a content personalization for future digital libraries [3]. In this work, we propose a methodology for personalizing user/topic directories. Personalization based on navigational patterns collected through users previous web search behavior collected from proxy servers. Since a user directory contains heterogeneous sets of web resources, the proposed methodology identifies groups of users with similar interest, named interest groups, and our aims personalization and mining tasks to each group individually. Frequent visit of a particular page by particular users or a group of users stored in a system as a navigation pattern through back and forth visits to the same categories of the directory by the user. Due to proxy servers and cached versions of the pages frequent visit by the user using 'Back' and 'Forth' option, the sessions are identified have many missed pages. Discover sequences of a set of categories to confine the inherent hierarchical organization of the categories in user directories. Thus, after personalized a process, each group of users is identified with a different view of the same user directory.

Key Contribution

1) To find a similarity metrics and methodology for discovering interest groups. Assign a user to an interest group if their navigation patterns during the browsing are similar to that of the other user groups. For this we propose a metric to compare the similarity between two

navigation patterns and we generate interest groups using clustering techniques.

2) Prepare a set of mining tasks for interest groups. These tasks are the detection of indecisive user behavior, the discovery of sequences of famous categories, and the discovery of sequences of categories in the directory.

3) Prepare a set of personalization task. (Through static offline/online shortcuts)

4) Experimental evaluation of the abovementioned methodology. We carry out several tasks to evaluate mining and personalization tasks. Our results exhibit the effectiveness of our approach.

And also we mention here all mining and personalization tasks have been implemented in working model maintaining the topic directory of ODP (Google directory).

Outline: The following section carries related work. In Section 2, discuss about related work. In section 3, we introduce the concepts regarding user directories and navigational patterns. The proposed mining and personalization tasks for user directories are presented in section 4 and 5. In section 6, we present the experimental result and evaluation of our methods. We conclude in section7 with our future scope of research.

2. RELATED WORK

In this, we introduce some related work in data mining algorithms and probabilistic models. In the recent years, there has been much research on Web usage mining. The related literature is very extensive and many of these approaches fall out of the scope of this paper. [1, 10, 12]. There are many approaches for discovering sequences of visits in a web site. Some of them are based on data mining techniques [2,18,22], whereas others use probabilistic models, such as Markov models to represent model the user's visit [5,8,11,25,27]. Such approaches aim at identifying browsing patterns describes the activity and help to customize the website. They do not propose any methods or idea for personalization of websites.

The existing approaches mentioned to personalize a website [3,13,15] does not distinguish between different users or user groups in order to perform the personalization. So the methods appear to be more appropriate to ours, in terms of identifying different interest groups and personalize the website on these profiles based on collaborative filtering.

Mostly Collaborating filtering systems [4, 5, 16, 19] is used to develop recommendations of e-commerce solutions based on the assumption that user with common interests and behavior during their web search. Thus the identifications of similar user profile enable the filtering of related content and produce recommendations based on the same. Similar to that, we identify users with similar interests and use this to personalize the user directory. In our work we do not prepare user profiles as vectors to find similar users. Instead we use clustering techniques to group users into interest groups to prepare certain user directories. Moreover, we propose the use of sequential pattern mining to generate recommendations. We also find sequential dependencies with in user's visit, is not the case with collaborative filtering systems.

All of the above mentioned approaches is to aim at personalizing generic web sites , but our approach focuses on the personalization of a more specific website , that of user directories. User directories organize the web content in to meaningful categories.

As mentioned earlier, our methodology does not limit its personalization services of individual users alone. To a certain extent, we identify a group of users with similar interest and behavior through their visits to particular categories and information portals. This is enabled by the collaborative filtering approaches [6,9] even though; those approaches fail to find the sequential dependencies between consecutive visits by the user.

3. SUB CATEGORIES AND RELATED CATEGORIES

A user/topic directory is a hierarchical organization of thematic categories contains information about a web links. A category may have sub categories that narrow down the content of parent categories and related categories contain similar resources, which may exist in the different part of the directory.

Navigation patterns

A concept of Navigation patterns is very much used to represent the navigational behavior of the user when he browses the directory. A navigation pattern is formed by the sequence of categories visited by a user during a session. Multiple occurrences of the same categories are also included in those pattern will result of users going back and forth within a root directory path that indicate the user may have interest in more than one topic .

