Personalizing Web Directories with Community Discovery Algorithm

Sriram K.P
ME Computer Science & Engineering,
SMK Fomra Institute of Technology,
Kelambakkam, Chennai-603103.
leosri888@gmail.com

Joel Robinson
Department of Computer Science & Engineering,
SMK Fomra Institute of Technology,
Kelambakkam, Chennai-603103.
joelnaz@yahoo.co.in

Abstract—This paper presents a community discovery Algorithm for the development of Community Web Directories applying personalization on Web pages and web directories. In this context, the Web directory is viewed as a hierarchy of web related information and personalization is realized by constructing user community on the basis of usage data by a particular User. In contrast to most of the work on Web usage mining, the usage data that are analyzed here correspond to user navigation throughout the Web, rather than a particular Web site, exhibiting as a result a high degree of diversity. For modeling the user communities, we introduce a novel methodology that combines the users' browsing behavior with information from the Web directories. Following this methodology, we enhance the clustering and probabilistic approaches presented in existing work and also present a new algorithm that combines these two approaches. The resulting community models take the form of Community Web Directories. From this web directory our proposed community discovery algorithm finds the best content to be displayed. The experiments also assess the effectiveness of the different machine learning techniques on the task.

I INTRODUCTION

At its current state, the Web has not achieved its goal of providing easy access to online information. As its size is increasing, the abundance of available information on the Web causes the frustrating phenomenon of "information overload" to Web users. Organization of the Web content into thematic hierarchies is an attempt to alleviate the problem. These hierarchies are known as Web Directories and correspond to listings of topics which are organized and overseen by humans. A Web directory, such as Yahoo (www.yahoo.com) and the Open Directory Project (ODP) (dmoz.org), allows users to find Web sites related to the topic they are interested in, by starting with broad categories and gradually narrowing down, choosing the category most related to their interests. However, the information for the topic that a user is seeking might reside very deep inside the directory.

Hence, the size and the complexity of the Web directory itself are cancelling out the gains that were expected with respect to the information overload problem, i.e., it is often difficult to navigate to the information of interest to a particular user. On the other hand, Web Personalization [1], i.e., the task of making Web-based information systems adaptive to the needs and interests of individual users, or groups of users, emerges as an important means to tackle information overload. However, in achieving personalization, we are confronted with the difficult task of acquiring and creating accurate and operational user models. Reliance on manual creation of these models, either by the users or by domain experts, is inadequate for various reasons, among which the annoyance of the users and the difficulty of verifying and maintaining the resulting models. Web Usage Mining [2] is an approach that employs knowledge discovery from usage data to automate the creation of user models [3]. We claim that we can overcome the deficiencies of Web directories and Web personalization by combining their strengths, providing a new tool to fight information overload. In particular, we focus on the construction of usable Web directories that model the interests of groups of users, known as user communities. The construction of user community models, i.e., usage patterns representing the browsing preferences of the community members, with the aid of Web Usage Mining has primarily been studied in the context of specific Web sites [4]. In our work, we have extended this approach to a much larger portion of the Web through the analysis of usage data collected by the proxy servers of an Internet Service Provider (ISP).

More specifically, we present a knowledge discovery framework for constructing community-specific Web directories. Community Web Directories exemplify a new objective of Web personalization, beyond Web page recommendations [5], [6], or adaptive Web sites [7]. The members of a community can use the community directory as a starting point for navigating the Web, based on the topics that they are interested in, without the requirement of accessing vast Web directories. Moreover, community Web directories can be exploited by Web search engines to provide personalized results to queries. The construction of community directories with usage mining raises a number of interesting research issues, which are addressed in this paper. One of the challenges is the analysis of large data sets in order to identify community behavior. Moreover and apart from the heavy traffic expected at a central node, such as an ISP proxy server, a peculiarity of the data is that they do not correspond to hits within the boundaries of a site, but record outgoing traffic to the whole of the Web. This fact leads to increased dimensionality and semantic incoherence of the data, i.e., the Web pages that have been accessed.

