Personalizing Web Directories with the Aid of Web Usage Data

Dimitrios Pierrakos, *Member*, *IEEE*, and Georgios Paliouras

Abstract—This paper presents a knowledge discovery framework for the construction of Community Web Directories, a concept that we introduced in our recent work, applying personalization to Web directories. In this context, the Web directory is viewed as a thematic hierarchy and personalization is realized by constructing user community models on the basis of usage data. 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 thematic diversity. For modeling the user communities, we introduce a novel methodology that combines the users' browsing behavior with thematic information from the Web directories. Following this methodology, we enhance the clustering and probabilistic approaches presented in previous work and also present a new algorithm that combines these two approaches. The resulting community models take the form of Community Web Directories. The proposed personalization methodology is evaluated both on a specialized artificial and a general-purpose Web directory, indicating its potential value to the Web user. The experiments also assess the effectiveness of the different machine learning techniques on the task.

Index Terms—Machine learning, Web mining, clustering, personalization.

1 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 canceling 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. Thus, personalization can be of particular benefit to large generic directories such as ODP, or Yahoo!. Personalized versions of these directories can also be employed by various services on the Web, such as Web portals, in order to offer their subscribers a personalized view of the Web. 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

• *The authors are with the Institute of Informatics and Telecommunications, NCSR "Demokritos," PO Box 60228, Ag. Paraskevi, 15310 Athens, Greece. E-mail: {dpie, paliourg}@iit.demokritos.gr.*

Manuscript received 14 Jan. 2009; revised 12 June 2009; accepted 8 July 2009; published online 17 July 2009.
Recommended for acceptance by D. Cook.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2009-01-0017.
Digital Object Identifier no. 10.1109/TKDE.2009.173.

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.

In previous work [8], [9], we have proposed the use of machine learning techniques for modeling the user communities based on clustering and probabilistic modeling. The experimental results illustrated the benefits of constructing personalized Web directories with the proposed methods. However, these personalized directories suffer from what we call the "local overload" problem. This problem is a side effect of pruning 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, upward in the hierarchy. This leads to increased information density of some leaf nodes of the personalized directory. Addressing this issue, in this paper, we introduce a novel methodology that combines usage data and thematic information from Web directories. This methodology enhances the proposed methods, using an estimate of the amount of thematic information included a priori in the categories of the Web directory. This information is exploited by the introduction of a new criterion in our methodology, leading to the construction of more fine-grained community Web directories, and thus, to an efficient tackling of the "local overload" problem. Besides the improvement of these methods, a new one is presented that combines clustering and probabilistic modeling by applying Probabilistic Latent Semantic Analysis (PLSA) to the results derived by a clustering algorithm.

In our previous work, we applied personalization on an artificial Web directory that was constructed from the Web pages included in the log files themselves, using document clustering. In this paper, we study the personalization of a "real" Web directory, namely, the ODP. The main difficulty in this effort 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 developed in [10]. In order to provide a link to our older experiments, we compare the results of the personalized ODP to the artificial Web directory.

The rest of this paper is organized as follows: Section 2 presents existing approaches to Web personalization with usage mining methods, as well as some earlier efforts on the personalization of Web directories. Section 3 describes the methodology for the discovery of community Web directories, while Section 4 explains the proposed solution to the local overload problem mentioned above. Section 5 presents in detail the discovery algorithms and the postprocessing of the community Web directories in order to become more operative and user-friendly. Section 6 provides the experimental results, and Section 7 summarizes the main conclusions of this work.

2 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. The resulting "task-sequences" were used for statistical analysis of user behavior, such as finding the most frequent tasks, or for generating recommendations. In [16], PLSA was exploited to build user profiles, where each profile consists of the distribution of Web pages over the set of the latent factors. Subsequently, user profiles supported personalized Web search. More recent work [17] used PLSA to cluster users in legitimate and malicious ("shilling") groups.

On the other hand, a number of studies exploit Web directories to achieve a form of personalization. In [18], users build their profiles by specifying a set of categories from the ODP hierarchy. Automatic profile construction is proposed in [19], [20], [21], and [22]. 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 [23], which was 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 related technique for mobile portal personalization was presented in [24], where the portal structure was adapted to the preferences of users. In [25], a Web directory was used as a "reference" ontology and the Web pages navigated by a user were mapped onto this ontology using document classification techniques, thus resulting in a personalized ontology. Finally, in recent work [26], the similarity between users, based on navigation data within the ODP, was used to create clusters of ODP categories. These clusters were further exploited to recommend shortcuts within the Web directory.

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 [26], 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.

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.

3 Discovery of Community Web Directories from Web Usage Data

The construction of community Web directories is a fully automated process, resulting in operational personalization knowledge, in the form of user models. 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 process of getting from the data to the community Web directories is summarized below:

Usage Data Preparation comprises the collection and cleaning of the usage data, as well as the identification of user sessions.

Web Directory Initialization provides the characterization of the Web pages included in the usage data, according to the categories of a Web directory. We compare two different approaches for the characterization of the Web pages. The first approach organizes Web pages into an artificial Web directory using hierarchical document clustering. The second approach classifies them onto an existing Web directory, like ODP.

Community Web Directory Discovery is the main process of discovering the user models from data, using machine learning techniques and exploiting these models to build the community Web directories.

The first two stages result in the construction of the required structures for the discovery of community Web directories. These stages are presented in Sections 3.1 and 3.2. An initial discussion of the third stage is presented in Section 3.3, while more details are provided in the sections that follow.

3.1 Usage Data Preparation

The usage data that form the basis for the construction of community Web directories are collected in the access log files of ISP cache proxy servers. These data record the navigation of the subscribers through the Web, and hence, they are usually diverse and voluminous. The outgoing traffic is much higher than the usual incoming traffic of a Web site and the Web pages more diverse semantically. The task of usage data preparation, detailed in [8], [9], is to assemble these data into a consistent, integrated, and concise view. The next stage is the identification of individual user sessions. The fact that we are focusing on the discovery of behavioral patterns in the data, rather than individual users, allowed us to overcome the lack of user registration data or other means of user identification, such as cookies, and led us to exploit a simple kind of user session. 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. More formally:

Definition 1 (User session). *Let* $\mathtt{P} = (p_1, p_2, \ldots, p_u)$ *is the sequence of Web pages accessed from a certain IP between* t_1 *and* t_u. *Then, a* user session $v(t_1, t_f), t_f \leq t_u$, is defined as: $v(t_1, t_f) = (p_1, p_2, \ldots, p_f) : (\Delta t = t_j - t_{j-1} \leq \delta,\ 1 < j \leq f) \wedge (f = u \vee t_{f+1} - t_f > \delta)$, *where* δ is a predefined time threshold.

User sessions are thus extracted from access logs as follows: 1) group the log entries by date and IP address; 2) select a threshold within which two consecutive records from the same IP address can be considered to belong to the same user session; and 3) group the Web pages accessed by the same IP address respecting the selected threshold to form sessions.

