Construction of Web Community Directories by Mining Usage Data

Dimitrios Pierrakos[1,3], Georgios Paliouras[1], Christos Papatheodorou[2],
Vangelis Karkaletsis[1], Marios Dikaiakos[4]

[1] Institute of Informatics and Telecommunications, NCSR "Demokritos",
15310 Ag. Paraskevi, Greece
{dpie, paliourg}@iit.demokritos.gr
[2] Department of Archive & Library Sciences, Ionian University
49100, Corfu, Greece
papatheodor@ionio.gr
[3] Department of Informatics and Telecommunications, University of Athens
15784, Athens, Greece
[4] Department of Computer Science, University of Cyprus
CY1678, Nicosia, Cyprus
mdd@ucy.ac.cy

Abstract. This paper introduces the concept of Web Community Directories, as a means of personalizing services on the Web, and presents a novel methodology for the construction of these directories by usage mining methods. The community models are extracted with the use of the Community Directory Miner, a simple cluster mining algorithm which has been extended to ascend a concept hierarchy, such as a Web directory, and specialize it to the needs of user communities. The construction of the communities is based on usage data collected by the proxy servers of an Internet Service Provider, which is also a task that has not been addressed in the literature. The examined data present a number of peculiarities such as their large volume and their semantic diversity. Initial results presented in the paper illustrate the use of the methodology and provide an indication of the behavior of the new usage mining method.

1. INTRODUCTION

As the size of the Web is increasing at a galloping pace, the information overload of its users emerges as one of the Web's major shortcomings. One approach towards the alleviation of this problem is the organization of Web content into thematic hierarchies, also known as *Web directories*. A Web directory, such as Yahoo! [17] or the Open Directory Project (ODP) [11], allows Web users to locate information that relates to their interests, through a hierarchy navigation process. Typically, a Web directory is maintained by human experts.

This approach suffers from a number of problems, the most important of which is the difficulty of the task of manually categorizing the entire Web, as there are millions of Web pages and the rate of expansion is very high. In addition, the size and complexity of the directories is canceling out any gains that were expected with respect to the information overload problem, i.e., it is often difficult to navigate to the information of interest to a particular user.

A different approach to the problem of information overload is the personalization of the services on the Web, which aims to make the Web a friendlier environment for its individual user. *Web Personalization* [8] focuses on the adaptability of Web-based information systems to the needs and interests of individual users, or groups of users. Typically, a personalized Web site recognizes its users, collects information about their preferences and adapts its services, in order to match the users' needs.

A major obstacle towards realizing Web personalization is the acquisition of accurate and operational models for the users. Reliance to manual creation of these models, either by the users or by domain experts, is inadequate for various reasons, among which the annoyance of the users and the difficulty of verifying and maintaining the resulting models. An alternative approach is that of *Web Usage Mining* [16], which uses data mining methods to create models, based on the analysis of usage data, i.e., records of how a service on the Web is used. Web usage mining provides a methodology for the collection and preprocessing of usage data, and the construction of models representing the behavior and the interests of users [14].

In this paper, we propose a new solution to the problem of information overload, by combining the strengths of Web Directories and Web Personalization, in order to address some of the above-mentioned issues. In particular we focus on the construction of usable Web directories that correspond to the interests of groups of users, known as *user communities*. The construction of user community models with the aid of Web Usage Mining has so far only been studied in the context of specific Web sites [12]. This approach is extended here 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). The final goal is the construction of community-specific Web Directories.

Web Community Directories can be employed by various services on the Web, such as Web portals, in order to offer their subscribers a more personalized view of the Web. 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. In this manner, the information overload of the user is reduced, while at the same time the service on the Web obtains added value for its customers.

The construction of community directories with usage mining raises a number of interesting research issues, which are addressed in this paper. The first challenge is the analysis of large datasets in order to identify community behavior. In addition to 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 the increased dimensionality and the semantic incoherence of the data, i.e., the Web pages that have been accessed. As a result, the task of constructing a thematic hierarchy as a community model, without examining the content of the Web pages that have been accessed, arises as an additional issue.