4. MINING TASKS

To personalize a user directory, we propose a data mining tasks over and above user's navigation patterns. To this the system need to understand users with similar navigation behavior and similar search interests. Such users frame interest groups. Mining tasks are performed for the same.

Interest group

The main part of the proposed work involved in recognizing how two navigation patterns are similar to each other . There we use a metric to measure the similarity between two navigation patterns. The metric is defined as the ratio of the number of the common categories that exist in the navigation patterns to the total number of distinct categories in the patterns.

Let P1 and P2 be the two navigation patterns. The similarity S between two navigation patterns (p1 and p2) is calculated as follows

$$S = \frac{\sum_{forallp} occuranceofp \in p1 \cap p2inp1 \cup p2}{|p1|+|p2|}$$

The similarity between the pattern { tamil, english, maths, science, commerce, science, commerce } and pattern {tamil, english, maths, computer science, commerce} is S = 9/12 = 0.75. We used the Hybrid clustering method of combinations of the k-means clustering algorithm [14] and hierarchical clustering algorithm, both combined to find the similarity metrics. It used to detect clusters of navigation patterns and to analyze navigational patterns is called agglomerative for the same clusters because it merges similar patterns in a cluster iteratively. It is helpful for the system produce the user interest groups. An interest group is created by the users whose navigation patterns are presence in the same cluster. Clustering of patterns contains users profiles used to prepare user community directories can help the user to access preferable websites according to their need.

Indecisive Users

Mining tasks import to identification of indecisive users. Users called as indecisive when their navigation behavior shows their back and forth visit to same WebPages in the directory. This is because of the user might not know the actual information what he/she wants to search. Apart from this the organization of categories in the directories are different from user interest results poor organization of topic sub directories. The navigation pattern {apple, grapes, orange, apple, orange, chicko, apple, guava, orange} indicate user as indecisive user, since he/she goes back and forth to visiting categories apple, orange. To detect indecisive users, B&F actions (Back and Forth) are allowed in the navigation patterns of users. A set of categories {N1, N2,.... Nk-1, Nk, Nk-1, …N2, N1,…, Nk } is a sub pattern of P. We denote P' a B&F chain in P. The following method illustrates how a k-B&F action is detected (where k is the length) within the navigation patterns. For example P is a navigation pattern of the form {A,B,C,D,C,B,A,B,C,D,E,F}, and 4-B&F action detected as follows;

```
Step1:function CheckBF(navigation pattern P,
int     k, init position i)
Step 2: startPosL= i
Step 3: endPosL= startPosL +(k-1)
Step 4: startPosR = i + (k-1)+(k-2)+1
Step5: endPosR= startPosR + (k-1)
Step6: if (Subpat(P, startposL, endposL) ==
Subpat( P,startPosR, endPosR)
Step7: and
Step8: ( Subpat( P, startPosL+1, endPosL-1) = =
        subpat ( p, startPosR-1, endPosL+ 1))
then
Step9: return True
Step10: else
Step11: return False
Step12: end if
Step13: end function
      function subpat (navigation pattern P,
         startPos,        endPos)
      return part of P from startPos to endPos
      end function
```

Fig 1: Detection Of B&F Actions

The above approach used to detect B& F actions, and calculate the frequency of occurrences and rank the users in descending order of this frequency. The higher rank in the frequency the higher the degree of user indecisiveness.

Popular Categories and sequential navigational subpatterns

The categories of a topic directory are hierarchically arranged. This accepted order should be established in the discovered navigation patterns. So we need to focuses on discovering frequent sequences of popular categories in navigation patterns. A category is popular because its number of visits is highly ranked. These sequences are called as sequential navigation subpatterns. These subpattern capture the idea of popular moves among categories. The order for a navigation pattern tells apart the discovery of frequent sequential navigation patterns of the discovery of association rules [17]. To identify frequent sequential navigation subpattern a k-apriori algorithm is used to mine frequent item sequences. The length of a sequential navigation subpattern is defined as the number of categories in the subpattern. If a sequential navigations subpattern length k means that k-sequential navigation subpattern. For example, { apple, orange, grapes} is a 3-Sequential navigation subpattern.