II RELATED WORK

Web usage mining has been used extensively for Web personalization. A number of personalized services employ machine learning methods, particularly clustering techniques, to analyze Web usage data and extract useful knowledge for the recommendation of links to follow within a site, or for the customization of Web sites to the preferences of the users. A thorough analysis of these methods, together with their pros and cons in the context of Web Personalization, is presented in [3], [11], and [12]. PLSA has been used in the context of Collaborative Filtering [13] and Web Usage Mining [14]. In the first case, PLSA was used to construct a model-based framework that describes user ratings. Latent factors were employed to model unobservable motives, which were then used to identify similar users and items, in order to predict subsequent user ratings. In [14], PLSA was used to identify and characterize user interests inside certain Web sites. The latent factors segmented user sessions to support a personalized recommendation process. A similar approach was followed in [15], where each user session was “mapped” onto a sequence of latent factors, named tasks, that correspond to a more abstract view of user behavior. More recent work [15] used PLSA to cluster users in legitimate and malicious (“shilling”) groups. In this paper, we propose a knowledge discovery framework for building Web directories according to the preferences of user communities.

Community Web directories are more appropriate than personal user models for personalization across Web sites, since they aggregate statistics for many users under a predefined thematic taxonomy, thus making it possible to handle a large amount of data, residing in a sparse dimensional space. To our knowledge, this is the first attempt to construct aggregate user models, i.e., communities, using navigational data from the whole Web. Compared to our earlier work on this topic, in this paper, we address the problem of “local overload.” We achieve this by combining thematic with usage information to model the user communities. On this basis, we present new versions of the approaches introduced in [8] and [9] and a new method that combines crisp clustering with probabilistic models.

III PROPOSED MODEL AND ALGORITHM

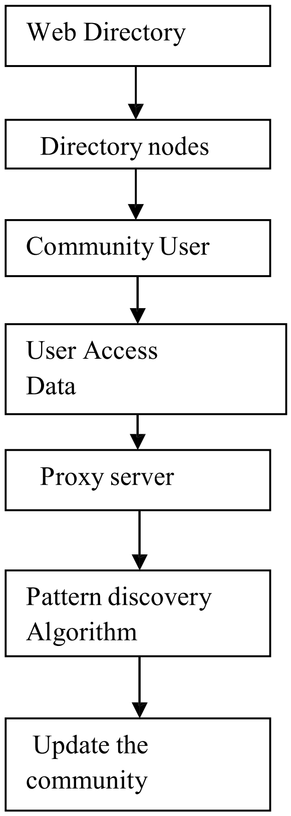


Fig 1: Work Flow Diagram

In the Fig 1, we describe the pattern discovery methodology that we propose for the construction of the community Web directories. This methodology aims at the selection of categories of the Web directory that satisfy the criteria mentioned in the previous sections, i.e., community as well as objective informativeness. The selected categories are used to construct the subgraph of the community Web directory. The input to the pattern discovery algorithms is the user sessions, which have been mapped to thematic user sessions. We note that there is a many-to-one mapping of the Web pages to the leaf categories of the Web directory. In other words, more than one Web page within a user session can be mapped to the same leaf category. Thus, the number of distinct entries in a thematic user session ui is generally smaller than its simple counterpart _i that contains Web pages. The removal of duplicates leads to the thematic session set which is defined as follows:

Definition (Thematic session set). Let $u(t_1,t_f)$ is a thematic user session, then its session set $\bar{u}(t_1,t_f)$ = $\{l_i,l_{i€}C\}$ is the set of unique categories in $v(t_1,t_f)$.

The assignment of user sessions to thematic session sets results in the loss of the sequential nature of the user’s browsing behavior. This has a limited effect on our community discovery methodology since we are focusing on the interests, rather than the navigation patterns of the user. The occurrence of categories in sessions is expressed by the appearance of the categories in thematic session sets. In our approach, we extend the relation between sessions and categories to cover the ancestor categories which are also assumed to characterize the accessed Web page.

Enhanced Community Directory Miner (ECDM):

