A Web Usage Mining Framework for Web Directories Personalization

Dimitrios Pierrakos*

Department of Informatics & Telecommunications., University of Athens, Greece
Institute of Informatics & Telecommunications., NCSR "Demokritos", Athens, Greece
dpie@iit.demokritos.gr

Abstract. In this thesis we propose a novel framework that combines Web personalization and Web directories, which results in the concept of Community Web Directories. Community Web directories is a novel form of personalization performed on Web directories, that correspond to "segments" of the directory hierarchy, representing the interests and preferences of user communities. The proposed approach is based on Web usage mining and the usage data that are analyzed here correspond to user navigation throughout the Web, rather than a particular Web site. For the construction of Community Web Directories, we introduce three novel techniques that combine the users' browsing behavior with thematic information from the Web directories. Finally, we present OurDMOZ, a system that builds and maintains community Web directories.

1 Introduction

Information overload is one of the Web's major shortcomings, placing obstacles in the way users access the required information. Web Personalization, i.e., the task of making Web-based information systems adaptive to the needs and interests of individual users or group of users like *user communities*, emerges as an important means to tackle information overload. The first step towards achieving personalization is the specification of user models. However, acquiring and creating accurate and operational user models is a difficult task. Web Usage Mining is one such approach which employs knowledge discovery from data to create user models, based on the analysis of usage data, as we presented in [11].

Another attempt to alleviate the problem of information overload is the organization of the Web content into thematic hierarchies. These hierarchies are known as *Web Directories*, and correspond to listings of topics which are organized and overseen by humans. However, the size and the complexity of the Web directory are canceling out the gains that were expected with respect to the information overload problem.

The contribution of this thesis is a novel approach to overcome the deficiencies of Web personalization and Web directories by combining their strengths,

* Dissertation Advisors: Yannis Ioannidis, Professor - Georgios Paliouras, Researcher, NCSR "Demokritos"

providing a new tool to fight information overload. In particular, we focus on the construction of usable Web directories that model the interests of user communities. The construction of user community models, with the aid of Web Usage Mining has primarily been studied in the context of specific Web sites [6]. In this thesis, 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).

In the course of this thesis, we developed and evaluated three community modeling techniques, based on clustering and probabilistic learning. These techniques allowed us to take advantage of existing Web directories and specialize them to the interests of particular communities. In addition to handling the "global information overload" problem, the proposed methods also deal effectively with the "local overload" problem. This problem is a side-effect of the pruning of a number of leaf nodes of the initial Web directory, which pushes the information that they contained, i.e., the terminal links to Web pages, upwards in the hierarchy. This leads to increased information density in some leaf nodes of the personalized directory. In order to address this issue, the proposed methods combine usage data with thematic information from the original Web directories.

The proposed methodology is tested on two types of Web directory: an *artificial Web directory*, that was constructed using document clustering from the Web pages included in the log files themselves, and a *"real" Web directory*, namely the Open Directory Project (ODP). The main difficulty in the latter approach was the association of usage data, i.e. the Web pages, to categories of the directory, given the small proportion of Web pages that are explicitly assigned (manually) to categories of the directory. We approached this problem by an automated page classification method.

Finally, we present OurDMOZ, a system that integrates and implements the various components of the proposed framework. The thesis includes the results of a user evaluation study, which assessed the potential benefits of OurDMOZ and consequently of community Web directories to the end user.

2 Related Work

A number of studies exploit Web directories to achieve a form of personalization. Automatic profile construction is proposed in [5]. The user profiles, linked to categories of the directory are used typically for personalized Web search, while the directory itself is not personalized. The personalization of Web directories is mainly represented by services such as Yahoo! and Excite (www.excite.com), which support the manual selection of interesting categories by the user. An initial approach to automate this process, was the Montage system [1].

Our work differs from the above cited approaches in several aspects. First, instead of using the Web directory for personalization, it personalizes the directory itself. Compared to existing approaches to directory personalization, it focuses on aggregate or collaborative user models such as user communities, rather than

content selection for single user. Furthermore, unlike most existing approaches, it does not require a small set of predefined thematic categories, which could complicate the construction of rich hierarchical models. Finally, the work presented in [2], which is closest to ours is limited to the recommendation of short navigation paths in the ODP hierarchy, rather than the personalization of the whole Web directory structure. Moreover, that method makes the assumption that usage data are collected from the navigation of users within the Web directory. Thus, its applicability to independent services such as a Web portal is questionable.