These issues are addressed by extending a community construction method to take advantage of existing Web directories and specialize them to the interests of particular communities. The basic data mining algorithm that has been developed for that purpose is called *Community Directory Miner* (CDM). It is an extension of the *cluster mining* algorithm, which has been employed for the construction of site-specific user communities in previous work [13]. The new method proposed here is able to ascend an existing Web directory in order to arrive at a suitable level of semantic characterization of the interests of a particular user community.

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

2. Related Work

In recent years, the exploitation of usage mining methods for Web personalization has attracted considerable attention, mainly for the recommendation of links to follow within a site (e.g. [5], [9], [10] and [18]), or for the customization of Web sites to the preferences of the users [15]. These systems use information from Web server log files to construct user models that represent the behavior of the users. Their differences are in the method that they employ for the construction of user models, as well as in the way that this knowledge, i.e., the models, is exploited. In [5], [9] and [18] the authors employ clustering methods to create user models that represent common usage patterns. Based on these models the systems recommend dynamically and in real time Web pages to the visitors of a Web site. Similarly to these systems, page recommendation is also provided by the SiteHelper tool [10]. However, instead of cluster analysis, SiteHelper employs classification techniques, where the Web pages that a user has visited are considered positive examples for the training of the classifier. The result of the process is a set of rules that model the user's interests. Having discovered these rules the system can recommend Web pages to the users according to their interests. The customization of a Web site is another application of the knowledge that can be extracted from usage data. A system that adopts this approach is the WUM tool [15], which employs a sequential pattern discovery method to extract navigation patterns from Web site logs. These patterns are subsequently used to customize the Web site.

Personalized Web directories, on the other hand, are mainly associated with services such as Yahoo! [17] and Excite [4], which support the manual personalization of their directories by the user. An initial approach to automate this process, with the aid of usage mining methods, is the Montage system [1]. This system is used to create personalized portals, consisting primarily of links to the Web pages that a particular user has visited, while also organizing the links into thematic categories according to the ODP directory. Web pages are extracted from a proxy server log file, and their topics are estimated using a probabilistic text classification method that assigns a Web page to the most likely category, out of a set of ODP categories. For the personalization of the service and the construction of the user model a number of heuristic metrics are used, such as the interest in a page or a topic, the probability of revisiting a page, etc.

An alternative approach is the construction of a directory of useful links (bookmarks) for an individual user, as adopted by the PowerBookmarks system [6]. The system collects Web pages that a user frequently visits, as well as pages that link to frequently visited pages (referrer information), and pages that are returned as query results from a search engine. All this information is considered bookmark information for a particular user. Text classification techniques are used for the assignment of labels to Web pages, using significant keywords from the text and generic classification systems such as LCC (Library of Congress Classification). The results of the classification process are used to organize bookmarks in a tree for the user.

An important issue regarding the above-mentioned methods is the scalability of the content-based classification methods that they use. These methods may be suitable for constructing models of what a user usually views, but their extendibility to aggregate user models is questionable. Furthermore, the requirement for a small set of predefined classes complicates the construction of rich hierarchical models.

In contrast to the existing work, this paper proposes a novel methodology for the construction of Web directories according to the preferences of user communities. The community models are aggregate user models,

and are constructed with the use of a simple cluster mining method, which has been extended to ascend a concept hierarchy, such as a Web directory, and specialize it to the preferences of the community. The construction of the communities is based on usage data collected by the proxy servers of an Internet Service Provider (ISP), which is also a task that has not been addressed in the literature. This type of data has a number of peculiarities, such as its large volume and its semantic diversity, as it records the navigational behavior of the user throughout the Web, rather than within a particular Web site. The methodology presented in this paper handles these problems, while it proposes a new way of exploiting the knowledge that is extracted by the usage mining process. Instead of link recommendation or site customization, it focuses on the construction of Web community directories, as a new way of personalizing the services on the Web.