3.2 Web Directory Initialization

The next stage toward the construction of community Web directories is the association of the users' browsing data with the Web directory. Generally, a Web directory can be defined as follows:

Definition 2 (Web directory). *A* Web Directory *is a directed acyclic graph* $G = (C, E)$, *where* C *is a finite set of nodes and* E *a set of directed edges connecting the nodes. The set* C *of nodes corresponds to the thematic Web directory categories.*

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. Although Web directories such as ODP and Yahoo! include a number of Web pages manually linked to the nodes of the hierarchy, their coverage of the Web is very small. Thus, 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 process of Web directory initialization, a crawler downloads the Web pages included in the usage data and encodes them using the vector space representation [27] of their contents, i.e., the terms of the Web pages. This representation allows us to categorize the Web pages onto the thematic taxonomy of the Web directory following one of the two approaches discussed below.

3.2.1 Artificial Web Directory

An artificial Web directory is constructed from the usage data themselves. In particular, by exploiting the vector space representation of the Web pages, a taxonomy is built using a hierarchical agglomerative approach (e.g., [28]) for document clustering. The resulting hierarchy is a binary tree, representing clusters of Web pages that form thematic *categories*. This hierarchy corresponds to the initial Web directory, which provides directly a mapping between the Web pages and the categories that the pages are assigned to. Details of how the artificial Web directory is constructed can be found in [8], [9].

3.2.2 "Real" Web Directory

The structure of the artificial directory is similar to that of existing Web directories since each node can be considered a category of the directory containing a set of semantically similar Web pages. Nevertheless, there are also notable differences between a "real" Web directory and the artificially constructed one. One such difference is that the initial directory (without personalization) is already strongly focused on the usage data, rather than being a general resource. As a result, the potential benefits of personalization are rather limited. A more practical difference concerns the artificiality of thematic categories (document clusters). The artificial directory is a binary tree, where each node clusters exactly two subnodes. In addition, the number of nodes has been chosen using only statistical properties of Web page contents. Finally, the artificial Web directory might not "cover" the semantics of new sessions due to "overfitting" of the document clustering approach on the initial data. These observations motivated us to 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 developed in [10].

More formally, let $P = \{p_1, p_2, \ldots, p_n\}$, the set of n Web pages in the usage data and $T = \{\tau_1, \tau_2, \ldots, \tau_m\}$, the set of m terms, in the n Web pages. Each Web page p_i is represented as a vector $\vec{p_i} = \{w_{i1}, w_{i2}, \ldots, w_{im}\}$, where w_{ik} is the weight of term τ_k. A vector space representation is also employed for the node categories of the ODP taxonomy. More formally, let $C = \{c_1, c_2, \ldots, c_M\}$, the set of M ODP categories "mapped" onto the same space of terms T as the Web pages. Each ODP category c_j is represented as a vector $\vec{c_j} = \{q_{j1}, q_{j2}, \ldots, q_{jm}\}$, where q_{jk} is the weight of the category term. The weights followed in both cases are calculated using the *Term Frequency-Inverse Document Frequency* (TF-IDF) scheme. The Web pages are classified onto the ODP hierarchy using cosine similarity, i.e.,

$$sim(p_i, c_j) = \frac{\vec{p_i} \cdot \vec{c_j}}{|\vec{p_i}| \cdot |\vec{c_j}|} = \frac{\sum_{k=1}^{k=m} w_{ik} q_{jk}}{\sqrt{\sum_{k=1}^{k=m} w_{ik}^2} \sqrt{\sum_{k=1}^{k=m} q_{jk}^2}}. \quad (1)$$

Note that a Web page might belong to more than one ODP categories. For reasons of simplicity, we assign the Web page to a single category, maximizing cosine similarity. Furthermore, the assumption is made that there is a sufficiently suitable ODP category for every Web page and no pages remain unclassified. This assumption is an interesting direction for further research.

The hierarchical classification of Web pages requires us to redefine the notion of user session in order to work with the categories in the Web directory rather than the Web pages themselves. In our approach, pages are mapped onto thematic categories of the hierarchy, and therefore, a user session is translated into a sequence of categories. We define the user sessions which result from this mapping as *thematic user sessions* since they do not contain the Web pages themselves, but rather their thematic representation.

Definition 3 (Thematic user session). *Let* $\mathtt{P} = (p_1, p_2, \ldots, p_u)$ *is the sequence of Web pages accessed from a certain IP between* t_1 *and* t_u *which is mapped onto the sequence of categories* $(c_1, c_2, \ldots, c_f)$*, where* $f \leq u$*. The* thematic *counterpart* $u(t_1, t_f)$ *of user session* $v(t_1, t_f)$ *(Definition 1) is defined as:* $u(t_1, t_f) = (c_1, c_2, \ldots, c_f) \vee \{c_i : c_i \in C \wedge f \leq u\}$ *in the case of the artificial Web directory, and* $u(t_1, t_f) = (c_1, c_2, \ldots, c_f) \vee \{(c_i : c_i \in C \wedge c_i = \arg\max_i(sim(p_u, c_i)\}$ *in the case of the ODP, where* sim *is a similarity function like cosine or another classifier.*

3.3 Community Web Directory Discovery

Having determined the mapping and the associations between Web pages, user sessions, and Web directory categories, we employ unsupervised learning to discover patterns of interest in the thematic user sessions. In our recent work, we employed two methods for the discovery of community Web directories. In [8], we presented an extension of the cluster mining algorithm, named *Community Directory Miner (CDM)*, while in [9], we presented an approach based on the discovery of latent semantics using PLSA. These algorithms are used to extract a subset of the categories of the initial Web directory that correspond to the community models, i.e., usage patterns that occur in data and represent the browsing preferences of community members. Each community model Θ is subsequently exploited to construct the community Web directory. The general process of discovering community Web directories can be seen as a construction of the subgraph G' of the Web directory G which corresponds to the community Web directory. More formally:

Definition 4 (Community Web directory). *Let* $\mathrm{G} = (\mathrm{C}, \mathrm{E})$ *be the Web directory as per Definition 2. Let also* Θ *be a set of community model categories. We define the* Community Web directory *as the subgraph* $G' = (C', E')$ *of* G *with the following properties:*

- $C' \subseteq \Theta \subseteq C$.
- *An edge* $e = (c_a, c_b) \in E'$ *from node* c_a *to node* c_b *is created in* G'*, if* $\{c_a, c_b\} \subseteq \Theta$ *and one of the following conditions is met:*
  - $\exists (c_a, c_b) \in E$
  - $\exists (c_a, c_{a+1}, \ldots, c_{b-1}, c_b)$ *a path in* E, $\wedge \{c_{a+1}, \ldots, c_{b-1}\} \bigcap \Theta = \emptyset$.

