Exploiting Probabilistic Latent Information for the Construction of Community Web Directories

Dimitrios Pierrakos[1,2] and Georgios Paliouras[1]

[1] Institute of Informatics and Telecommunications, NCSR “Demokritos”,
15310 Ag. Paraskevi, Greece,
{dpie, paliourg,}@iit.demokritos.gr

[2] Department of Informatics and Telecommunications, University of Athens,
Panepistimiopolis, Ilissia Athens 15784, Greece

Abstract. This paper improves a recently-presented approach to Web Personalization, named Community Web Directories, which applies personalization techniques to Web Directories. The Web directory is viewed as a concept hierarchy and personalization is realized by constructing user community models on the basis of usage data collected by the proxy servers of an Internet Service Provider. The user communities are modeled using Probabilistic Latent Semantic Analysis (PLSA), which provides a number of advantages such as overlapping communities, as well as a good rationale for the associations that exist in the data. The data that are analyzed present challenging peculiarities such as their large volume and semantic diversity. Initial results presented in this paper illustrate the effectiveness of the new method.

1 Introduction

The hypergraphical architecture of the Web has been used to support claims that the Web will make Internet-based services really user-friendly. However, due to its almost unstructured and heterogeneous environment, as well as its galloping growth, the Web has not realized the goal of providing easy access to online information. Information overload is one of the Web’s major shortcomings that place obstacles in the way users access the required information.

An approach towards the alleviation of this problem is the organization of Web content into thematic hierarchies, also known as Web directories, such as Yahoo! [14] or the Open Directory Project (ODP) [11], that allow users to locate required information. However their size and complexity are canceling out any 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. A different approach is Web personalization [8], which focuses on the adaptability of Web-based information systems to the needs and interests of individual users, or groups of users. A major obstacle though towards realizing this solution is the acquisition of accurate and operational models for the users. Reliance to 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. An alternative approach is that of Web Usage Mining [13], which provides a methodology for the collection and preprocessing of usage data, and the construction of models representing the behavior and the interests of users [10].

In recent work [9], we proposed, the concept of *Community Web Directories*, which combines the strengths of Web Directories and Web Personalization, in order to address some of the above-mentioned issues. Community Web Directories are usable Web directories that correspond to the interests of groups of users, known as user communities. 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. For the construction of Community Web directories, we have presented the *Community Directory Miner, (CDM)* algorithm, which was able to produce a suitable level of semantic characterization of the interests of a particular user community. This approach, similar to other clustering approaches, is based on relations between the users, that correspond to observable patterns in the usage data. However, there also exists a number of latent factors that are responsible for the observable associations. These latent factors can be thought of as the motivation of a particular user accessing a certain page, and therefore groups of users, can be constructed sharing common latent motives. In the case of Web directories this method could provide a more thorough insight of the patterns that exist in the usage data. A common method for discovering latent factors in data is Probabilistic Latent Semantic Analysis (PLSA), a technique that has been used extensively in Information Retrieval and Indexing [5]. In this work we employ this method in order to identify user communities.

The rest of this paper is organized as follows: Section 2 presents existing approaches to Web personalization with usage mining methods, as well as approaches to the construction of personalized Web directories. Section 3 presents our methodology for the construction of Community Web directories. Section 4 provides results of the application of the methodology to the usage data of an Internet Service Provider (ISP). Finally section 5 summarizes interesting conclusions of this work and presents promising paths for future research.

2 Related Work

In recent years, Web personalization has attracted considerable attention. To realize this task, a number of applications employ machine learning methods and in particular clustering techniques, that analyze Web usage data and exploit the extracted knowledge for the recommendation of links to follow within a site, or for the customization of Web sites to the preferences of the users. In [10] a thorough analysis of the above methods is presented, together with their pros and cons in the context of Web Personalization. Personalized Web directories, on the other hand, are mainly associated with services such as Yahoo! [14] and Excite [4], which support the manual personalization of their directories by the user. An initial approach to automate this process, is the Montage system [1], which is used to create personalized portals, consisting primarily of links to the Web pages that a particular user has visited, while also organizing the links into thematic categories according to the ODP directory. A technique for WAP portal personalization is presented in [12], where the portal structure is adapted to the preferences of individual users. A related approach is presented in [3], where a Web directory, is used as a "reference" ontology and the web pages navigated by a user are mapped onto this ontology using document classification techniques, resulting in a personalized ontology.

The scalability of the content-based classification methods and their questionable extensibility to aggregate user models such as user communities, raise important issues for the above methods.