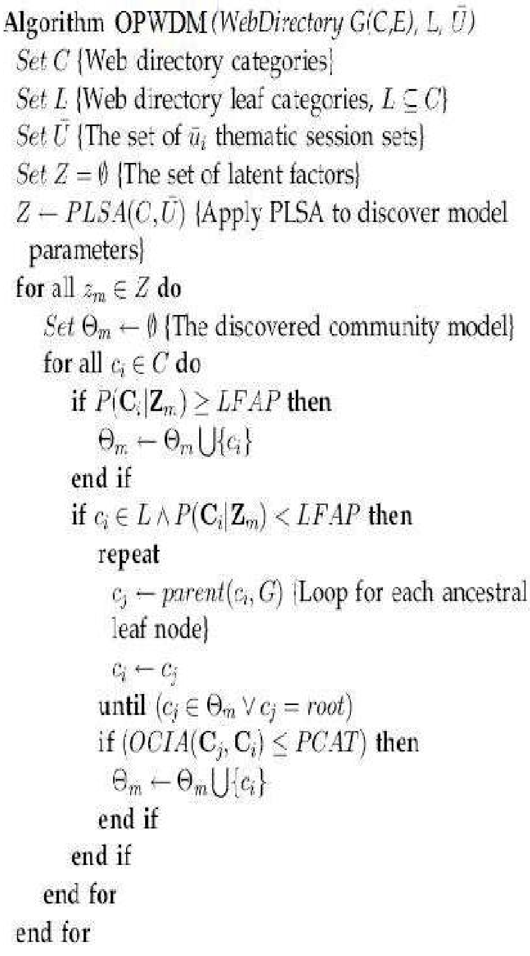
The first machine learning method that we employed for community discovery is the CDM algorithm [8]. The enhanced version of CDM, incorporating the OCIA criterion, is named as Enhanced-Community Directory Miner (ECDM). Similar to CDM, ECDM is based on the cluster mining algorithm which has been employed earlier [4] for site-specific community discovery. Cluster mining discovers patterns of common behavior by looking for all maximal fully connected subgraphs (cliques) of a graph that represents the users’ characteristic features, i.e., thematic categories in our case. The algorithm starts by constructing the graph, the vertices of which correspond to the categories, while the edges to category co-occurrence in thematic session sets. Vertices and edges are associated with weights, which are computed as the category occurrence and co-occurrence frequencies, respectively.

The connectivity of the graph is usually high. For this reason, we make use of a connectivity threshold that reduces the edges of the graph. This threshold is related to the frequency of co-occurrence of the thematic categories in the data. Once the connectivity of the graph has been reduced, the weighted graph is turned to an unweighted one. Finally, all maximal cliques of the unweighted graph are generated, each one corresponding to a community model. One important advantage of this approach is that each user may be assigned to many communities, unlike most crisp user clustering methods. Moreover, the clusters generated by ECDM group together characteristic features of the user. Each clique discovered by ECDM is thus already a community model, i.e., a set of interesting categories. The ECDM algorithm incorporating the OCIA criterion can be summarized in the following steps,

Step 1. Compute frequencies of categories that correspond to the weights of the vertices and co-occurrence frequencies between categories that correspond to the weights of the edges
Step 2. Introduce a connectivity threshold to remove the edges of the graph with weights less than or equal to its value.
Step 3. Turn the weighted graph of categories into an unweighted one by removing all the weights from the nodes and the edges, and find all the maximal cliques, e.g., as proposed in [10].
Step 4. Select informative leaves for the hierarchy.

Objective Probabilistic Web Directory Miner (OPWDM)

In the ECDM algorithm discussed in the previous section, the constructed patterns are based solely on the "observable" behavior of the users, as this is recorded in the usage data. However, it is rather simplifying to assume that relations between users are based only on observable characteristics of their behavior. Generally, users' interests and motives are less explicit.