The support σ of a k-sequential navigation sub pattern defined as the fraction of navigation patterns that contain no of subpattern in an interest group. The support of this subpattern is calculated using a popular category X and the 1-sequential navigation pattern {X} have the same support σ (X) is equal to a threshold is called as minimum support.

Let P is the set of navigation patterns of an interest group. Give a set S = {S1, S2, Sn} of k-sequential navigation subpatterns of P, the support of each Si is defined as follows:

$$\sigma(S_i) = \frac{|\{nav.patternp \in p : S_i issubseqofp\}|}{|P|}$$

For a minimum support $\sigma_{\min}$ and a length k, we recognize the frequent k- sequential navigation subpatterns whose support is more than $\sigma_{\min}$. For above mention mining tasks achieved through k-apriori method are a combination of k-means clusterings and apriori association.

5. PERSONALIZATION

Mining tasks is more useful in personalization of the topic directory in terms of navigation behavior of the user and their interest groups. The result of personalization tasks are a set of shortcuts among categories in the directory. A shortcut A → B is a direct link from A to B. These shortcuts are the used to navigating the directory based on navigation behavior of different users or interest group. We used two modes of personalization to create a shortcut is online mode and offline mode.

Offline mode

In the offline mode the system analyze the navigation patterns of interest groups and recommends the shortcuts. Then the admin decides which of the shortcut optimum is for the group is created at the last. This type of shortcuts is static shortcuts are given to all members of an interest group.

We represent two personalization tasks to create static shortcuts based on detecting a) frequent B &F chains, b) frequent sequential navigation sub patterns.

Personalization based on frequent B&F search

As mentioned earlier, the occurrence of B&F categories in a user navigation pattern point out that the user in indecisive nature. The mining task of identifying frequent B&F chains helps to determine such part in the directory. The proposed personalization tasks (shortcuts) are performed for each interest group separately. Our proposed method recommends a shortcut in the directory for each frequent B & F chain presence in the navigation pattern. The category presence in the beginning of a frequent B&F chain is the stat point of the shortcut and the category presence in the last of frequent B & F is the end point of the shortcut.

For example, consider the following navigation pattern {movie, dance, movie, dance, drama}. In that {movie, dance, movie, dance} is a frequent B&F chain. Our proposed method recommends the shortcut as movie →drama in the directory.

Personalization based on frequent sequential navigation subpatterns

By recognizing the popular categories for an interest group, the interested topics of users groups are identified. The sequences of popular categories indicate the popular move between these topics. To personalize topic directories, the system recommends a static shortcut. These shortcuts recommend the user to move directly to these popular categories.

For a given interest group and threshold, the system identifies the frequent 2-sequential navigation subpattern {a, b}. The system recommends a static shortcut a→b. Here we need not want to consider frequent k-sequential navigation subpatterns, with k>3, (k =length), k-Apriori algorithm put up these subpatterns based on a set of extracted frequent 2-sequential navigation subpatterns. The following frequent sequential navigation subpatterns for a given

interest group are {arts, science} and {science, maths}. The system identifies arts → science and science → maths is the candidate shortcuts.
Our system recommends only arts → science and there is an edge of science to maths is already available.

Online mode
In online modes the system considers the categories visited by the user and the navigation patterns of their interest groups. Shortcuts are created in real time are known as dynamic shortcuts and it does not involve any actions from the directory administrator. Different dynamic shortcuts are provided to each individual user. Our aim to prepare a personalization task for dynamic environment using sequential navigation subpatterns.
We use a sliding window (w) is called active navigation window for each user interest group. All fixed size window slides over the user navigation in the active session, have the last |w| popular categories visited by the user. To personalize the directory, the system creates shortcuts by matching each slide window w with the sequential navigation sub patterns of users' interest groups, respectively. A same method introduced in [20]. Our approach from multiple windows instead of single window and matches each of them against the sequential navigation subpatterns of the corresponding interest group. In this approaches, we extract and store in advance all sequential navigation sub patterns having the length |w|+1and a given a support threshold with fixed window size |w|. The system dynamically creates shortcuts for all users whose confidence value is equal or greater than a given minimum confidence threshold. Then we describe the confidence for dynamic shortcuts.
Let A → B is a shortcut w an active navigation window for an interest group such that A is the last category of w. The confidence α of A → B is defined as:

$$\alpha(A \to B) = \frac{\sigma\left(\omega o \ \{B\}\right)}{\sigma\left(\omega\right)}$$

Where o denotes the concatenation operator.