The community directories include only a subset of the categories of the initial Web directory that represent the browsing preferences of the community. Therefore, the categories selected in this manner, i.e., through pattern discovery algorithms, reveal what we define as the *commu-*

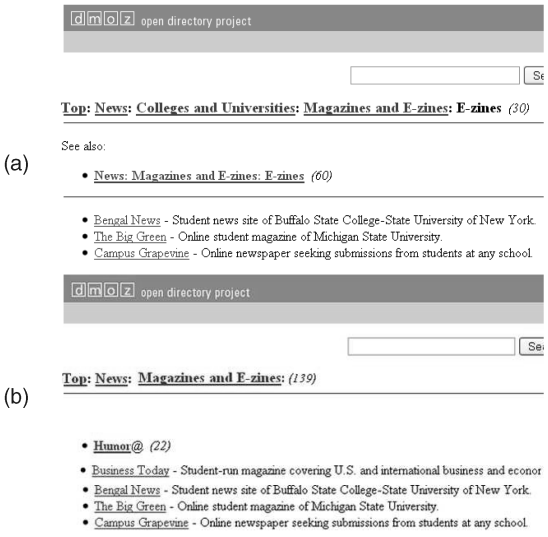


Fig. 1. ODP directory path and ODP community directory path.

nity informativeness of each category. Community informativeness is a measure of the importance of a particular category for the community, based on the frequency of the users' visits to the category. The exact measure of community informativeness varies according to the particular discovery algorithm that is used. This approach results in a substantial reduction of the initial Web directory, which is now personalized to the interests of the community.

As a working example, in Fig. 1a, we focus on a particular path of the ODP directory to the leaf category "Top/News/Colleges_and_Universities/Magazines_ and_E-zines/E-zines." Its personalized view is depicted in Fig. 1b, where the categories that are not included in the community model are also not included in the community Web directory. The shorter path "Top/News/Magazines_ and_E-zines" allows the members of the particular community to arrive more directly at the required information.

However, the approach that we proposed in our past work had an undesirable side effect: high-level categories that become leaves aggregate the Web pages of all of their subcategories 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. This situation is depicted in Fig. 1b, where the pages are cumulated in the category "Top/News/Magazines_and_E-zines." Therefore, although the "global" overload problem seems to be tackled well, a "local"overload arises. The "local" overload problem motivated us to enhance our methodology for the discovery of community models, as we present in the following section.

4 Objective Category Informativeness

To alleviate the "local" information overload problem discussed above, we introduce an additional criterion in the discovery of user communities. This criterion incorporates a measure of a priori informativeness of 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 proposed measure of objective informativeness is based on the entropy of the category. More formally, let $c_i \in C$ a category of the Web directory. This category corresponds to a node of the Web directory and can be represented by a Boolean random variable $\mathbf{C}_i$.[1] This variable is *true* for Web pages of the category and *false* for the remaining Web pages in the directory. Thus, the probability distribution of the variable $p(\mathbf{C}_i)$ is estimated across the number of Web pages in the directory as:

$$p(\mathbf{C}_i = true) = \frac{\#\ pages\ in\ the\ category}{total\ number\ of\ pages}.$$

The a priori amount of information in this category before examining the users' browsing behavior can be estimated by its entropy $H(\mathbf{C}_i)$, i.e.,

$$H(\mathbf{C}_i) = \sum_{k \in \{true, false\}} p(\mathbf{C}_i = k) \log p(\mathbf{C}_i = k). \quad (2)$$

Exploiting further the above reasoning, we proceed to the following definition:

Definition 5 (Objective category informativeness). *We define the* Objective Category Informativeness (OCI) *of a category c_j of the Web directory, as the conditional entropy of the category given its parent c_i, i.e.,*

$$OCI(c_j) = H(\mathbf{C}_j|\mathbf{C}_i). \quad (3)$$

This approach allows us to measure the a priori informativeness of the Web directory categories in order to address the local overload problem. The rationale behind measuring informativeness through the conditional entropy of each category is that it provides a rather "objective" measure of the importance of the categories in the Web directory. It is objective in the sense that a category with a large number of Web pages might convey the same amount of information as a category with a small number of Web pages. This fact allows us to directly compare and examine the dependance between the categories of the Web directory. In particular, using (3), we introduce a new selection criterion which is based on the OCI of a category compared to that of its children categories. This criterion is combined with the selection criterion of the knowledge discovery method when applied to the leaves of the Web directory. More specifically, in evaluating whether a leaf l_n should be included in the community Web directory, we measure how "similar" in terms of OCI the leaf category is to its parent. Leaf categories with OCI close to that of their parents are excluded from the model. Leaves that are less similar, i.e., have a weak association to their parents introduce significant variation, and thus, they add to the community models the required specialization.[2] The new criterion is called *Objective Category Informativeness Association (OCIA)*, and is based on a measure of the mutual dependence of the leaf category l_n, to its parent category c_i. *Mutual Information* (MI) is defined as

1. We follow this notation for the rest of the paper, i.e., bold letters represent random variables.
2. By definition, leaves are less general than their parents.

$$MI(\mathbf{L}_n; \mathbf{C}_i) = H(\mathbf{L}_n) + H(\mathbf{C}_i) - H(\mathbf{C}_i, \mathbf{L}_n). \quad (4)$$

MI provides an indication of how strong the association between the parent category c_i and the leaf category l_n is. An improved version of MI is the *Symmetrical Uncertainty* (SU) measure [29], which normalizes MI by dividing the sum of the entropies of $\mathbf{C}_i$ and the leaf $\mathbf{L}_n$:

$$SU(\mathbf{C}_i, \mathbf{L}_n) = 2.0 \times \left[\frac{(H(\mathbf{L}_n) + H(\mathbf{C}_i) - H(\mathbf{C}_i, \mathbf{L}_n))}{H(\mathbf{C}_i) + H(\mathbf{L}_n)}\right]. \quad (5)$$

The value range of symmetrical uncertainty is [0..1]. Values closer to 0 indicate a weak association between the parent and the leaf category. Following the rationale explained above, 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 toward leaf categories that contain a large number of Web pages.

Definition 6 (Objective category informativeness association). *We define the OCIA as the measure of similarity between a category c_i and its descendant leaf node l_n in the Web directory, given by the following equation:*

$$OCIA(\mathbf{C}_i, \mathbf{L}_n) = \frac{N_{l_n}}{N_{c_i}} \times SU(\mathbf{C}_i, \mathbf{L}_n). \quad (6)$$

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

As a concrete example, we consider the situation depicted in Fig. 2. In Fig. 2a, we present a snapshot of the community Web directory extracted without the use of the OCIA criterion. In this figure, we assume that all leaf categories together with their direct ancestor (Publications), i.e., grayed-out hyperlinks, are not considered interesting according to the community informativeness measure. Thus, the category c_{ML} = "Top/.../Machine_Learning" will become the leaf category of the community Web directory, potentially causing a local overload problem due to the cumulation of Web pages. Leaves l_{BIB} = "Top/.../Bibliographies/," l_{BKS} = "Top/.../Books/," l_{JNL} = "Top/.../Journals/," and l_{PAP} = "Top/.../Paper_Repositories/" having values OCIA(c_{ML}, l_{BIB}) = 0.4, OCIA(c_{ML}, l_{BKS}) = 0.2, OCIA(c_{ML}, l_{JNL}) = 0.1, and OCIA(c_{ML}, l_{PAP}) = 0.3, respectively, are examined for selection, as children of c_{ML} = "Top/.../Machine_Learning." If a *PCAT* threshold of 0.3 is introduced, leaves l_{BKS} and l_{JNL} (dark hyperlinks in Fig. 2a) will not be pruned, while leaves l_{BIB} and l_{PAP} (grayed-out hyperlinks in Fig. 2b) are assumed to be closer to c_{ML} and will not be included in the final community Web directory. In this manner, the OCIA criterion is incorporated in the community discovery algorithms in order to include a number of leaves that would, otherwise, be pruned from the Web directories. Details of the enhanced versions of the algorithms are presented in the following section. In addition, we propose a new method that combines clustering and PLSA, by initially applying a common clustering algorithm, like *k-means*, and subsequently discovering latent information in the derived clusters.