3. Constructing Web community Directories

The construction of Web community directories is seen here as the end result of a usage mining process on data collected at the proxy servers of a central service on the Web. This process consists of the following steps:

- *Data Collection and Preprocessing*, comprising the collection and cleaning of the data, their characterization according to an existing Web directory, such as ODP, and the identification user sessions.
- *Pattern Discovery*, comprising the extraction of user communities from the data with a suitably extended cluster mining technique, which is able to ascend a thematic hierarchy, in order to discover interesting patterns.
- *Knowledge Post-Processing*, comprising the translation of community models into Web community directories and their evaluation.

An architectural overview of the discovery process is given in Figure 1, and described in the following sections.

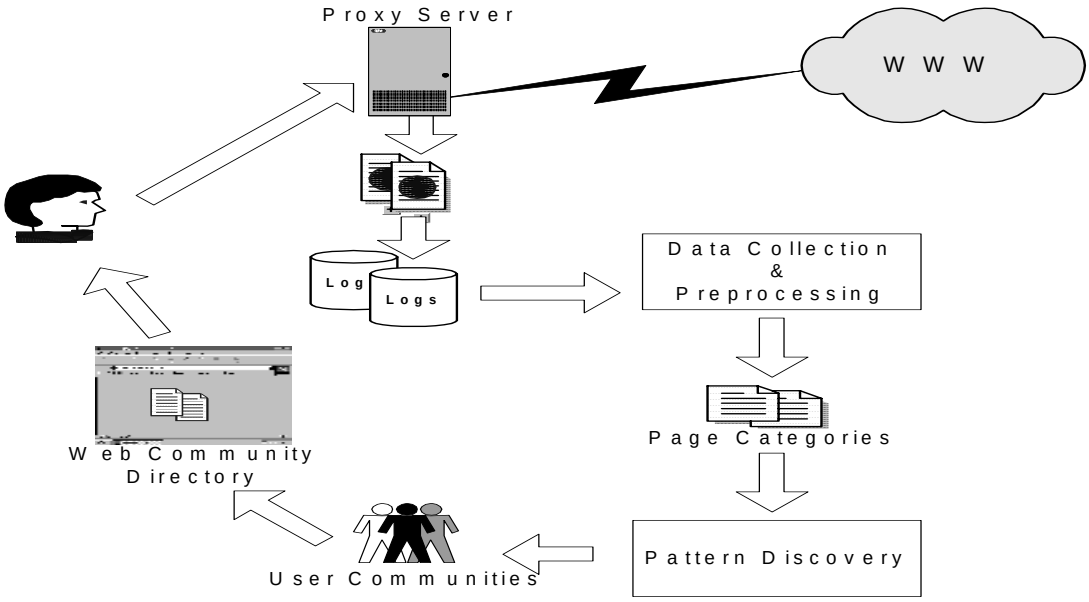

Figure 1: The process of constructing Web Community Directories

3.1 Data Collection & Preprocessing

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 behavior of the subscribers, during their navigation through the Web. The information that is recorded in the logs and exploited by our approach consists of a date and time stamp, the IP that has been allocated to the subscriber, the Web resource that has been requested, the response status and the Content-Type header of the resource, e.g. text/html. No record of the user's identification is being kept or used, in order to avoid privacy violations.

The usage data collected in the logs are usually diverse and voluminous. The outgoing traffic is much higher than the usual incoming traffic of a Web site and the Web pages less coherent semantically. The task of data preprocessing is to assemble these data into a consistent, integrated and comprehensive view, in order to be used for pattern discovery.

The first stage of data preprocessing involves data cleaning. The aim is to remove as much noise from the data as possible, in order to keep only the Web pages that are directly related to the user behavior. This involves the filtering of the log files to remove data that are downloaded without a user explicitly requesting them, such as multimedia content, advertisements, Web counters, etc. Records with HTTP error codes that correspond to bad requests, or unauthorized accesses are also removed. The result of the data cleaning phase is a file that contains only the information required in the subsequent phases.

The second stage of data preprocessing involves the categorization of Web pages into thematic categories, reducing thus the dimensionality and the semantic diversity of the data. The task of Web page categorization has been studied in the literature, e.g. [3], [7]. Typically, in these approaches text classification methods are used to construct models for a small number of known thematic categories of a Web directory, such as that of Yahoo!. These models are then used in order to assign each Web page to a category. This process has a number of limitations with respect to the methodology proposed here, such as its low coverage of the Web directory and the requirement to analyze the content of each Web page in the log.