3 Discovery of Community Web Directories from Web Usage Data

The construction of community Web directories is seen in this thesis as a fully automated process, powered by operational knowledge, in the form of user models that are generated by Web usage mining. User communities are formed using data collected from Web proxies as users browse the Web. The goal is to identify interesting behavioral patterns in the collected usage data and construct community Web directories based on those patterns. The stages of getting from the data to the community Web directories (Figure 1) are summarized below and described in the following sections:

- *Usage Data Preparation*, comprising the collection and cleaning of the usage data.
- *Web Directory Initialization*, providing the characterization of the Web pages included in the usage data, according to the categories of a Web directory. There are two approaches for the characterization of the Web pages. The first approach is to classify them on an existing Web directory, like ODP. The second approach is to map them onto an artificial Web directory constructed from the Web pages themselves using a hierarchical document clustering approach.
- *Community Web Directory Discovery*, being the main process of discovering the user models from data, using machine learning techniques and exploiting these models to build the community Web directories.

3.1 Usage Data Preparation

The usage data that form the basis for the construction of the communities are collected in the access log files of proxy servers, e.g. ISP cache proxy servers. These data record the navigation of the ISP subscribers through the Web. The first stage of data preparation involves data cleaning. The next stage is the identification of individual user sessions. The lack of user registration data or other means of user identification, such as cookies, led us to exploit a simpler definition of user sessions. A user session is defined as a sequence of log entries, i.e., accesses to Web pages by the same IP address, where the time interval between two subsequent entries does not exceed a certain time threshold.

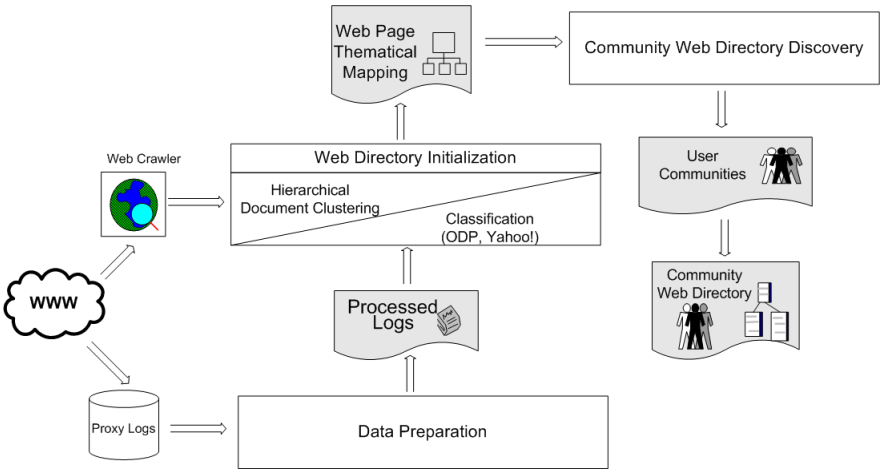


Fig. 1: The process of constructing Community Web directories.

3.2 Web Directory Initialization

The next stage towards the construction of community Web directories is the association of the users' browsing data with the Web directory. In order to personalize the Web directory, we need to "initialize" it with the users' data, i.e., "map" the Web pages onto the Web directory structure. This mapping requires the thematic categorization of the Web pages to the categories of the Web directory, since it is very unlikely to find many of the Web pages that appear in a log file in any directory.

As a first step in the categorization process, a crawler downloads the Web pages included in the usage data and encodes them using the vector space representation, by extracting the most important terms. There are two methodologies that we follow to realize the mapping of Web pages. The first one is based on document clustering and the second is based on document classification. Document clustering is performed following a hierarchical agglomerative approach, whilst document classification requires as a preliminary step, the extraction of keywords from the Web pages included initially in the categories of the Web directory.

The document clustering approach constructs a new Web directory from the usage data themselves. We call this directory an *Artificial Web directory*. The resulting hierarchy is a binary tree, representing clusters of Web pages that form thematic *categories*. This hierarchy corresponds to the initial, non-personalized Web directory, which provides directly a mapping between the Web pages and the categories that the pages are assigned to.