The confidence of a shortcut A → B, with A being the last category of the active window w, refers to conditional probability that the user will visit B given that she has visited all the categories of w.

6. EVALUATION

We have implemented a model system to evaluate the mining and personalization task proposed in this paper. Developing a web application is the first module where users are allowed to register and browse the ODP (Open Directory Project). The ODP categories and the categories of web pages were stored in a RDBMS. The Second module is a stand- alone applications are developed to execute the mining tasks and then perform personalization tasks.
We present the results of evaluation of interest groups. We allow 12 users (Two groups of each 6) to register in our system and browse the directory relevant to topics like sports, news, cooking, and mother Theresa. We assign a topic to 3users first then we repeat the process by switching the topic between groups. We identified different clusters of navigation patterns. We use hybrid clustering methods (combination of both K-means and hierarchical clustering) techniques to find meaningful clusters of navigation pattern are formed of 10 clusters. We calculated precision Pr and recall R values for the 10 clusters by micro average method. High precision implies high accuracy of the clustering task, while low recall means that there are many patterns not assigned to the correct cluster. High precision and high recall indicate excellent clustering quality.
The calculated Pr and R values are shown in Table1.
All navigation patterns wrongly assigned to cluster 2 and 10 were as misclustered patterns. High value indicates the effectiveness of the interest group.

Table 1: Precision And Recall Values For Interest

Cluster No	Precision (Pr)	Recall (R)
1	1.0	1.0
2	0.00	----
3	1.00	0.57
4	0.79	0.85
5	0.82	0.93
6	1.00	0.92
7	1.00	0.93
8	0.98	1.0
9	1.00	0.97
10	0.00	---

Groups Detected

Table 2: Hit Ratio Of Shortcuts Per Interest Group

Interest group	Hit Ratio
1	1.10
2	0.54
3	---
4	0.80
5	1.20
6	0.78
7	0.58
8	0.92
9	0.53
10	---
Avg for the 8 valid group	0.72
Number of shortcut inserted	26

To evaluate the personalization tasks in the offline mode, we calculated the hit ratio of the created shortcuts per interest groups. The high rates of hit ratios demonstrate the high utilization of shortcuts, and, in consequence, the effectiveness of personalization tasks is achieved. We created a synthetic data set including to visits to a part of ODP for online mode. Each navigation pattern of the test data set is divided in two parts. The first part is used for generating shortcuts using our methodology. The second part is used for evaluation. We calculate the ratio of numbers of popular categories involved in shortcuts as targets, to the number of created shortcuts for each navigation pattern of the test data set. By averaging the computed values the precision of the personalization task is calculated.

7. CONCLUSION

In this paper, we bring in a methodology for personalizing user directories along with the navigation behavior of the users. We introduce a set of mining task on navigation patterns which are captured from the user navigation behavior during their web browsing. User interest represented in terms of visited categories and retrieved resources. Furthermore, we recommend a set of personalization tasks that customize the organization of the topic directory as interest group. The creation of links called shortcuts between categories in the directory. At last we run many experiments to evaluate the proposed mining and personalization tasks, bring out the effectiveness of our approach. Our future work will focus on semantic information available in web directories and to extend for online personalization by studying real user navigations.

REFERENCES:

[1] B. Mobasher, R. Cooley, and J. Srivastava, “Automatic Personalization Based on Web Usage Mining,” Comm. ACM, vol. 43, no. 8, pp. 142-151, 2000.

[2] J. Srivastava, R. Cooley, M. Deshpande, and P.T. Tan, “Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data,” SIGKDD Explorations, vol. 1, no. 2, pp. 12-23, 2000.

[3] D. Pierrakos, G. Paliouras, C. Papatheodorou, and C.D. Spyropoulos, “Web Usage Mining as a Tool for Personalization: A Survey,” User Modeling and User-Adapted Interaction, vol. 13, no. 4, pp. 311-372, 2003.