As a result, we have adopted a different approach for categorizing Web pages, based only on usage information, i.e., the Uniform Resource Locator (URL) of the Web page. Instead of looking at the content of each Web page, we extract its domain from the URL and then search for the category of that domain within an existing Web directory, such as the ODP [11]. This method provides a mapping between the domains and the categories as they appear in the directory. The simplifying assumption underlying this approach is that the category of the top-level domain of a Web site characterizes the Web pages within a site. This simplification is supported by the fact that the categories that are listed in Web directories refer, to a large extent, to the "home" pages of the Web sites that have been classified in the directory and thus, to the domain of the site. Furthermore, it is a generalizing assumption that may be tolerated in the construction of aggregate user models, such as the user communities.

The third stage of data preprocessing involves the extraction of access sessions. An access session is a sequence of log entries, i.e., accesses to Web pages, for the same IP address, where the time interval between two subsequent entries does not exceed a certain time interval. Access sessions are the main input to the pattern discovery phase, and are extracted using the following procedure:

1. Grouping the logs by date and IP address.

2. Selecting a time-frame within which two records from the same IP address can be considered to belong in the same access session.
3. Grouping the categories accessed by the same IP address within the selected time-frame to form a session.

Finally, access sessions are translated into binary attribute vectors. Each attribute in the vector represents the presence of a particular thematic category in that session.

3.2 Extraction of Web Communities

Once the data have been translated into attribute vectors, they are used to discover patterns of interest, in the form of community models. This is done by the *Community Directory Miner* (CDM), an enhanced version of the cluster mining algorithm. This data mining approach is based on the work presented in [13] for site-specific communities.

Cluster mining discovers patterns of common behavior by looking for all maximal fully-connected subgraphs (cliques) of a graph that represents the user's characteristic attributes, which in our case correspond to page categories. The method starts by constructing a weighted graph $G(A,E,W_A,W_E)$. The set of vertices A corresponds to the descriptive attributes used in the input data. The set of edges E corresponds to attribute co-occurrence as observed in the data. For instance, if the user visits pages belonging to the categories "Business" and "Arts" an edge is created between the relevant vertices. The weights on the vertices W_A and the edges W_E are computed as the attribute occurrence frequencies and co-occurrence frequencies respectively.

Figure 2 shows an example of such a graph. The connectivity of the graph is usually very high. For this reason we make use of a *connectivity threshold* aiming to reduce the edges of the graph. This threshold is related to the frequency of the categories inside the data. In our example in Figure 2, if the threshold is 0.07 the edge ("Business", "Games") is dropped. Once, the connectivity of the graph has been reduced, all maximal cliques of the graph are generated, each one corresponding to a community model. One important advantage of this approach to community modeling is that each user may be assigned to many categories, unlike most user clustering methods.

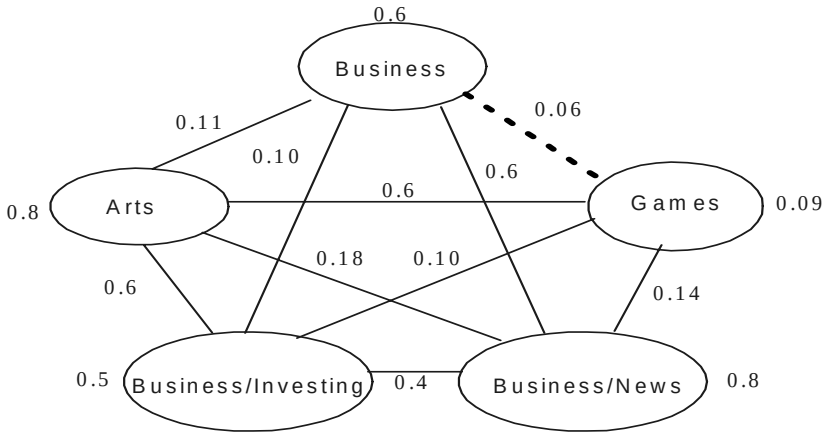

Figure 2: An example of a graph for cluster mining

CDM enhances cluster mining so as to be able to ascend a hierarchy of topic categories. This is achieved by updating the weights of the vertices and the nodes in the graph. Initially, each category is mapped onto a set of categories, corresponding to all of its parents and grandparents in the thematic hierarchy. Thus, the category:

"Computers/Internet/Searching/Directories/Yahoo"

is mapped onto the following categories:

"Computers",
"Computers/Internet",
"Computers/Internet/Searching",
"Computers/Internet/Searching/Directories",
"Computers/Internet/Searching/Directories/Yahoo".

The frequency of each of these categories is increased by the frequency of the initial child category. Thus, the frequency of each category corresponds to its own original frequency, plus the frequency of all of its children. The CDM algorithm can be summarized in the following steps:

Step 1*: Compute frequencies of categories that correspond to the weights of the vertices*. More formally, if a_{ij} is the value of an attribute *i* in the binary attribute vector *j*, and there are *N* vectors, the weight of the vertice w_i for that attribute is calculated as follows:

$$w_i = \frac{\sum_{j=1}^{N} a_{ij}}{N}$$

Step 2: *Compute co-occurrence frequencies between categories that correspond to the edges of the graph*. If a_{ik}^{j} is a binary indicator of whether attributes *i* and *k* co-occur in vector *j*, then the weight of the edge w_{ik} between these two attributes is calculated as follows:

$$w_{ik} = \frac{\sum_{j=1}^{N} a_{ik}^{j}}{N}$$

Step 3: *Update the weights of categories, i.e. vertices, by adding the frequencies of their children*. More formally, if w_p is the weight of a parent vertex *p* and w_i is the weight of a child vertex *i*, the final weight w'_p of the parent is computed as follows:

$$w'_p = w_p + \sum_{i} w_i$$

This calculation is repeated recursively ascending the hierarchy of the Web directory. Similarly, the edge weights are updated, as all the parents and grandparents of the categories that co-occur in a session, are also assumed to co-occur.

Step 4: *Find all maximal cliques in the graph* [*] *of categories, as in cluster mining.*

The underlying assumption for this 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 that exists in the data, and ascend the topic hierarchy accordingly. The result is the construction of the topic tree, even if only a few nodes of the tree exist in the usage data.

4. Post-Processing and Model Evaluation

The discovered patterns are sets of categories that are organized into topic trees and represent the community models, i.e., behavioral patterns that occur frequently in the data. These models are directly usable as Web community directories, and can be delivered by various means to the users of a community. A pictorial view of such a Web directory is shown in Figure 3, where the community directory is "superimposed onto a larger Web directory, such as ODP. Grey boxes represent the categories that belong to a particular community, while white boxes represent the rest of the categories in the Web directory. Note, that the categories "Business", "Society" and "Sports" do not necessarily occur frequently in the usage data. However, they appear in the community model, due to the frequency of their children. Furthermore, some of their children, e.g. "Society/Relationships" and "Society/Religion" (the spotted grey boxes) may also not be sufficiently frequent to appear in the model. Nevertheless, they force their parent category, i.e., "Society" into the model.