Probabilistic Latent Semantic Analysis has already been used for Web Personalization, in the context of Collaborative Filtering [6], and Web Usage Mining [7]. 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 [7], PLSA was used to identify and characterize user interests inside certain Web sites. The latent factors derived by the modeling process were employed to segment user sessions and personalization services took the form of a recommendation process.

In this paper we extend the methodology of our previous work for building Web directories according to the preferences of user communities which are now modeled with the use of PLSA. The methodology presented in this paper proposes a new way of exploiting the knowledge that is extracted by the analysis of usage data. Instead of link recommendation or site customization, it focuses on the construction of Community Web directories, as a new way of personalizing the Web. The construction of the communities is based on usage data collected by the proxy servers of an Internet Service Provider (ISP). This type of data, in contrast to the works mentioned above, has a number of peculiarities, such as its large volume and its semantic diversity, as it records the navigational behavior of the user throughout the Web, rather than within a particular Web site. The methodology presented in this paper handles these problems, focusing on the new personalization modality of Community Web directories.

3 Constructing Community Web Directories

The construction of Community Web directories is seen here as the end result of an analysis of Web usage data collected at the proxy servers of a central service on the Web. The details of the process are described in our previous work [9]. In brief, this process involves the thematic categorization of Web pages, thus reducing the dimensionality and semantic diversity of the data. A hierarchical agglomerative clustering approach [15], is used to build a taxonomy from Web pages included in the log files, based on terms that are frequently encountered in the Web pages. The resulting taxonomy of thematic categories forms the base Web directory. Furthermore usage data are transformed into access sessions, where each access session is a sequence of accesses to Web pages by the same IP address, when the time interval between two subsequent entries does not exceed a certain time interval. Pages are mapped onto thematic categories that correspond to the leaves of the hierarchy and therefore an access session is translated into a sequence of categories from the Web directory.

3.1 Building the Probabilistic Model

Most of the work on Web usage mining uses clustering methods or association rule mining, in order to identify navigation patterns based solely on the "observable" behavior of the users, as this is recorded in the usage data. For instance, pages accessed by users are a typical observable piece of navigational behavior. However, it is rather simplifying to assume that relations between users are based only on observable characteristics of their behavior. We are relaxing this assumption and consider a number of latent factors, that control the user behavior and are responsible for the existence of associations between users. These latent factors can be exploited to construct clusters of users which share common motivation. The existence of latent factors that rule user behavior provides a more generic approach for the identification of patterns in usage data and thus provides better insight into the users' behavior.

As an example, assume that user u navigates through Web pages that belong in the *category*="computer companies" because of the existence of a latent cause-factor *z*. This cause might be the user's interest in finding information for business-to-business commerce. However, another user *u'* might arrive at the same category because she is interested in job offers. Hence, the interest of the second user corresponds to the existence of 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.

A commonly used technique for the identification of latent factors in data is the PLSA method [5], which is supported by a strong statistical model. Applying PLSA to our scenario of Web directories we consider that there exists a set of user sessions $U=\{u_1, u_2,\ldots,u_i\}$, a set of Web directory categories $C=\{c_1, c_2,\ldots,c_j\}$, as well as their binary associations (u_i, c_j) which correspond to the access of a certain category c_j during the session u_i. The PLSA model is based on the assumption that each instance, i.e. each observation of a certain category inside a user session, is related to the existence of a latent factor, z_k that belongs to the set $Z= \{z_1, z_2,\ldots,z_k\}$. We define the probabilities $\mathrm{P}(u_i)$: the a priori probability of a user session u_i, $\mathrm{P}(z_k|u_i)$: the conditional probability of the latent factor z_k being associated with the user session u_i and $\mathrm{P}(c_j|z_k)$: the conditional probability of accessing category c_j, given the latent factor z_k. Using these definitions, we can describe a probabilistic model for generating session-category pairs by selecting a user session with probability $\mathrm{P}(u_i)$, selecting a latent factor z_k with probability $\mathrm{P}(z_k|u_i)$ and selecting a category c_j with probability $\mathrm{P}(c_j|z_k)$, given the factor z_k. This process allows us to estimate the probability of observing a particular session-category pair $(u_i$,$c_j)$, using joint probabilities as follows:

$$P(u_i, c_j) = P(u_i)P(c_j|u_i) = P(u_i)\sum_k P(c_j|z_k)P(z_k|u_i). \tag{1}$$

Using Bayes's theorem we obtain the equivalent equation:

$$P(u_i, c_j) = \sum_k P(z_k)P(u_i|z_k)P(c_j|z_k). \tag{2}$$

Equation 2 leads us to an intuitive conclusion about the probabilistic model that we exploit: each session-category pair is observed due to a latent generative factor that corresponds to the variable z_k and hence it provides a more generic association between the elements of the pairs. However, the theoretic description of the model does not make it directly useful, since all the probabilities that we introduced are not available a priori. These probabilities are the unknown parameters of the model, and they can be estimated using the *Expectation-Maximization* (EM) algorithm, as described in [5].