However and despite the similarity of the artificial directory to existing Web directories, there are also notable differences such as the artificial Web directory might not "cover" the semantics of new sessions, due to "overfitting" of the document clustering approach to the initial data. Thus we study the personalization

of a "real" Web directory and in particular the ODP. The main difficulty in this effort was the association of usage data, i.e., the Web pages, to categories of the Web directory, given the small proportion of Web pages that are explicitly assigned (manually) to categories of the directory. We approached this problem by an automated page classification method, which is described below.

Recall from the previous section that each Web page is represented as a vector of terms, we follow a similar vector space representation, for the node-categories of the ODP taxonomy. The Web pages are then classified onto the ODP hierarchy using cosine similarity.

3.3 Objective Category Informativeness

Each community directory includes only a subset of the categories of the initial Web directory, that represent the browsing preferences of the community. However, due to the structure of the Web directory, the community selection process leads to an undesirable side-effect: high-level categories that become leaves aggregate the Web pages of all of their sub-categories that are not in the community model, leading to a "cumulation" of information at the leaves of the reduced directory, which is bound to be overwhelming for its users. Therefore, although the "global" overload problem seems to be tackled well, a "local" overload arises.

To alleviate the "local" information overload problem, we introduce an additional criterion in the discovery of user communities. This criterion incorporates a measure of a-priori informativeness of the categories, which is taken into account when pruning leaf nodes from the Web directory. The inclusion of leaves that satisfy this criterion selectively reduces the generality of the directories, making them reflect more "fine-grained" user interests and resulting in a better distribution of the information that is indexed.

The new criterion is called *Objective Category Informativeness Association, (OCIA)*, and is based on a measure of the *Mutual information, (MI)* of the leaf category l_n, to its parent category c_i. An improved version of MI is the *Symmetrical Uncertainty* (SU) measure, which normalizes MI by dividing by the sum of the entropies of $\mathbf{C}_i$ and the leaf $\mathbf{L}_n$:

$$SU(\mathbf{C}_i, \mathbf{L}_n) = 2.0 \times \left[\frac{(H(\mathbf{L}_n) + H(\mathbf{C}_i) - H(\mathbf{C}_i, \mathbf{L}_n)}{H(\mathbf{C}_i) + H(\mathbf{L}_n)} \right]. \quad (1)$$

The value range of symmetrical uncertainty is [0..1]. Values closer to 0 indicate a weak association between the parent and the leaf category. Thus, leaf categories with a low association to their parents should be included in the community Web directories. OCIA is estimated by normalizing SU further, by the ratio of the number of pages of the leaf to the pages of the parent category, N_{l_n}, N_{c_i} respectively, in order to remove the bias towards leaf categories that contain a large number of Web pages. OCIA is given by the following equation:

$$OCIA(\mathbf{C}_i, \mathbf{L}_n) = \frac{N_{l_n}}{N_{c_i}} \times SU(\mathbf{C}_i, \mathbf{L}_n). \quad (2)$$

OCIA is the criterion that is used to decide whether a leaf node should be included in the community model. Only leaves for which *OCIA* is smaller than a designated *Parent-Children Association Threshold, (PCAT)*, are selected. Thus, the subset $L_i^{'} \subseteq L_i$ of these leaves is defined as:

$$L_i^{'} = \{l_n \in L_i |\ OCIA(\mathbf{C}_i, \mathbf{L}_n) \leq PCAT\}. \quad (3)$$

4 The *Objective Community Directory Miner (OCDM)* Algorithm

In this section we present the three algorithms that have been developed for the construction of community Web directories, as presented in [8], [9], and [10].

5 The *Objective Community Directory Miner (OCDM)* Algorithm

The first machine learning method that we employed for pattern discovery is the *Objective Community Directory Miner* (OCDM), an enhanced version of the cluster mining algorithm [7]. 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.