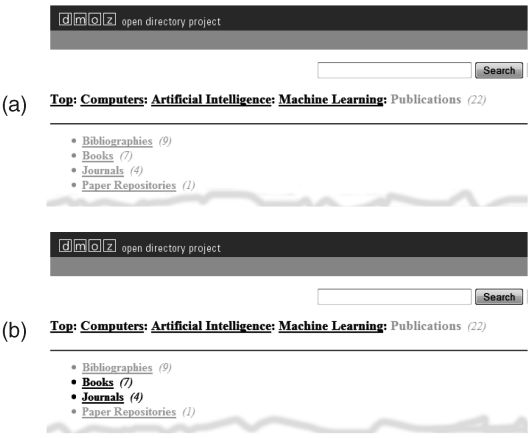


Fig. 2. An example of the OCIA criterion.

5 Community Web Directory Discovery Algorithms

In this section, 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, as per Definition 3. From this definition, 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 u_i is generally smaller than its simple counterpart v_i that contains Web pages. The removal of duplicates leads to the *thematic session set* which is defined as follows:

Definition 7 (Thematic session set). *Let $u(t_1, t_f)$ is a thematic user session, then its* session set $\bar{u}(t_1, t_f) = \{l_i, l_i \in C\}$ *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. This extended relation is formally defined below.

Definition 8 (Thematic session tree). *Let the set of thematic session sets $\bar{U} = \{\bar{u}_1, \bar{u}_2, \ldots, \bar{u}_\nu\}$, and the set of Web directory categories C, as per Definition 2. We define a Thematic Session Tree as the binary relation $\Re = (\bar{U}, C)$, where each $(\bar{u}_\nu, c_i)$ pair represents the access of a certain category $c_i \in C$ or one of the leaves L_i of its subtree, during the session set $\bar{u}_\nu$. The* indicator function *of the relation $\Re$, $\mathbf{1}_\Re : \Re \rightarrow \{0, 1\}$, is defined as*

$$\mathbf{1}_\Re(\bar{u}_\nu, c_i) = \begin{cases} 1, & if\ (\bar{u}_\nu, c_i) \in \Re, \\ 0, & otherwise. \end{cases} \quad (7)$$

For example, if the leaf category "Top/Science/Social_Sciences/Education" is included in the thematic session set $\bar{u}_\nu$, then the following pairs are in $\Re$: ($\bar{u}_\nu$,"Top/Science"), ($\bar{u}_\nu$,"Top/Science/Social_Sciences"), and ($\bar{u}_\nu$,"Top/Science/Social_Sciences/Education"). For reasons of simplicity, the proposed formalization ignores the frequency of a particular category in a session, which may be an additional indicator of user interest. The formalization and subsequent knowledge discovery methods could be extended to take frequency into account.

5.1 Objective Community Directory Miner (OCDM)

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 *Objective-Community Directory Miner* (OCDM). Similar to CDM, OCDM 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 OCDM group together characteristic features of the user. Each clique discovered by OCDM is thus already a community model, i.e., a set of interesting categories. The OCDM algorithm incorporating the OCIA criterion can be summarized in the following steps and presented in Algorithm 1 in the Appendix:

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*. We exploit the thematic session tree $\Re$ as per Definition 8 calculating the weight w_i for the vertex corresponding to category c_i and the weight of the edge w_{ij} as follows:

$$w_i = \frac{\sum_{k=1}^{\nu} \mathbf{1}_\Re(\bar{u}_k, c_i)}{\nu}, \quad (8)$$

$$w_{ij} = \frac{\sum_{k=1}^{\nu} \mathbf{1}_\Re(\bar{u}_k, c_i) \cdot \mathbf{1}_\Re(\bar{u}_k, c_j)}{\nu}, \quad (9)$$

where $(\bar{u}_\kappa, c_i), (\bar{u}_\kappa, c_j) \in \Re$ and ν are the total number of the thematic session sets.

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 [30].* One can trivially show that each clique generated by CDM contains the complete path of each of the categories, i.e., the category itself as well as its ancestral categories. Thus, each clique corresponds to a community Web directory.

Step 4. *Select informative leaves for the hierarchy*. For each community model, Θ_r, i.e., clique, we search for the leaves in the initial Web directory that are not in the clique. Then, we apply the OCIA criterion to each such leaf against its closest ancestor that is included in the Θ_r. More formally, let L be the set of leaf categories of the Web directory, $l_n \in L \wedge l_n \notin \Theta_r$. if $c_i \in \Theta_r$ the closest ancestor of l_n in Θ_r, then l_n is added to Θ_r iff $OCIA(\mathbf{C}_i, \mathbf{L}_n) \leq PCAT$, where *OCIA* is the measure defined in Section 4.

5.2 Objective Probabilistic Directory Miner (OPDM)

In the *OCDM* 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. 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.

A powerful statistical methodology for identifying latent factors in data is PLSA [31]. Similar to the approach that we followed for the OCDM algorithm, we recall the relation $\Re = (\bar{U}, C)$, with the pair $(\bar{u}_i, c_j) \in \Re$ representing the access of category c_j or one of its leaf subcategories during the

thematic session set $\bar{u}_i$. The PLSA model is based on the assumption that there exists a set $Z = \{z_1, z_2, \ldots, z_k\}$ of latent factors such that each instance $(\bar{u}_i, c_j) \in \Re$, i.e., each observation of a certain category inside a thematic session tree, is related to a latent factor $z_k \in Z$. To further formalize our assumption, we define the following probabilities: $p(\bar{\mathbf{U}}_i)$, the a priori probability of the thematic session set $\bar{u}_i$, i.e., the number of times the session appears in the usage data; $p(\mathbf{Z}_k|\bar{\mathbf{U}}_i)$, the conditional probability of the latent factor z_k motivating the observation of session $\bar{u}_i$ and $p(\mathbf{C}_j|\mathbf{Z}_k)$, the conditional probability of category c_j being accessed, given the latent factor z_k. Using these definitions, we can describe a probabilistic model for generating the categories "observed" in sessions as follows: select a thematic session set with probability $p(\bar{\mathbf{U}}_i)$, select a latent factor z_k with probability $p(\mathbf{Z}_k|\bar{\mathbf{U}}_i)$, and select a category c_j with probability $p(\mathbf{C}_j|\mathbf{Z}_k)$. The results of the above process allow us to estimate the probability of observing a particular (session-category) pair ($\bar{u}_i$,c_j), using joint probabilities and Bayes's theorem as follows:

$$p(\bar{\mathbf{U}}_i, \mathbf{C}_j) = \sum_k p(\mathbf{Z}_k) p(\bar{\mathbf{U}}_i|\mathbf{Z}_k) p(\mathbf{C}_j|\mathbf{Z}_k). \quad (10)$$