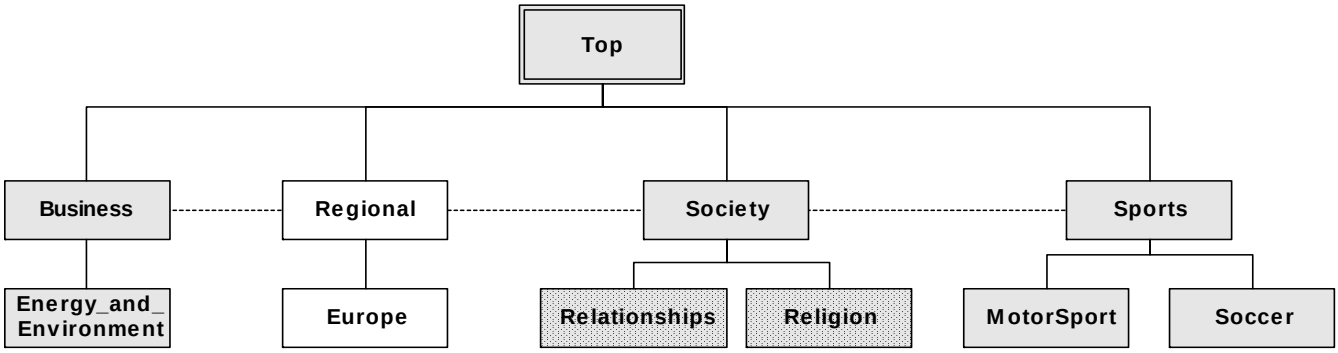

Figure 3: Web Community Directory Example.

Having generated the community models, we need to decide on the desired properties of these models, in order to evaluate them. For this purpose, we use ideas from existing work on community modeling and in particular the measure of *distinctiveness* [13]. When there are only small differences between the models, accounting for variants of the same community, the segmentation of users into communities is not interesting. Thus, we are interested in community models that are as distinct from each other as possible. We measure the distinctiveness

[*] We use the algorithm of [2] for finding all maximal cliques.

of a set of models *M* by the ratio between the number of distinct attributes that are covered and the size of the model set *M*. Thus, if there are *J* communities in *M*, A_j the attributes used in the j-th model, and A' the different attributes appearing at least in one model, distinctiveness is given by the following equation:

$$Distinctiveness(M) = \frac{|A'|}{\sum_j |A_j|}$$

As an example, if there exits a community model with the following simple (one-level) communities:

Community 1: Business, Regional, Society

Community 2: Society, Sports

Community 3: Business, Sports

then the number of distinct attributes are 4, i.e., Business, Regional, Society, and Sports, while the total number of attributes is 7. Thus, the distinctiveness of the model is 0.57. In the case of more complex community directories, each node of the directory counts as a separate attribute. The optimization of distinctiveness by a set of community models indicates the presence of useful knowledge in the set. Additionally, the number of distinct categories that are used in a set of community models, i.e., A', is also of interest as it shows the extent to which there is a focus on a subset of categories by the whole population of users. These two measures are used in the experimental results presented in the following section.

5. Experimental Results

The methodology introduced in this paper for the construction of Web community directories has been tested in the context of a research project, which focuses on the analysis of usage data from the proxy server logs of an Internet Service Provider. In an initial evaluation phase, we analyzed log files consisting of 781,069 records, and the results of this analysis are presented here.

In the stage of pre-processing, data cleaning has been performed in order to remove records that are unlikely to have been requested explicitly by the user, such as images or ads, as well as to remove “bad requests”. The remaining data has been characterized by mapping pages to domains and domains to thematic categories. Based on these characterized data, we constructed 3,037 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 characterization of the sessions uses 259 distinct ODP categories.

The resulting sessions were translated into binary vectors and were analyzed by the CDM algorithm, in order to identify community models, in the form of thematic trees. The resulting models were evaluated using the two measures that were mentioned in section 3.3, i.e. the distinctiveness and the number of categories, while varying the connectivity threshold. Figures 4 and 5 present the results of this process.

Figure 4 shows how the distinctiveness of the resulting community models increases as the connectivity threshold increases, i.e., as the requirement on the frequency of occurrence/co-occurrence becomes “stricter”. The rate of increase is higher for smaller values of the threshold and starts to dampen down for values above 200. This effect is justified by the high rate of decrease for the number of distinct categories in the models, as shown in Figure 5. Less than 50 categories have a frequency of occurrence of 200 and above (threshold 0.06),

while the level of distinctiveness at that threshold value exceeds 0.6, i.e., 60% of the categories that appear in the models are distinct. At that level, 18 community directories are constructed, which is a reasonable figure for the customers of a medium-sized ISP.