OCDM enhances cluster mining so as to take into account the hierarchy of topic categories. This is achieved by updating the weights of the vertices and the nodes in the graph. Each category is mapped onto a set of categories, corresponding to its parent and grandparents in the thematic hierarchy. The frequency of each of these categories is increased by the frequency of the initial child category. The underlying assumption for the update of the weights is that if a certain category exists in the data, then its parent categories should also be examined for the construction of the community model. In this manner, even if a category (or a pair of categories) have a low occurrence (co-occurrence) frequency, their parents may have a sufficiently high frequency to be included in a community model. This enhancement allows the algorithm to start from a particular category and ascend the topic hierarchy accordingly. The result is the construction of a topic tree, even if only a few nodes of the tree exist in the usage data.

The connectivity of the resulting 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 simplified to an unweighted one. Finally all maximal cliques of the unweighted graph are generated, each one corresponding to a community model. Θ_r. Then, for each community model, Θ_r i.e., clique, we examine the informativeness of the leaf categories of the initial Web directory that are not in the

clique. Using the OCIA criterion, we compare each such leaf against its closest ancestor that is included in the Θ_r.

6 The *Objective Probabilistic Directory Miner (OPDM)* Algorithm

In the *OCDM* algorithm discussed above, the constructed patterns are based solely on the "observable" behavior of the users, as this is recorded in the usage data. Generally, users' interests and motives are less explicit. We are considering that the user's choices are motivated by a number of latent factors that correspond to these subsets. These factors are responsible for the associations between users. The advantage of this approach is that it allows us to describe more effectively the multi-dimensional characteristics of user interests.

A powerful statistical methodology for identifying latent factors in data is Probabilistic Latent Semantic Analysis (PLSA) [4]. 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 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 P(u_i): the a priori probability of a user session u_i, P($z_k|u_i$): the conditional probability of the latent factor z_k being associated with the user session u_i and 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 P(u_i), selecting a latent factor z_k with probability P($z_k|u_i$) and selecting a category c_j with probability 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:

$$P(u_i, c_j) = P(u_i)P(c_j|u_i) = P(u_i)\sum_k P(c_j|z_k)P(z_k|u_i). \quad (4)$$

Using Bayes's theorem we obtain the equivalent equation:

$$P(u_i, c_j) = \sum_k P(z_k)P(u_i|z_k)P(c_j|z_k). \quad (5)$$

Equation 5 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.

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:

$$P(c_j|z_k) \geq LFAP. \tag{6}$$

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.The *LFAP* criterion is combined with the *OCIA* criterion an the initial Web directory is traversed and each category-node which does not satisfy the LFAP and the PCAT thresholds is pruned. 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.

7 Objective Clustering and Probabilistic Directory Miner (OCPDM)

In an attempt to combine the advantages of clustering and probabilistic modeling, we 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 clustering algorithm, for the creation of an initial set of communities. This approach differs from OCDM clustering, as it produces non-overlapping clusters, i.e., each category belongs to a single cluster. However, as we have explained above for PLSA, the explicit modeling of latent factors is considered advantageous. Thus, we assume that in addition to the k-means clusters, further hidden associations exist in the data, i.e., sub-communities inside each cluster that are not directly observable. To discover this hidden knowledge, 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 a combination of observable and latent associations in the data, and potentially allow us to better model the interests of users.Thus, the new algorithm *Objective Clustering and Probabilistic Directory Miner (OCPDM)* invokes OPDM for each of the K clusters. The categories of the cluster, on which each latent factor has the maximum impact, are selected using the LFAP threshold.

8 Community Web Directory Refinement

The result of the pattern discovery methods presented in Sections 5 to 7 is a set of hierarchies that correspond to the community Web directories, i.e., to a prototypical model for each community, which is representative of the participating

users. The construction of the directory is based on the selection of the categories by each algorithm and their mapping onto the original Web directory. However, the construction of useful community Web directories needs to go beyond the selection of categories by the pattern discovery algorithms. Further processing is required to improve the structure of the directory and this is achieved by the following operators: *Shortcut Operator*. If a category has a single descendant node, then a "shortcut" is created from the parent to the leaf node. *Absorb*. This operator applies to categories that are leaves in the community Web directory, but not in the initial Web directory. Since all of their descendant categories are excluded from the community Web directory, all the Web pages of the eliminated descendant leaves are absorbed by the shrinking category. This operator ensures that no information is lost, even when the "original" leaves are not included in the community Web directory. In the case though, where at least one descendant leaf is included in the community models, this operator is not applied, assuming that the users are not interested in the other leaf categories.