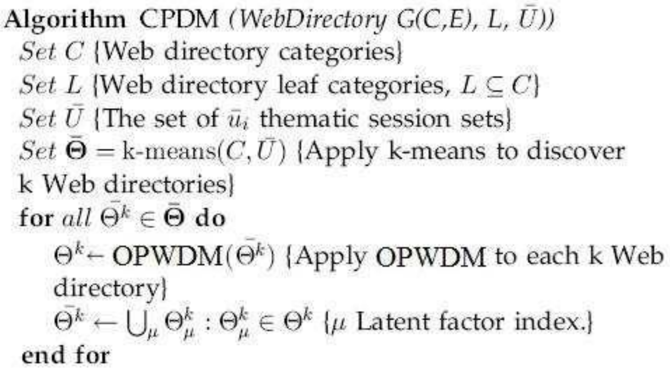


```
Algorithm OPWDM (WebDirectory G(C,E), L, $\bar{U}$)
 Set $C$ {Web directory categories}
 Set $L$ {Web directory leaf categories, $L \subseteq C$}
 Set $\bar{U}$ {The set of $\bar{u}_i$ thematic session sets}
 Set $Z = \emptyset$ {The set of latent factors}
 $Z \leftarrow PLSA(C, \bar{U})$ {Apply PLSA to discover model
   parameters}
 for all $z_m \in Z$ do
     Set $\Theta_m \leftarrow \emptyset$ {The discovered community model}
     for all $c_i \in C$ do
         if $P(\mathbf{C}_i|\mathbf{Z}_m) \geq LFAP$ then
             $\Theta_m \leftarrow \Theta_m \bigcup \{c_i\}$
         end if
         if $c_i \in L \wedge P(\mathbf{C}_i|\mathbf{Z}_m) < LFAP$ then
             repeat
               $c_j \leftarrow parent(c_i, G)$ {Loop for each ancestral of
               leaf node}
               $c_i \leftarrow c_j$
             until $(c_j \in \Theta_m \vee c_j = root)$
             if $(OCIA(\mathbf{C}_j, \mathbf{C}_i) \leq PCAT)$ then
               $\Theta_m \leftarrow \Theta_m \bigcup \{c_i\}$
             end if
         end if
     end for
 end for
```

The algorithm is shown above where the functioning of the OPWDM is given, two users might visit pages from a particular category of the Web directory, not because they have been motivated by the exact same interests, but only by a common "subset" of them. Thus, in this method, we follow the rationale introduced in [9] and assume that the users' choices are motivated by a number of latent factors that correspond to these subsets. These factors are responsible for the associations between users. The presence of latent factors that justify user interests provides a generic approach to the identification of patterns in usage data and can be used for grouping the users. The advantage of this methodology is that it allows us to describe more effectively the multidimensional characteristics of user interests. As an example, assume that a user navigates through Web pages that belong to the category "Top/ Computer/Companies" because of the existence of a latent factor z. This action might have been motivated by the user's interest in e-commerce. Another user might arrive at the same category because she is interested in job offers. The interest of the second user corresponds to a different motive, represented by a different latent factor z. Despite the simplicity of this example, we can see how different motives may result in similar observable behavior in the context of a Web directory.

Clustering and Probabilistic Directory Miner (CPDM)

In addition to the enhanced methods presented above, we also introduce here a new hybrid method for the discovery of community models. This method combines a clustering algorithm with PLSA. We apply the popular k-means [13] clustering algorithm on the relation R=(Ū,C) for the creation of the initial communities. This approach differs from CDM clustering, as it produces nonoverlappingclusters, i.e., each category belongs to only one cluster. However, as we have explained above for PLSA, the explicit modeling of latent factors is considered advantageous. Thus, we can assume that in addition to the k-means clusters, further hidden associations exist in the data, i.e., subcommunities inside the cluster that are not directly observable.



```
Algorithm CPDM (WebDirectory G(C,E), L, $\bar{U}$))
  Set $C$ {Web directory categories}
  Set $L$ {Web directory leaf categories, $L \subseteq C$}
  Set $\bar{U}$ {The set of $\bar{u}_i$ thematic session sets}
  Set $\bar{\Theta}$ = k-means$(C, \bar{U})$ {Apply k-means to discover
  k Web directories}
  for all $\bar{\Theta^k} \in \bar{\Theta}$ do
      $\Theta^k \leftarrow$ OPWDM$(\bar{\Theta^k})$ {Apply OPWDM to each k Web
      directory}
      $\bar{\Theta^k} \leftarrow \bigcup_\mu \Theta^k_\mu : \Theta^k_\mu \in \Theta^k$ {$\mu$ Latent factor index.}
  end for
```

The algorithm shown above is to discover the hidden knowledge that we map each cluster derived by k-means onto a new space of latent factors. In this manner, the community Web directories are constructed using both observable and latent associations inthe data, and potentially allow us to better model theinterests of users.In order to discover the community models, we enhance the k-means community construction process by identifying latent associations inside each k-means cluster.

IV EXPERIMENTAL ANALYSIS AND RESULTS

The methodology introduced in this paper for the construction of community Web directories has been tested in the context of a research project,which focused on the analysis of usage data from the proxy server logs of an Internet Service Provider. The evaluation procedure is described in the following sections.