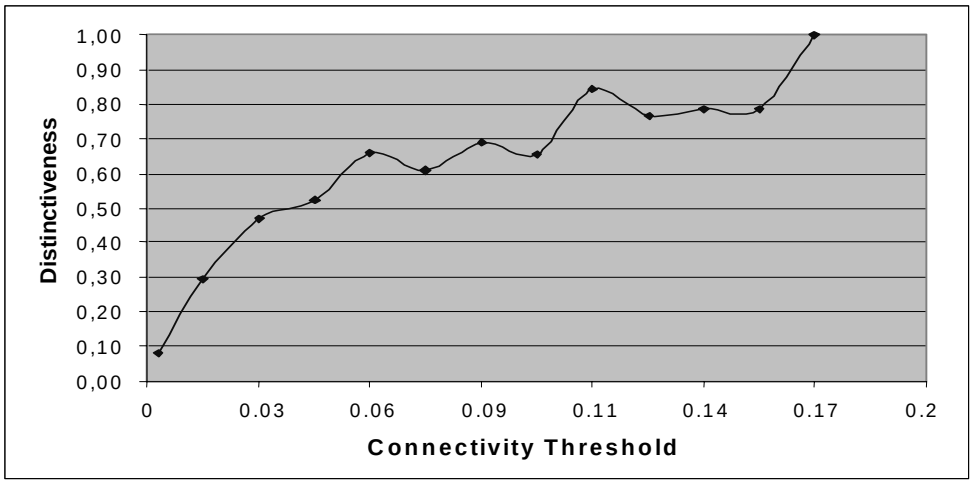

Figure 4: Distinctiveness as a function of Connectivity Threshold.

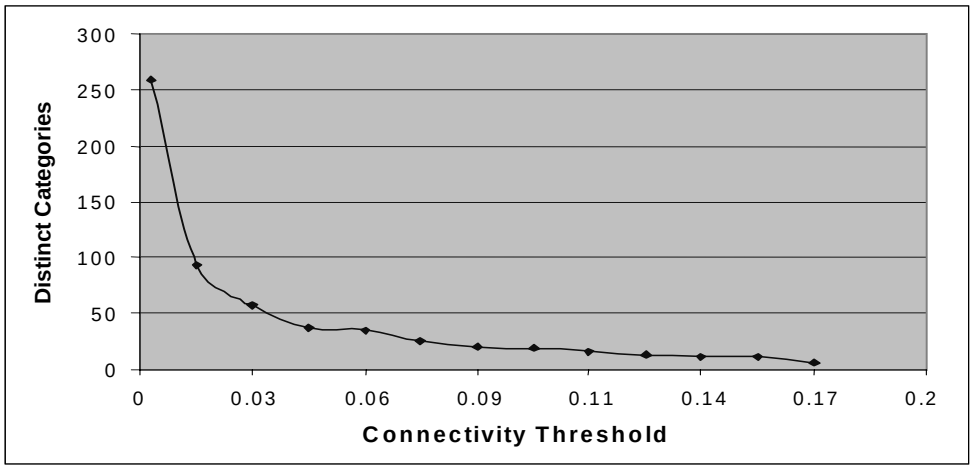

Figure 5: Distinct Categories as a function of Connectivity Threshold.

These figures provide an indication of the behavior and the effectiveness of the community modeling algorithm. At the same time, they assist in the process of selecting an appropriate value for the connectivity threshold and a corresponding set of community directories. Although these are initial results and further experimentation is required, they are very encouraging for the exploitation of the proposed methodology.

6. Conclusions and Future Work

This paper has presented a novel methodology for the personalization of Web Directories with the aid of Web usage mining methods. The concept of a Web community directory has been introduced, which corresponds to a usable directory of the Web, customized to the needs and preferences of user communities. User community

models take the form of thematic hierarchies and are constructed by a cluster mining algorithm, which has been extended to take advantage of existing Web directories, ascending their hierarchical structure appropriately.

We have tested this methodology by applying it on usage data collected at the proxy servers of an ISP and have provided initial results, indicative of the behavior of the mining algorithm. Proxy server usage data have introduced a number of interesting challenges, such as their size and their semantic diversity. The proposed methodology handles these problems by reducing the dimensionality of the problem, through the categorization of individual Web pages into the categories of an existing directory. In this manner, the corresponding community models are constructed in the form of thematic hierarchies, without the requirement for analysis of the content of Web pages.