3.2 Extraction of User Communities

Using the above probabilities we can assign categories to clusters based on the k factors z_k that are considered responsible for the associations between the data. This is realized by introducing a threshold value, named *Latent Factor Assignment Probability, (LFAP)* for the probabilities $P(c_j|z_k)$ and selecting those categories that are above this threshold. More formally, with each of the latent factors z_k we associate the categories that satisfy:

$$P(c_j|z_k) \geq LFAP. \tag{3}$$

In this manner and for each latent factor, the selected categories are used to construct a new Web directory. This corresponds to a topic tree, representing the community model, i.e., usage patterns that occur due to the latent factors in the data. A number of categories from the initial Web directory have been pruned, resulting in a reduced directory, named community Web directory. This approach has a number of advantages. First is the obvious shrinkage of the initial Web directory, which is directly related with the interests of the user community, ignoring all other categories that are irrelevant. Second, the selected approach allows us to construct overlapping patterns, i.e. a category might belong to more than one community directories, i.e. affected by more than one latent factor. A pictorial view of a “snapshot” of such a directory is shown in Figure 1, where the community directory is “superimposed” onto the original Web directory. For the sake of brevity we choose to label each category using a numeric coding scheme, representing the path from the root to the category node, e.g. “1.4.8.19” where “1” corresponds to the root of the tree. Each category also labelled by the most frequent terms of the Web pages that belong in it. Grey boxes represent the categories that belong in a particular community directory, while big white boxes represent the rest of the categories in the Web directory, which are not included in the model. The categories “12.2” and “18.79” (the smaller white boxes) have been pruned, since they do not have more than one child in the community directory and are thus redundant.

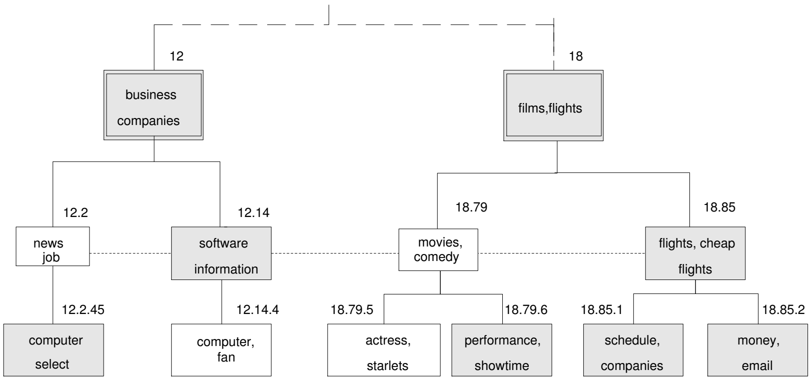

Fig. 1. An example of a Community Web Directory.

4 Experimental 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 focuses on the analysis of usage data from the proxy server logs of an Internet Service Provider. We analyzed log files consisting of 781,069 records and using hierarchical agglomerative clustering [15] we obtained 998 distinct categories. We also constructed 2,253 user sessions, using a time-interval of 60 minutes as a threshold on the "silence" period between two consecutive requests from the same IP. At the next step we built the PLSA models, varying the number of the latent factors. We used 10-fold cross validation, in order to obtain an unbiased estimate of the performance of the method. We train the model 10 times, each time leaving out one of the subsets from training, and employ 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, compared to the community directories derived by the PLSA model. This was measured via the average path length of the original directory and the community directories. These values were computed by calculating the number of nodes from the root to each leaf of a directory. The results are depicted in Figure 2, taken for various values of the LFAP threshold discussed in Section 3.2. From these results we can derive that the length of the paths is dramatically reduced, up to 50%, as the threshold increases. This means that the users have to follow much shorter paths to arrive at the their interests. Furthermore, the method seems to be robust to the choice of the number of factors.