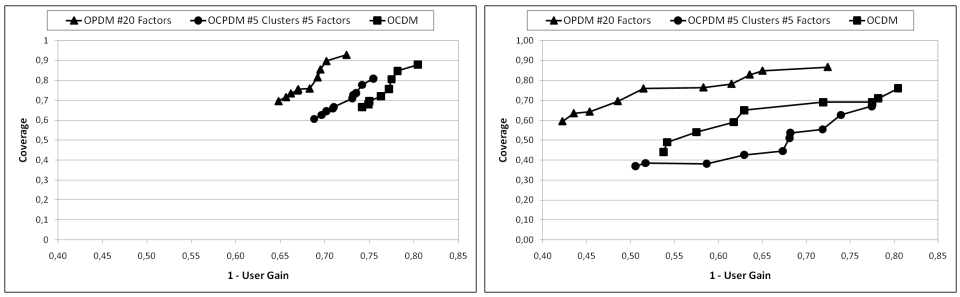
9 Experimental Setup

The evaluation process assessed the effectiveness of the algorithms on categories of the artificial Web directory and on the ODP categories. The evaluation employs mainly two measures: *Coverage* and *User Gain*. Coverage corresponds to the predictiveness of our model, i.e., the number of target Web pages that are actually *covered* by the session-specific community directories. On the other hand, user gain is an estimate of the actual gain that a user would have by following the community Web directory, instead of the initial non-personalized Web directory to get to the desired Web page.

An interesting measure of the effectiveness of our approach is the trade-off between coverage and user gain. The usual choice for such a trade-off measure is the use of Receiver Operating Characteristics (ROC) curves and 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. In Figure 2, we present the trade-off curves for the algorithms for the artificial Web directory and the ODP. From this figure we conclude that the user seems to be benefiting from the personalization, both in terms of Coverage and User Gain. Regarding the comparison of the three directory methods, OPDM clearly outperforms the other two in both directories. The performance of the "hybrid" OCPDM method is lower in terms of coverage, due to the non-overlapping nature of the k-means algorithm.

Comparing the behavior of the methods on the two different directories, we have obtained a higher user gain at a smaller cost in coverage in the ODP directory. In particular, for the ODP directory and for the OPDM algorithm which exhibits the best performance overall, we obtain user gain around 0,50, maintaining coverage at the level of 0,75. For the artificial Web directory, the same algorithm results in a coverage of almost 0,90 but with a user gain value of 0,27. This level of user gain in the ODP directory, is attainable, due to the

size and the generic nature of the ODP. Thus, the use of a real directory has revealed the power of personalizing the directory.



(a) Artificial Web directory. (b) ODP Web directory.

Fig. 2: Web Directory Coverage-User Gain Trade-Off with OCIA (Average PCAT Thresholds).

The results presented in this section provide a detailed picture of the benefits of our approach to personalizing Web directories. Regarding the various discovery methods that we tested, the "pure" PLSA technique (OPDM) outperforms the simple clustering algorithm (OCDM) and the combination of clustering and PLSA (OCPDM).

10 OurDMOZ

In this section, we present *OurDMOZ*, a system that implements and integrates the various components of the proposed methodology. In particular, OurDMOZ collects and processes usage data, maps the data onto the Web directory, uses machine learning techniques to extract the community models and finally builds the community Web directories. The main contribution of OurDMOZ is that it offers, through its Web application, a personalized view of ODP and consequently a personalized view of the Web. In OurDMOZ, a user can join a particular community either by specifying her preferences, or by using the system for some time and letting it decide on the most suitable community models. Thus, there is no requirement for personal information, or other private data, to be provided to the system.

We perform an actual user evaluation, where OurDMOZ is given to a set of users who interact with the system and use its personalization functionalities. One of the scenarios followed considered the fill-in of a small questionnaire. The questions included in the questionnaire were answered in a seven-level Likert scale from "Strongly disagree" to "Strongly agree". The results of the users' responses to this questionnaire are presented in Figure 3 and show that the users found OurDMOZ easy and helpful.

