Construction of Web Community Directories using Document Clustering and Web Usage Mining

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

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

Abstract. This paper presents the concept of Web Community Directories, as a means of personalizing services on the Web, together with a novel methodology for the construction of these directories by document clustering and 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, and specialize it to the needs of user communities. The initial concept hierarchy is generated by a content-based document clustering method. Communities are constructed on the basis of usage data collected by the proxy servers of an Internet Service Provider. These data present a number of peculiarities such as their large volume and semantic diversity. Initial results presented in the paper illustrate the use of the methodology and provide an indication of the behavior of the new mining method.

1 Introduction

The hypergraphical architecture of the Web has been used to support claims that the Web will make Internet-based services really user-friendly. However, at its current state, the Web has not achieved its goal of providing easy access to online information. Being an almost unstructured and heterogeneous environment it creates an information overload and places obstacles in the way users access the required information.

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! [19] or the Open Directory Project (ODP) [14], allows Web users to locate information that relates to their interests, through a hierarchy navigation process. This approach suffers though from a number of problems. The manual creation and maintenance of the Web directories leads to limited coverage of the topics that are contained in those directories, since 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 for a particular user to navigate to interesting information.

An alternative solution is the personalization of the services on the Web. *Web Personalization* [10] focuses on the adaptability of Web-based information systems to the needs and interests of individuals or groups of users and aims to make the Web a friendlier environment. 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* [18], 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 [16].

In this paper, we propose a 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 [15]. 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 is reduced, while at the same time the service offers added value to 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, 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. In order to address these issues we create a thematic hierarchy of the Web pages by examining their content, and assign the Web pages to the categories of this hierarchy. An agglomerative clustering approach is used to construct the hierarchy with nodes representing content categories. A community construction method then exploits the constructed hierarchy and specializes it to the interests of particular communities. The basic data mining algorithm that has been developed for that purpose, the *Community Directory Miner* (CDM), is an extension of the *cluster mining* algorithm, which has

been used for the construction of site-specific communities in previous work [15]. The new method proposed here is able to ascend an existing directory in order to arrive at a suitable level of semantic characterization of the interests of a particular community.

The rest of this paper is organized as follows. Section 2 presents existing approaches to Web personalization with usage mining methods that are related to the work presented here. Section 3 presents in detail our methodology for the construction of Web community directories. Section 4 provides 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 and a number of 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. Clustering methods, e.g. [7], [11] and [20], classification methods, e.g. [13], and sequential pattern discovery, e.g. [17], have been employed to create user models. These models are subsequently used to customize the Web site and recommend links to follow. Usage data has also been combined with the content of Web pages in [12]. In this approach content profiles are created using clustering techniques. Content profiles represent the users' interests for accessed pages with similar content and are combined with usage profiles to support the recommendation process. A similar approach is presented in [6]. Content and usage data are aggregated and clustering methods are employed for the creation of richer user profiles. In [4], Web content data are clustered for the categorization of the Web pages that are accessed by users. These categories are subsequently used to classify Web usage data.

Personalized Web directories, on the other hand, are mainly associated with services such as Yahoo! [19] and Excite [5], which support manual personalization 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, organized into thematic categories according to the ODP directory. For 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 [8]. The system collects "bookmark" information for a particular user, such as frequently visited pages, query results from a search engine, etc. Text classification techniques are used for the assignment of labels to Web pages. An important issue regarding these methods is the scalability of the classification methods that they use. These methods may be suitable for constructing models of what a single 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 existing work, this paper proposes a novel methodology for the construction of Web directories according to the preferences of user communities, by combining document clustering and usage mining techniques. A hierarchical clustering method is employed for document clustering using the content of the Web pages. Subsequently, the hierarchy of document categories is exploited by the Web usage mining process and the complete paths of this hierarchy are used for the construction of Web community directories. This approach differs from the related work mentioned above, where the content of the Web pages is clustered in order to either enhance the user profiles or to assign Web usage data to content categories.

The community models are aggregate user models, 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 adequately 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 users throughout the Web, rather than within a particular Web site. The methodology presented here addresses these issues and proposes a new way of exploiting the extracted knowledge. Instead of link recommendation or site customization, it focuses on the construction of Web community directories, as a new way of personalizing 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 using the content of the Web pages, and the identification of user sessions. Note that this step involves a separate data mining process for the discovery of content categories and the characterization of the pages.
- *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.