The combination of two different approaches to the problem of information overload on the Web, i.e. thematic hierarchies and personalization, as proposed in this paper, introduces a promising research direction, where many open issues arise. Various components of the methodology could be replaced by a number of alternatives. Most importantly, more sophisticated methods for extracting the categories from usage data, in addition to the use of an existing Web directory, would make the mapping of pages to domains and then to categories more accurate and complete. Furthermore, other data mining methods could be adapted to the task of discovering community directories and compared to the algorithm presented here. Finally, additional evaluation is required, in order to test the robustness of the data mining algorithm to a changing environment and the usability of the resulting community directories.

7. ACKNOWLEDGEMENTS

This research has been funded by the Greek-Cyprus Research Cooperation project "Web-C-Mine: Data Mining from Web Cache and Proxy Log Files".

8. REFERENCES

1. Anderson, C. R. and Eric Horvitz. Web Montage: A Dynamic Personalized Start Page. In Proceedings of the 11th World Wide Web Conference (WWW 2002), 2002
2. Bron, C. and J. Kerbosch. Algorithm 457---finding all cliques of an undirected graph. Communications of the ACM, 16, 9, 575-577, 1973
3. Chen, H. and S. T. Dumais. Bringing order to the web: automatically categorizing search results. In Proceedings of CHI'00, Human Factors in Computing Systems, 145-152, 2000
4. Excite, http://www.excite.com
5. Kamdar, T., and A. Joshi. On Creating Adaptive Web Sites using WebLog Mining. Technical Report TR-CS-00-05. Department of Computer Science and Electrical Engineering University of Maryland, Baltimore County, 2000

6. Li W-S., Q. Vu, E. Chang, D. Agrawal, Y. Hara, and H. Takano. PowerBookmarks: A System for Personalizable Web Information Organization, Sharing, and Management. In Proceedings of the Eighth International World Wide Web Conference, 1999\
7. Mladenic, D. Turning Yahoo into an Automatic Web-Page Classifier. In Proceedings of the 13th European Conference on Artificial Intelligence, ECAI'98, 473-474, 1998
8. Mobasher, B., R. Cooley, and J. Srivastava. Automatic personalization based on Web usage mining. TR-99010, Department of Computer Science. DePaul University, 1999
9. Mobasher, B., R. Cooley, and J. Srivastava. Creating Adaptive Web Sites Through Usage-Based Clustering of URLs. In Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99), 1999.
10. Ngu, D. S. W., and X. Wu. SiteHelper: A Localized Agent that Helps Incremental Exploration of the World Wide Web. Computer Networks and ISDN Systems: The International Journal of Computer and Telecommunications Networking, 29**,** 8, 249-1255, 1997
11. Open Directory Project (ODP). http://dmoz.org
12. Paliouras, G., C. Papatheodorou, V. Karkaletsis, and C.D Spyropoulos. Clustering the Users of Large Web Sites into Communities. In Proceedings of Intern. Conf. on Machine Learning (ICML), Stanford, California, 719-726, 2000
13. Paliouras, G., C. Papatheodorou, V. Karkaletsis, C.D. Spyropoulos. Discovering User Communities on the Internet using Unsupervised Machine Learning Techniques., Interacting with Computers Journal, 14,6, 761-791, 2002
14. Pierrakos, D., Paliouras, G., Papatheodorou, C., Spyropoulos, C.D.: Web Usage Mining as a Tool for Personalization: a survey , User Modeling and User Adapted Interaction, to appear
15. Spiliopoulou, N., and L. C. Faulstich. WUM: A Web Utilization Miner. In International Workshop on the Web and Databases. Valencia, Spain, 1998
16. Srivastava, J., R. Cooley, M. Deshpande, and Tan, P. T. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. In SIGKDD Explorations, 1, 2, 2000
17. Yahoo! http://www.yahoo.com
18. Yan, T. W., M. Jacobsen, H. Garcia-Molina, and U. Dayal. From User Access Patterns to Dynamic Hypertext Linking. In Proceedings of the 5th WWW Conference, Paris, France, 1996