Using the Expectation-Maximization (EM) algorithm [31], we estimate the required probabilities $p(\mathbf{Z}_k)$, $p(\bar{\mathbf{U}}_i|\mathbf{Z}_k)$, and $p(\mathbf{C}_j|\mathbf{Z}_k)$, which correspond to the probability of a latent factor z_k, the probability of a thematic session set $\bar{u}_i$, given the latent factor, and the probability of accessing Web pages of a certain category c_j of the Web directory, given the latent factor, respectively. Note that the probability $p(\mathbf{C}_j|\mathbf{Z}_k)$ explains the rationale discussed above. A latent factor can describe a subset of the user's interests, i.e., categories in a session, while it might not describe other interests at all since these categories might be attributed to other latent factors. The above probabilistic model is further exploited for selecting the characteristic categories of the latent factors, i.e., the community models. This is realized by introducing a threshold value, named as *Latent Factor Assignment Probability (LFAP)*, on the probabilities $p(\mathbf{C}_j|\mathbf{Z}_m)$ and selecting those categories that satisfy this threshold. More formally, with each of the latent factors z_m, we associate the categories that satisfy:

$$p(\mathbf{C}_j|\mathbf{Z}_m) \geq LFAP. \quad (11)$$

The *LFAP* and *OCIA* criteria are combined in the new version of algorithm named as *Objective Probabilistic Directory Miner (OPDM)*. The initial Web directory is traversed and each category node which does not satisfy the LFAP and the PCAT thresholds is pruned. Similarly to the OCDM algorithm, it can be trivially shown that the categories of a community model form a Web directory. The process is described in Algorithm 2 in the Appendix.

In addition to the probabilistic OPDM soft clustering approach, we developed and tested a variant that uses fuzzy clustering based on the popular Fuzzy c-means [32]. However, the results for the latter approach were very poor and are not included here. This can be explained by the large number of categories that are present in the directory. The result of this for PLSA is that the vector of each category, represented by $P(\mathbf{C}|\mathbf{Z})$, contains many near-zero probability values. In the case of the fuzzy c-means, many of the corresponding fuzzy degrees are zero leading to a very sparse vector. This implies that the uncertainty that we deal with can better be modeled by a probabilistic rather than a fuzzy approach.

5.3 Objective Clustering and Probabilistic Directory Miner (OCPDM)

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 [33] clustering algorithm on the relation $\Re = (\bar{U}, C)$ for the creation of the initial communities. This approach differs from CDM clustering, as it produces nonoverlapping clusters, 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. 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 both observable and latent associations in the data, and potentially allow us to better model the interests 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. In particular, we define a relation $\Re^k \subseteq \Re$, where the ordered pair $(\bar{u}_i, c_j^k) \in \Re^k$ represents the access of category c_j^k or one of its leaf subcategories, belonging to cluster k, during thematic session set $\bar{u}_i$. Note that the fact that we have chosen nonoverlapping clusters simplifies our methodology since it relates the ordered pair $(\bar{u}_i, c_j^k)$ with a single cluster. Assuming the existence of a set $Z^k = \{z_1^k, z_2^k, \ldots, z_\mu^k\}$ that constitutes the new space of μ latent factors inside cluster k, the methodology is similar to the OPDM method described in Section 5.2, resulting in the estimation of the following probabilities: $p(\bar{\mathbf{U}}_i|\mathbf{Z}_\mu^k)$, the conditional probability of a thematic user session set given the factor z_μ^k, inside the cluster; $p(\mathbf{C}_j^k|\mathbf{Z}_\mu^k)$, the conditional probability of a cluster's category given the factor z_μ^k; and $p(\mathbf{Z}_\mu^k)$, the probability of a latent factor inside cluster k.

The new algorithm *Objective Clustering and Probabilistic Directory Miner (OCPDM)* invokes OPDM for each of the K clusters. In particular, the categories of the cluster on which each latent factor has the maximum impact are selected using the LFAP threshold. Thus, if Θ_μ^k is the community model of the latent factor z_μ^k, then:

$$\forall\, c_j^k \in \Theta_\mu^k,\ \exists\ j \in (1, \ldots, \mu)\ : p\big(\mathbf{C}_j^k|\mathbf{Z}_\mu^k\big) \geq LFAP. \quad (12)$$

Finally, the OCIA criterion is also applied as discussed in Section 5.2. This process leads to a more "specialized" community Web directory than the one that would be produced by k-means. The community Web directory includes only the most "dominant" categories of the community, i.e., the categories that satisfy not only the explicit associations, but also the hidden knowledge that exists in the data. The categories that do not satisfy these

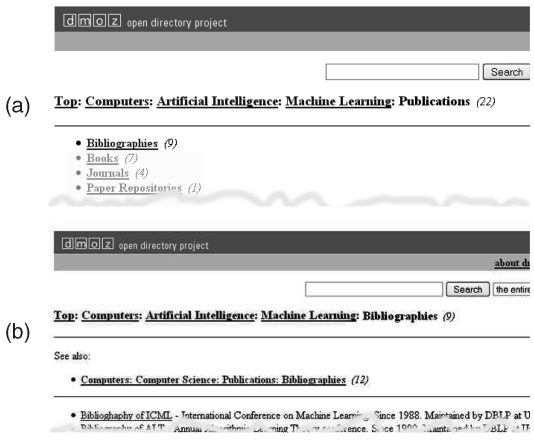


Fig. 3. *Shorten* Operator. Grayed-out categories are not included in the directory.

associations are dropped from the community model. The process is depicted in Algorithm 3 in the Appendix.

5.4 Community Web Directory Refinement

The result of the aforementioned pattern discovery methods is a hierarchy that corresponds to the community Web directory, 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:

Shorten operator. If a category has a single descendant node, then the category is removed. The application of the operator is depicted in Fig. 3. In Fig. 3a, we notice that the category labeled "Top/.../Publications/" deterministically leads to the leaf category "Top/.../Bibliographies." The *Shorten* operator will remove the category "Top/.../Publications". The resulting community Web directory is presented in Fig. 3b.