[4] G. Paliouras, C. Papatheodorou, V. Karkaletsis, and C.D. Spyropoulos, “Discovering User Communities on the Internet Using Unsupervised Machine Learning Techniques,” Interacting with Computers J., vol. 14, no. 6, pp. 761-791, 2002.

[5] G. Xu, Y. Zhang, and Y. Xun, “Modeling User Behaviour for Web Recommendation Using lda Model,” Proc. IEEE/WIC/ACM Int’l Conf. Web Intelligence and Intelligent Agent Technology, pp. 529-532, 2008.

[6] W. Chu and S.-T.P. Park, “ Personalized Recommendation on Dynamic Content Using Predictive Bilinear Models,” Proc. 18th Int’l Conf. World Wide Web (WWW), pp. 691-700, 2009.

[7] The Adaptive Web, Methods and Strategies of Web Personalization, P. Brusilovsky, A. Kobsa, and W. Neijdl, eds. Springer, 2007.

[8] D. Pierrakos, G. Paliouras, C. Papatheodorou, V. Karkaletsis, and M. Dikaiakos, “Web Community Directories: A New Approach to Web Personalization,” Web Mining: From Web to Semantic Web, B. Berendt et al., eds., pp. 113-129, Springer, 2004.

[9] D. Pierrakos and G. Paliouras, “Exploiting Probabilistic Latent Information for the Construction of Community Web Directories,” Proc. 10th Int’l Conf. User Modeling, L. Ardissono, P. Brna, and A. Mitrovic, eds., pp. 89-98, 2005.

[10] C. Christophi, D. Zeinalipour-Yazti, M.D. Dikaiakos, and G. Paliouras, “Automatically Annotating the ODP Web Taxonomy,” Proc. 11th Panhellenic Conf. Informatics (PCI ’07), 2007.

[11] P.I. Hofgesang, "Online Mining of Web Usage Data: An Overview," Web Mining Applications in E-Commerce and E-Services, pp. 1-24, Springer, 2009.

[12]G.Castellano,A.M.Fanelli, and M.A. Torsello, "Computational Intelligence Techniques for Web Personalization," Web Intelligence and Agent Systems, vol. 6, no. 3, pp. 253-272, 2008.

[13] . D.R. Ferreira, C. Alves " Discovering User Communities in Large Event Logs" citeseerx

[14] J. Borges and M. Levena, Data Mining of user navigation patterns , Lecture Notes in computer Science, 1836 (1999).

[15] J. Callan, A. Smeaton, M. Beaulieu, P. Borlund,P. Brusilovsky, M. Chalmers, C. Lynch, J. Ried,B. Smyth, U. Straccia, and E. Toms, Personalisation and recommender systems in digital libraries, May2003, Joint NSF-EU DELOS Working Group Report.

[16] M. Deshpande and G. Karypis, Selective markov models for predicting web-page accesses, ACM Transactions on Internet Technology 4 (2004), no. 2

[17] M. Eirinaki and M. Vazirgiannis, Web mining for web personalization., ACM Trans. Internet Techn. 3 (2003), no. 1.

[18] L. Fern´andez, J. Alfredo S´anchez, and A. Garc´ıa,Mibiblio: personal spaces in a digital library universe. In The 5th ACM Conference on Digital Libraries,(ACM DL), 2000, pp. 232–233.

[19] B. Mobasher, Data mining for personalization. in theadaptive web: Methods and strategies of web personalization., (2006).

[20] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa,Using sequential and non-sequential patterns in predictive web usage mining tasks., International Conference on Data Mining (ICDM), 2002.

[21]W.Park,W. Kim, S. Kang, H. Lee, and Y.-K. Kim, Personalized digital e-library service using users' profile information., 10th European Conference on Digital Libraries (ECDL), 2006, pp. 528–531.

[22] D. Pierrakos, G. Paliouras, C. Papatheodorou, and C. D. Spyropoulos, Web usage mining as a tool for personalization: A survey., User Model. User-Adapt. Interact. 13 (2003), no. 4, 311–372.

[23] E. Rasmussen, Clustering algorithms, 1992, in: W.Frakes, R. Baeza-Yates (Eds.), Information Retrieval:Data Structures and Algorithms, Prentice Hall.

[24] Bamshad Mobasher, Robert Cooley, Jaideep Srivastava' "Automatic Personalization Based on Web Usage Mining".