Experimental Setup:

The evaluation process assessed the performance of the algorithms on the ODP categories. In particular, we examined the first six levels of the ODP thematic taxonomy, which include 59,863 categories. We analyzed log files consisting of 781,069 records, i.e., Web page requests. Data cleaning was performed and the remaining data, i.e., 18,459 Web pages, were downloaded locally using a Web crawler. Based on log data, we constructed 3,286 user sessions using a time interval of 60 minutes as a threshold on the "silence" period between two consecutive requests from the same IP. The initial mapping of Web pages to the ODP hierarchy was achieved with the method described in [10]. Using these data, we built the community models with the three pattern discovery algorithms.

For the ECDM algorithm, the results presented in this work correspond to the communities created by varying the values of the connectivity threshold. Similarly, the LFAP threshold is varied in the other two methods. In the case of the OPWDM approach, the models were built using 5, 10, 15, and 20 latent factors. For the CPDM, five clusters were built by the k-means algorithm and the PLSA enhancement involved the modeling of five latent factors per cluster, leading to a comparable number of communities as OPWDM clustering. The PCAT threshold was also varied for all of the discovery methods to measure the impact of the OCIA criterion. The results are also compared against the application of the same algorithms to the artificial Web directory that was used in our past work [9]. In this scenario, the Web pages were clustered using hierarchical agglomerative clustering. Following the criteria discussed in [14] for the selection of the number of clusters, the process resulted in the creation of 998 distinct categories. Based on these data, we constructed 2,253 user sessions, using the same time interval of 60 minutes.

In all of the experiments, we used 10-fold cross validation in order to obtain an unbiased estimate of the performance of the methods. For each pattern discovery method, we trained the model 10 times, each time leaving out one of 10 subsets of the data, and used the omitted subset to evaluate the model. Therefore, the results that we present are always the average of 10 runs for each experiment. As an initial measure of performance, we measured the shrinkage of the original Web directory, achieved by the pattern discovery algorithms. This was measured by comparing the Average Path Length of the original directory to that of the community directories. The Average Path Length was computed by calculating the average number of nodes from the root to the leaves of a directory. Additionally, we examined the effectiveness of the discovered models, i.e., the way that users benefit from the resulting community Web directories. In order to measure effectiveness, we followed an approach commonly used for recommendation systems [15].

We have started with the assumption that users are ultimately looking for Web pages inside the Web directory. We have hidden one Web page in each test user session and used the rest of the session, actually its thematic counterpart, to choose the most appropriate community directory. We call the hidden Web page "target" as it is the one driving the evaluation. More specifically, we examined whether and how the user can get to the target page, using the community Web directory to which the particular thematic user session is assigned. The assignment of a user session to a community directory is based on the observed Web pages of the session. Motivating this choice, one can consider the extreme scenario where a thematic session set contains a single category, i.e., all the Web pages of the user session have been mapped to the same category. In that case, we cannot identify target and observed categories, and thus, we cannot evaluate the session.

Step 1. For each of the observed categories in the test session, identify the community directories that contain it, if any.

Step 2. Since the categories in the session might belong in more than one community directory, identify the three most prevalent community directories for the session. In the case of the ECDM algorithm, we select the directories that contain the largest portion of the categories in a particular session, while in the case of OPWDM and CPDM, we select for each category c_j the three community directories that maximize the probability $p(C_j|Z_k)$.

Step 3. From all the selected community directories, a new session-specific directory is constructed by merging the hierarchies. This approach enables the transition from community to session-specific user models.

Results:

Using the methodology and metrics presented above, we performed experiments to evaluate the three discovery methods. The results are also compared to those obtained with versions of the algorithms that do not use the OCIA criterion. The experiments have been performed for a large number of value pairs for the Connectivity/LFAP and the PCAT thresholds. In the case of the OPWDM, we obtained very similar results for different numbers of factors and present here only the results for 20 latent factors. To begin with, we examine the percentage of shrinkage of the Web directories, achieved by our personalization methodologies. This was measured by comparing the Average Path Length of the original directories to that of the community directories. Due to the lack of space, we present here only the results for the ODP community Web directory using the OPWDM algorithm in Fig. 2. In this figure, we first observe that the size of the ODP directory can be reduced drastically.