Absorb operator. This operator applies to categories that are leaves in the community Web directory, but not in the initial Web directory. For these categories, if all of their descendant categories are excluded from the community Web directory, then all the Web pages that were contained in the descendant leaves are absorbed. 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. In Fig. 4a, we assume that all the descendant categories labeled "Top/.../Bibliographies," "Top/.../Books," "Top/.../Journals," and "Top/.../Paper_Repositories" of the parent category Top/.../Publications are not included in the community Web directory. By applying the operator, all the Web pages in these categories will be absorbed by the parent "Top/.../Publications" category. The resulting community Web directory is presented in Fig. 4b.

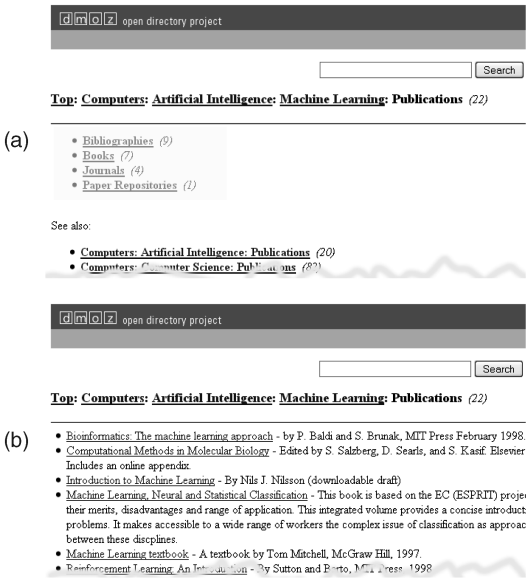


Fig. 4. *Absorb* Operator. Grayed-out categories are not included in the directory.

6 Evaluation of Community Web Directories

The methodology introduced in this paper for the construction of community Web directories has been tested in the context of a research project,[3] 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.

6.1 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 OCDM 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 OPDM approach, the models were built using 5, 10, 15, and 20 latent factors. For the OCPDM, 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 OPDM 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

3. WEB-C-Mine: Data Mining from Web Cache and Proxy Log Files.

scenario, the Web pages were clustered using hierarchical agglomerative clustering. Following the criteria discussed in [34] 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 [35]. 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. The requirement for the existence of target and observed Web pages (and corresponding categories) led us to use the thematic user sessions (Definition 3) instead of the thematic sessions sets (Definition 7). 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. The evaluation process is described below and in Algorithm 4 in the Appendix.

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 OCDM algorithm, we select the directories that contain the largest portion of the categories in a particular session, while in the case of OPDM and OCPDM, we select for each category c_j the three community directories that maximize the probability $p(\mathbf{C}_j|\mathbf{Z}_k)$.

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.

The evaluation is repeated iteratively by setting each different Web page in a test session as a "target" Web page. As a concrete example, consider the user session $v = (p_1, p_2, p_3)$ mapped onto the thematic user session u: ("Top/Business/Management," "Top/News/Newspapers," and "Top/Computers/Companies"). The evaluation process starts with the Web page p_1 as the "target" page. We exclude the category "Top/Business/Management," and the community Web directories that are selected for the two other categories (six directories) are used to construct the session-specific community Web directory, against which the target category is evaluated. The experiment is repeated considering as target each of the three pages in the session.

The evaluation of effectiveness employs two measures: *Coverage* and *User Gain*. Coverage corresponds to the predictiveness of our model, i.e., the number of target Web pages that are *covered* by the session-specific community directories. In particular, if p_t is the examined target Web page in a user session v_i and Θ_i is the session-specific Web directory constructed as described in Algorithm 4, the $Coverage$ of p_t is given by the following function:

$$Coverage(p_t) = \begin{cases} 1, & if\ p_t \in \Theta_i, \\ 0, & otherwise. \end{cases}$$

Thus, if A is the total number of target Web pages in the user sessions and A' the number of target pages that are covered by the session-specific community Web directories, the coverage of the set of the community Web directories $\mathbf{\Theta}$ is given by (13):

$$Coverage_{\mathbf{\Theta}} = \frac{\|A'\|}{\|A\|}. \tag{13}$$

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 Web directory to get to the desired Web page. In order to measure this, we developed a simple estimate of the effort a user is required to exert in order to arrive at the target page. We based this estimate on the path a user would have to follow to get to an interesting Web page, i.e., the sequence of categories from the root of the Web directory to the category that contains the Web page. In previous work [8], [9], we introduced a metric, named as *ClickPath*, which takes into account the depth of the navigation path, as well as the branching factor at each step. However, since we used only binary trees, the branching factor was always 2. This is not the case in the real Web directories, where the branching factor plays an important role. Additionally, at the leaf nodes, we measure the branching factor by the number of Web pages in the node. More formally, the *ClickPath* of a target page p_t is given by the following equation:

$$ClickPath(p_t) = d(p_t) + \sum_{j=1}^{d(p_t)} b_j, \tag{14}$$

where $d(p_t)$ is the depth of the path to p_t and b_j is the branching factor at the jth step of the path from the root to p_t. When $j = d(p_t)$, the branching factor corresponds to the number of Web pages that the leaf node c_t contains. Hence, for each target Web page that is covered by the corresponding session-specific community directory, we calculate the ClickPath in the community ($ClickPath_{CD}(p_t)$) and the

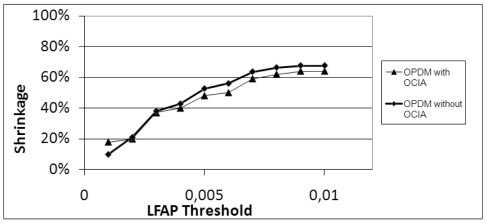


Fig. 5. ODP Web directory shrinkage using OPDM.

original Web directory ($ClickPath_{WD}(p_t)$). The *UserGain* for each target page is defined as follows:

$$UserGain(p_t) = \frac{ClickPath_{WD}(p_t) - ClickPath_{CD}(p_t)}{ClickPath_{WD}(p_t)}. \quad (15)$$

The average $UserGain_{A'}$ for all covered target pages is

$$\{UserGain\}_{A'} = \frac{\sum_{p_t \in A'} UserGain(p_t)}{|A'|}. \quad (16)$$

6.2 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 OPDM, 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 OPDM algorithm in Fig. 5. 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. Particularly interesting is the effect of the OCIA criterion which increases only slightly the size of the community Web directories. 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 the effectiveness 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. In Fig. 6, we present the trade-off curve for the old versions of the algorithms. Each curve in the two figures is generated by measuring Coverage and User Gain for the same values of the LFAP and the connectivity thresholds. Similar to the ROC curves, the optimal position is the top-left corner, where Coverage and

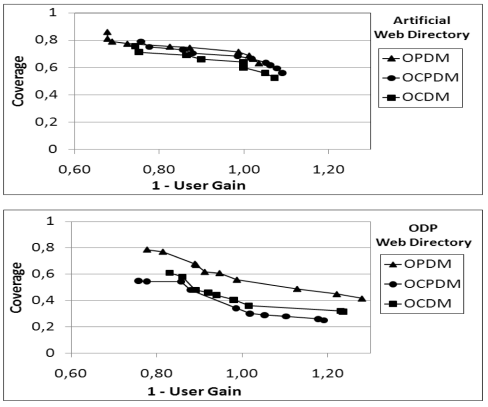


Fig. 6. Coverage/user gain trade-off without OCIA.