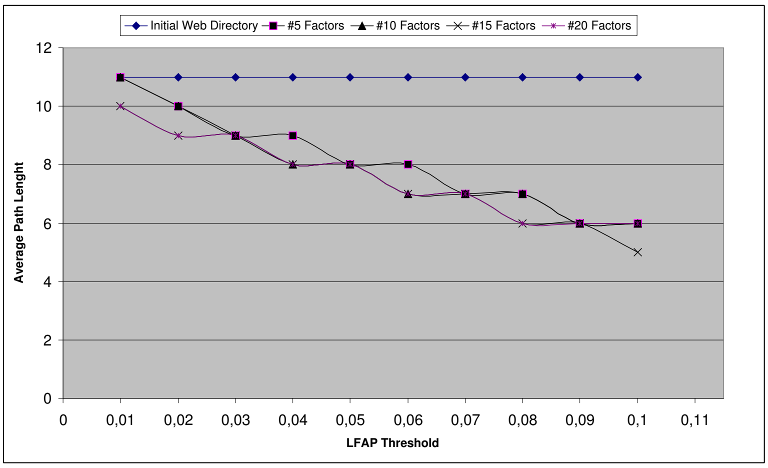

Fig. 2. Average Path Length.

The community directories are further processed for the evaluation tasks as follows: a user session is assigned to a community directory based on its conditional probability against the latent factor, $P(u_i|z_k)$, that defines the community directory. However, the PLSA model, allows user sessions to belong to more than one community directory, and hence we identified the most prevalent community directories, i.e. the community directories where user sessions have the highest conditional probabilities for the respective factors. For our scenario we select the three most prevalent directories. From these prevalent community directories a new and larger directory is constructed by joining the respective hierarchies, resulting in a session-specific community directory. The resulting, session-specific directories are used in further evaluation.

The next stage of the evaluation process consisted of analyzing the usability of our models, i.e. the way that users benefit from the community Directories. We have focused on: (a) how well our model can predict what the user is looking for, and (b) what the user gains by using a community directory against using the original directory. In order to define suitable metrics, we followed a common approach used for recommendation systems [2], i.e., we have hidden each hit, i.e. category, of each user session, and tested to see whether and how the user can get to it, using the community directory to which the particular user session is assigned. The hidden category is called the “target” category here. The first metric that we used measured the coverage of our model, which corresponds to its predictiveness, i.e. the number of target categories that are covered by the corresponding community directories. This is achieved by counting the number of user sessions, for which the community directory, as explained before, covers the target category. The second metric that we used was an estimate of the actual gain that a user would have by following the community directory structure, instead of the complete directory. In order to realize this, we followed a simple approach that is based on the calculation of the effort that a user is required to exert in order to arrive at the target category. We estimated this effort based on the user's navigation path inside a directory, towards the target category. This is estimated by a metric, introduced in [9], named *ClickPath*, which takes into account the depth of the navigation path as well as the branching factor at each step. More formally:

$$ClickPath = \sum_{j=1}^{d} j * b_j, \tag{4}$$

where d the depth of the path and b_j the branching factor at the j-th step.

We have performed experiments varying the number of latent factors. Due to lack of space, we only present (Figure 3), the results obtained for 20 latent factors, varying the values of the LFAP threshold defined in Equation 3. The results for different number of factors are almost identical to the ones shown in Figure 3, starting at the same coverage and gain levels for small LFAP values and deviating slightly as the LFAP value increases. The largest difference was obtained for LFAP=0.1, for which the results are shown in Table 1. Even at that level, the choice of the number of factors does not have a large effect in the two measures. The small increase in both coverage and user gain can be explained by the fact that, as the number of latent factors increases, more community directories are created. As the number of community directories increases, more categories get the chance to appear in one of the directories, thus increasing coverage. At the same time, smaller community directories are constructed, resulting in higher user gain. However, even this small effect disappears as the number of community directories increases above 15. We also provide the results of applying the CDM algorithm to the same dataset (Figure 4).

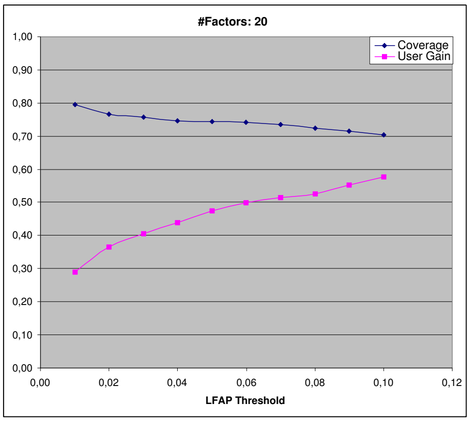

Fig. 3. PLSA Results

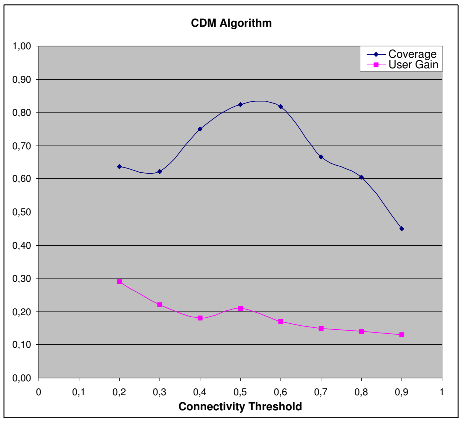

Fig. 4. CDM Results.