The average path length of the directory is reduced up to almost 60 percent, as the values of the LFAP threshold increase. This is an indication that the method is very selective in the leaf nodes that it chooses to maintain. Continuing the evaluation process, we turn to theeffectiveness of the approach. The measures that we examine are coverage and user gain. Typically, there is a trade-off between coverage and user gain. It is interesting to measure this trade-off and identify good operating points. The usual choice for such trade-off measure is the use of Receiver Operating Characteristics (ROCs) curves that have been used extensively in evaluating diagnostic systems. Adapting the idea of ROC curves to our measures, we plot coverage against (1-User Gain). We name this plot a trade-off curve since we are not measuring exactly sensitivity and specificity as commonly done in ROC analysis.

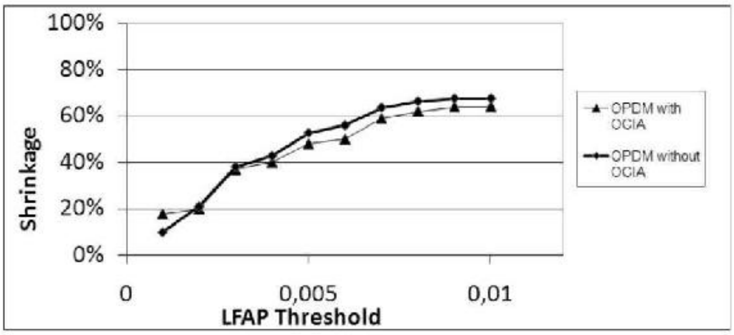


Fig 2: ODP Web Directory Shrinkage using OPWDM

V CONCLUSION

This paper advocates the concept of a community Web directory, as a Web directory that specializes to the needs and interests of particular user communities. Furthermore, it presents the complete methodology for the construction of such directories with the aid of machine learning methods. User community models take the form of thematic hierarchies and are constructed by employing clustering and probabilistic learning approaches. We applied our methodology to the ODP directory, as well as to an artificial Web directory, which was generated by clustering Web pages that appear in the access log of a Web proxy. For the discovery of the community models, we introduced a new criterion that combines the a priori thematic informativeness of the Web directory categories with the level of interest observed in the usage data. In this context, we introduced and evaluated three learning methods. We have tested the methodology using access logs from the proxy servers of an Internet Service Provider and provided results that are indicative of the behavior of the algorithms and the usability of the community Web directories. Proxy server logs have introduced a number of interesting challenges, such as the handling of their size and semantic diversity. The proposed methodology addresses these issues by reducing the dimensionality of the problem, through the classification of individual Web pages into the categories of the directory.

The proposed methodology provides a promising research direction, where many new issues arise. An analysis regarding the parameters of the community models, such as PLSA, is required. Moreover, additional evaluation on the robustness of the algorithms to a changing environment would be interesting. Furthermore, other knowledge discovery methods could be adapted to the task of discovering community directories and compared to the algorithms presented here. In addition, other classification methods could be exploited for the initial mapping of the Web pages to the Web directory.

ACKNOWLEDGMENT

The authors would like to thank all those who have extended their support for successful completion of this work.

REFERENCES

[1] DimitriosPierrakos, Member, IEEE, and Georgios Paliouras, “Personalizing Web Directories with the Aid of Web Usage Data” IEEETransactions on knowledge and data engineering, vol. 22, no. 9, september 2010

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

[3] 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.

[4] 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.

[5] G. Paliouras, C. Papatheodorou, V. Karkaletsis, 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.

[6] 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.

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

[8] D. Pierrakos, G. Paliouras, C. Papatheodorou, V. Karkaletsis, andM. 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 LatentInformation 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, “ComputationalIntelligence Techniques for Web Personalization,” Web Intelligence and Agent Systems, vol.6, no. 3, pp. 253-272, 2008.

[13] T. Hofmann, “Learning What People (Don’t) Want,” Proc. 12thEuropean Conf. in Machine Learning, pp. 214-225, 2001.

[14] X. Jin, Y. Zhou, and B. Mobasher, “Web Usage Mining Based on Probabilistic Latent Semantic Analysis,” Proc. ACM SIGKDD, pp. 197-205, Aug. 2004.

[15] X. Jin, Y. Zhou, and B. Mobasher, “Task-Oriented Web UserModeling for Recommendation,” Proc. 10th Int’l Conf. User Modeling, L. Ardissono, P. Brna, and A. Mitrovic, eds., pp. 109- 118, 2005.