User Gain reach their maximum values. From this figure, we notice an interesting property of our older approach that did not handle the local overload problem. As the thresholds increase and more categories are "pruned" from the community Web directories, the user gain decreases and reaches even negative values (values above 1.00 in the x-axis of the figures). This property is more conspicuous in the ODP Web directory (bottom figure) and it means that after a certain threshold, the user does not benefit from the community Web directory. This is due to the fact that the community Web directories include only high-level nodes, each of which "absorbing" a large number of Web pages, from its children that have been left out.

Fig. 7 presents the trade-off curves for the new versions of the algorithms. The phenomenon of negative User Gain values does not appear here due to the handling of the local overload problem. Thus, the user seems to be benefiting from the personalization, even at high levels of coverage. Comparing the behavior of the methods on the two different directories, we have obtained a higher user gain at a small cost in coverage in the ODP directory. In particular, for values of user gain between 25 and 35 percent, the coverage of the ODP directory (bottom figure) reaches comparable values to the coverage of the artificial Web directory (top figure), i.e., from 60 to 80 percent. However, for the ODP directory, we obtain user gain around 50 percent, maintaining coverage at the level of 75 percent. This level of user gain is attainable due to the size and the generic nature of the

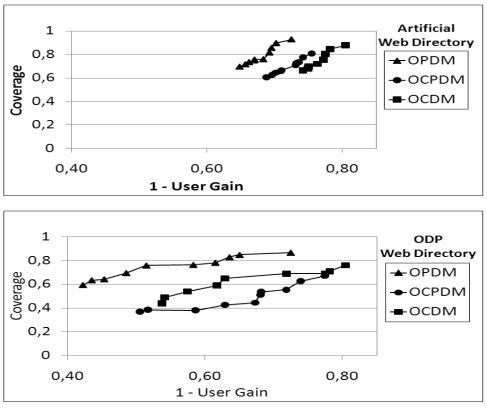


Fig. 7. Coverage/user gain trade-off with OCIA (average over PCAT threshold).

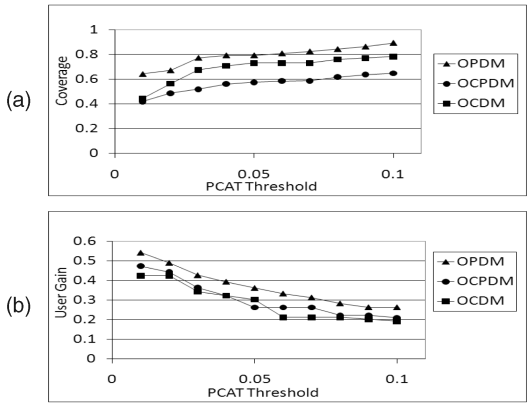


Fig. 8. Coverage and user gain (average over LFAP and connectivity thresholds).

ODP. Thus, as originally foreseen, the use of a real directory has revealed the power of personalizing the directory. Regarding the comparison of the three directory methods, OPDM clearly outperforms the other two in both cases. The performance of the "hybrid" OCPDM method is lower in terms of coverage due to the nonoverlapping nature of the k-means algorithm, which results in a more "strict" assignment of the categories to these models.

Looking more closely into the effect of the new OCI criterion, Fig. 8a shows the increase in coverage as the PCAT threshold increases. This is due to the fact that the selection becomes less strict and more leaf node categories are included in the community models. On the other hand, user gain drops, Fig. 8b, as more leaf nodes are retained in the models and the paths become longer. Similar results are obtained for the Artificial Web directory.

The results presented in this section provide a detailed picture of the benefits of our approach to personalizing Web directories. Regarding the discovery methods that we tested, the "pure" PLSA technique (OPDM) outperforms the simple clustering algorithm (OCDM) and the combination of the clustering and PLSA (OCPDM). The results also confirm our initial assumptions about the need to handle the local overload problem. The inclusion of thematically informative leaf nodes in the personalized directory leads to higher user gain values since the information is distributed and located in leaf nodes containing a small number of Web pages each. Moreover, we notice that the user gain obtained from the personalization of the ODP is higher than that for the artificial Web directory, with which we have tested our approach in the past. This is explained by the fact that the artificial is already focused on the usage data from which it was constructed. Thus, there is more room for personalization in a generic directory such as the ODP.

7 Conclusions

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 acquired results lead us to the conclusion that although we have obtained good performance by all methods, the use of PLSA for the personalization of Web directories appears to be the most promising. It helps identifying latent information in the users' choices and derives high-quality community directories that provide significant benefits to their users. Moreover, the use of objective (a priori) information about the directory categories helps tackling the local information overload problem that has been encountered in our earlier work. The results presented here provide 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 approximated the gain of the end user and have not taken into account the cost of "losses" that could be encountered in the case that the users do not find what they are looking for in the personalized directory. This issue requires the evaluation of community Web directories in user studies which are in our immediate plans for future work.

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.

Appendix

Algorithm 1. OCDM *(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 $\bar{u}_k$ thematic session sets}
Set $\mathbf{\Theta} \leftarrow \emptyset$ {The set of community Web directories}
Array $W_{ij} \leftarrow \emptyset$ {The adjacency matrix of graph G}
for all $c_i, c_j \in C, \bar{u}_k \in \bar{U}$ **do**

$w_{ij} \leftarrow Calculate\ (c_i, c_j, \bar{u}_k)$ {Calculate occurrence and co-occurrence values according to Equations 8, 9}
if $w_{ij} \leq Connectivity\ Threshold$ **then**
$w_{ij} \leftarrow 0$
else
$w_{ij} \leftarrow 1$
end if
end for
$\mathbf{\Theta} \leftarrow FindMaximalCliques(G, W_{ij})$ {Use Algorithm [30]}
for *all* $\Theta_r \in \mathbf{\Theta}$ **do**
for *all* $l_n \in L$ **do**
if $l_n \notin \Theta_r$ **then**
$c_i \leftarrow l_n$
repeat
$c_j \leftarrow parent(c_i, G)$ {Loop for each ansectral of leaf node}
$c_i \leftarrow c_j$
until $(c_j \in \Theta_r \vee c_j = root)$
if $(OCIA(\mathbf{C}_j, \mathbf{L}_n) \leq PCAT)$ **then**
$\Theta_r \leftarrow \Theta_r \bigcup \{l_n\}$
end if
end if
end for
end for

Algorithm 2. OPDM *(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

Algorithm 3. OCPDM *(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 $\bar{\mathbf{\Theta}} = \text{k-means}(C, \bar{U})$ {Apply k-means to discover k Web directories}
for *all* $\bar{\Theta}^k \in \bar{\mathbf{\Theta}}$ **do**
$\Theta^k \leftarrow$ OPDM $(\bar{\Theta}^k)$ {Apply OPDM to each k Web directory}
$\bar{\Theta}^k \leftarrow \bigcup_\mu \Theta^k_\mu : \Theta^k_\mu \in \Theta^k$ {μ Latent factor index.}
end for