Table 1. Coverage and User Gain for LFAP 0.1

	#5 Factors	**#10 Factors**	**#15 Factors**	**#20 Factors**
Coverage	0.63	0.67	0.71	0.70
User Gain	0.47	0.50	0.55	0.57

From the above figures we conclude that at least 70% of the target categories are included in the community directories. At the same time the user gain reaches values higher than 50%, as more categories are pruned from the original Web directory. Practically this means that the user requires half of the effort to arrive at the required information. The results also follow a rather deterministic behaviour, i.e. large values of coverage are related with small values of user gain, as compared with the results of the CDM algorithm. These figures provide an indication of the behavior and the effectiveness of the application of PLSA to community modeling. Furthermore, they give us an initial measure of the benefits that we can obtain by personalizing Web directories to the needs and interests of user communities. However, we have only estimated the gain of the end user and we have not weighted up any "losses" that could be encountered in the case that the users would not find the interesting category that they are looking for in the personalized directory. This issue will be examined in future work.

5 Conclusions and Future Work

This paper has presented work on the concept of the Community Web Directory, introduced in our recent work, as a Web Directory that specializes to the needs and interests of particular user communities. In this case, user community models take the form of thematic hierarchies and are constructed by employing Probabilistic Latent Semantic Analysis. The initial directory is generated by a document clustering algorithm, based on the content of the pages appearing in an access log. We have tested this methodology by applying it on access logs collected at the proxy servers of an ISP and have provided initial results, indicative of the behavior of the mining algorithm and the usability of the resulting Community Web Directories. Initial results lead us to the conclusion the the application of PLSA to the analysis of user behavior in Web directories appears to be a very promising method. It has allowed us to identify latent information in the users' behavior and derive high-quality community directories that provide significant gain to their users.

In general, the combination of two different approaches to the problem of information overload on the Web, i.e. thematic hierarchies and personalization, as proposed in this paper, together with the exploitation of PLSA for the construction of community models, introduces a promising research direction, where many new issues arise. Further analysis of the PLSA method could be performed and compared with other machine learning methods, in the task of discovering community directories. Another important issue that will be examined in further work is the scalability of our approach, to larger datasets, i.e. for larger log files that would result in a larger number of sessions. However, the performance of the whole process, together with the PLSA modeling, gave us promising indications for the methods's scalability. Finally, more sophisticated metrics could also be employed for examining the usability of the resulting community directories.

References

1. C. R. Anderson and E. Horvitz. Web montage: A dynamic personalized start page. In *11th WWW Conference*, May 2002.
2. J. S. Breese, D. Heckerman, and C.M Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *UAI'98*, pages 43–52, 1998.
3. J. Chaffee and S. Gauch. Personal ontologies for web navigation. In *9th CIKM '00*, pages 227–234, 2000.
4. Excite. http://www.excite.com.
5. T. Hofmann. Probabilistic latent semantic analysis. In *UAI*, 1999.
6. T. Hofmann. Learning what people (don't) want. In *12th European Conference in Machine Learning*, pages 214–225. Springer-Verlag, 2001.
7. X. Jin, Y Zhou, and B Mobasher. Web usage mining based on probabilistic latent semantic analysis. In *KDD*, pages 197–205, 2004.
8. B. Mobasher, R. Cooley, and J. Srivastava. Automatic personalization based on web usage mining. *Communications of the ACM*, 43(8):142–151, 2000.
9. D. Pierrakos, G. Paliouras, C. Papatheodorou, V. Karkaletsis, and M. Dikaiakos. Web community directories: A new approach to web personalization. In Berendt B. et al., editor, *Web Mining: From Web to Semantic Web, EMWF 2003*, volume 3209 of *LNCS*, pages 113–129. Springer, 2004.
10. 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*, 13(4):311–372, 2003.
11. Open Directory Project. http://dmoz.org.
12. B. Smyth and C. Cotter. Personalized adaptive navigation for mobile portals. In *15th European Conference on Artificial Intelligence*. IOS Press, 2002.
13. J. Srivastava, R. Cooley, M. Deshpande, and P. T Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2):12–23, 2000.
14. Yahoo. http://www.yahoo.com.
15. Y. Zhao and G Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *CICM*, 2002.