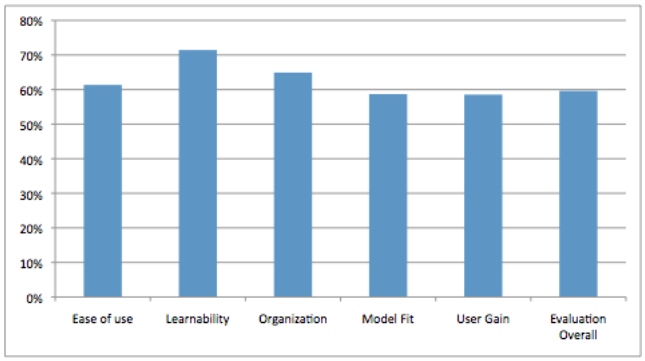


Fig. 3: Questionnaire Results.

11 Conclusions

Community Web directories exemplify a new type of Web personalization, beyond common Web personalization functions such as Web page recommendations and adaptive Web services. In this thesis, we present a 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.

More specifically, the thesis has contributed in the following areas:

1. Presentation of a roadmap of the Web personalization.
2. Analysis of the main stages of the Web usage mining process and their relation to the Web personalization.
3. Extension of the Web usage mining approach to a much larger portion of the Web, through the analysis of usage data collected by proxy servers. These data correspond to traffic throughout the Web and they are not restricted within the context of a single Web site. The proposed methodology addresses the high dimensionality of the problem, through the classification of individual Web pages onto the categories of the directory.
4. Proposal of three novel pattern discovery algorithms based on clustering and probabilistic approaches (PLSA) for the extraction of community models from the usage data. These methods take into account, not only the browsing behavior of users, but also the structure and the distribution of information within a Web directory.
5. Proposal of a methodology for converting of community models to community Web directories.
6. Development of a complete system, named OurDMOZ, that constructs community Web directories and exploits them for offering personalization functionalities to Web users. These functionalities include not only a customized view of the Web directory, but also they offer recommendation services to Web users.

We hope that this thesis will contribute to the move from Web site personalization, to real Web personalization. In this direction, several issues remain open. These issues are related with all the stages of the community Web directory construction process, such as the exploitation of new techniques that might offer a more accurate view of the users' behavior, whilst respecting user's privacy.

References

1. C. R. Anderson and E. Horvitz. Web montage: A dynamic personalized start page. In *11th WWW Conference*, May 2002.
2. T. Dalamagas, P. Bouros, T. Galanis, M. Eirinaki, and T. Sellis. Mining user navigation patterns for personalizing topic directories. In *9th annual ACM international workshop on Web information and data management*, pages 81–88, 2007.
3. J. A. Hartigan. *Clustering Algorithms*. A Wiley-Interscience Publication, New York: Wiley, 1975, 1975.
4. T. Hofmann. Probabilistic latent semantic analysis. In *UAI*, 1999.
5. T. Oishi, K. Yoshiaki, M. Tsunenori, H. Ryuzo, F. Hiroshi, and M. Koshimura. Personalized search using odp-based user profiles created from user bookmark. In *10th Pacific Rim International Conference on Artificial Intelligence*, pages 839–848, 2008.
6. G. Paliouras, C. Papatheodorou, V. Karkaletsis, and C. D. Spyropoulos. Discovering user communities on the internet using unsupervised machine learning techniques. *Interacting with Computers Journal*, 14(6):761–791, 2002.
7. M. Perkowitz and O. Etzioni. Adaptive web sites: automatically synthesizing web pages. In *Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*, pages 727–732. American Association for Artificial Intelligence, 1998.
8. D. Pierrakos and G. Paliouras. Exploiting probabilistic latent information for the construction of community web directories. In *User Modeling*, pages 89–98, 2005.
9. D. Pierrakos and G. Paliouras. Personalizing web directories with the aid of web usage data. *IEEE Transactions on Knowledge and Data Engineering*, 22:1331–1344, 2010.
10. D. Pierrakos, G. Paliouras, C. Papatheodorou, V. Karkaletsis, and M. Dikaiakos. Web community directories: A new approach to web personalization. In B. B. et al., editor, *Web Mining: From Web to Semantic Web, EMWF 2003*, volume 3209 of *LNCS*, pages 113–129. Springer, 2004.
11. 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.