Algorithm 4. CountTargetCategoriesCovered()
Set U {The set of u thematic user sessions}
Set Θ {The set of Community Web Directories}
$counterCovered \leftarrow 0$
for *all* $u_i \in U$ **do**
Set $\Theta_i \leftarrow \emptyset$ {The session-specific community Web directory of thematic user session u_i}
for *all* $c_j \in u_i$ **do**
Set $\Theta_{ij} \leftarrow \emptyset$ {The community Web directory of category c_j in session u_i}
$k \leftarrow 1$
$\Theta' \leftarrow \Theta$
while $k \leq 3 \wedge (\exists \Theta_k \in \Theta' : c_j \in \Theta_k)$ **do**
$\Theta_{ij} \leftarrow \Theta_{ij} \bigcup \Theta_k : \Theta_k \in \Theta' \wedge \arg\max_k(p(\mathbf{C}_j|\mathbf{Z}_k))$
$\Theta' \leftarrow \Theta' \setminus \Theta_k$
$k \leftarrow k + 1$
end while
end for
for *all* $c_m \in u_i$ **do**
$\Theta_i \leftarrow \bigcup_j \Theta_{ij} : j \neq m$
if $c_m \in \Theta_i$ **then**
$counterCovered \leftarrow counterCovered + 1$
end if
end for
end for

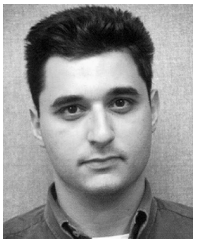
ACKNOWLEDGMENTS

This research has been partially funded by the Greece-Cyprus Research Cooperation project "Web-C-Mine: Data Mining from Web Cache and Proxy Log Files." The authors would like to thank M. Dikaiakos for helping them to obtain the test data, as well as the following people who collaborated with them in the project: Vangelis Karkaletsis, Christos Papatheodorou, and Christiana Christophi.

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] T. Hofmann, "Learning What People (Don't) Want," *Proc. 12th European 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 User Modeling for Recommendation," *Proc. 10th Int'l Conf. User Modeling,* L. Ardissono, P. Brna, and A. Mitrovic, eds., pp. 109-118, 2005.
[16] D. Chen, D. Wang, and F. Yu, "A PLSA-Based Approach for Building User Profile and Implementing Personalized Recommendation," *Proc. Joint Ninth Asia-Pacific Web Conf. (APWeb '07) and Eighth Int'l Conf. Web-Age Information Management (WAIM '07),* pp. 606-613, 2007.
[17] B. Mehta and N. Wolfgang, "Unsupervised Strategies for Shilling Detection and Robust Collaborative Filtering," *User Modeling and User-Adapted Interaction,* vol. 19, nos. 1/2, pp. 65-97, 2009.
[18] P.A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschütter, "Using odp Metadata to Personalize Search," *Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval,* pp. 178-185, 2005.
[19] A. Sieg, B. Mobasher, and R. Burke, "Ontological User Profiles for Representing Context in Web Search," *Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence and Intelligent Agent Technology—Workshops,* pp. 91-94, 2007.
[20] Z. Ma, G. Pant, and O.R.L. Sheng, "Interest-Based Personalized Search," *ACM Trans. Information Systems,* vol. 25, no. 1, article no. 5, Feb. 2007.
[21] T. Oishi, K. Yoshiaki, M. Tsunenori, H. Ryuzo, F. Hiroshi, and M. Koshimura, "Personalized Search Using odp-Based User Profiles Created from User Bookmark," *Proc. 10th Pacific Rim Int'l Conf. Artificial Intelligence,* pp. 839-848, 2008.
[22] J. Garofalakis, T. Giannakoudi, and A. Vopi, "Personalized Web Search by Constructing Semantic Clusters of User Profiles," *Proc. 12th Proc. Int'l Conf. Knowledge-Based Intelligent Information and Eng. Systems,* pp. 238-247, 2008.
[23] C.R. Anderson and E. Horvitz, "Web Montage: A Dynamic Personalized Start Page," *Proc. 11th Int'l Conf. World Wide Web,* pp. 704-712, May 2002.
[24] B. Smyth and C. Cotter, "Personalized Adaptive Navigation for Mobile Portals," *Proc. 15th European Conf. Artificial Intelligence,* 2002.
[25] J. Chaffee and S. Gauch, "Personal Ontologies for Web Navigation," *Proc. Ninth Int'l Conf. Information and Knowledge Management (CIKM '00),* pp. 227-234, 2000.
[26] T. Dalamagas, P. Bouros, T. Galanis, M. Eirinaki, and T. Sellis, "Mining User Navigation Patterns for Personalizing Topic Directories," *Proc. Ninth Ann. ACM Int'l Workshop Web Information and Data Management,* pp. 81-88, 2007.
[27] G. Salton and M. McGill, *Introduction to Modern Information Retrieval.* McGraw-Hill, 1986.
[28] Y. Zhao and G. Karypis, "Evaluation of Hierarchical Clustering Algorithms for Document Datasets," *Proc. ACM Int'l Conf. Information and Knowledge Management (CIKM),* pp. 515-524, Nov. 2002,
[29] H.W. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in C,* second ed. Cambridge Univ. Press, 1992.
[30] C. Bron and J. Kerbosch, "Algorithm 457-Finding All Cliques of an Undirected Graph," *Comm. ACM,* vol. 16, no. 9, pp. 575-577, 1973.
[31] T. Hofmann, "Probabilistic Latent Semantic Analysis," *Proc. 15th Conf. Uncertainty in Artificial Intelligence (UAI' 99),* pp. 289-296, 1999.
[32] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms.* Kluwer Academic Publishers, 1981.
[33] J. Hartigan, *Clustering Algorithms.* John Wiley & Sons, 1975.
[34] Y. Zhao and G. Karypis, "Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering," *Machine Learning,* vol. 55, no. 3, pp. 311-331, 2004.
[35] J.S. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," *Proc. 14th Conf. Uncertainty in Artificial Intelligence (UAI '98),* pp. 43-52, 1998.



Dimitrios Pierrakos received the BSc degree in physics from the University of Athens, Greece, and the MSc degree in information technology from the University College London, UK. He is currently working toward the PhD degree in computer science at the Department of Informatics and Telecommunications, University of Athens. He works as a research assistant in the Institute of Informatics and Telecommunications of the NCSR "Demokritos." His research interests lie in the areas of user modeling, Web mining, and Web personalization. He is a member of the IEEE.

Georgios Paliouras received the PhD degree in computer science from the University of Manchester, UK. He is a senior researcher in the Institute of Informatics and Telecommunications of NCSR "Demokritos." His research focuses on machine learning and knowledge discovery for ontology learning, user modeling, event recognition, information extraction, and text classification.