3.1 Data Collection and 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 navigation of the subscribers through the Web. No record of the user's identification is being used, in order to avoid privacy violations. However, the 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 visited 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.

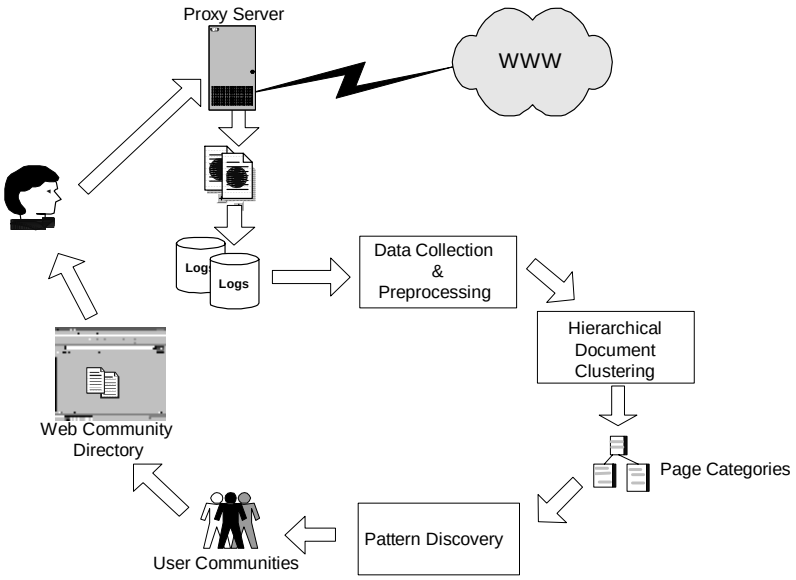

Fig. 1. The process of constructing Web Community Directories.

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 second stage of data preprocessing involves the thematic categorization of Web pages, thus reducing the dimensionality and the semantic diversity of data. Typically, Web page categorization approaches, e.g. [3] and [9], use text classification methods 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 to assign each visited page to a category. The limitation of this approach with respect to the methodology proposed here, is that it is based on a dataset for training the classifiers, which is usually limited in scope, i.e., covers only part of the directory. Furthermore, a manually-constructed Web directory is required, suffering from low coverage of the Web.

In contrast to this approach, we build a taxonomy of Web pages included in the log files. This is realized by a document clustering approach, which is based on terms that are frequently encountered in the Web pages. Each Web page is represented by a binary feature vector, where each feature encodes the presence of a particular term in

the document. A hierarchical agglomerative approach [21] is employed for document clustering. The nodes of the resulting hierarchy represent clusters of Web pages that form thematic categories. By exploiting this taxonomy, a mapping can be obtained between the Web pages and the categories that each page is assigned to. Moreover, the most important terms for each category can be extracted, and be used for descriptive labeling of the category. For the sake of brevity we choose to label each category using a numeric coding scheme, representing the path from the root to the category node, e.g. "1.4.8.19" where "1" corresponds to the root of the tree.

This document clustering approach has the following advantages: first a hierarchical classification of Web documents is constructed without any human expert intervention or other external knowledge; second the dimensionality of the space is significantly reduced since we are now examining the page categories instead of the pages themselves; and third the thematic categorization is directly related to the preferences and interests of the users, i.e. the pages they have chosen to visit.

The third stage of preprocessing involves the extraction of access sessions. An access session is 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 interval. In our approach, pages are mapped onto thematic categories and therefore an access session is translated into a sequence of categories. Access sessions are the main input to the pattern discovery phase, and are extracted as follows:

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 Web pages (thematic categories) accessed by the same IP address within the selected time-frame to form a session.

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

3.2 Extraction of Web Communities

Once the data have been translated into feature 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 approach is based on the work presented in [15] 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 users' characteristic features, i.e., thematic categories in our case. The method starts by constructing a weighted graph $G(A,E,W_A,W_E)$. The set of vertices A corresponds to the descriptive features used in the input data. The set of edges E corresponds to feature co-occurrence as observed in the data. For instance, if the user visits pages belonging to categories "1.3.5" and "1.7.8" an edge is created between the relevant vertices. The weights on the vertices W_A and the edges W_E are computed as the feature occurrence 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 thematic

categories in the data. In our example in Figure 2, if the threshold is 0.07 the edge ("1.3.5", "1.3.6") 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 is that each user may be assigned to many communities, unlike most user clustering methods.

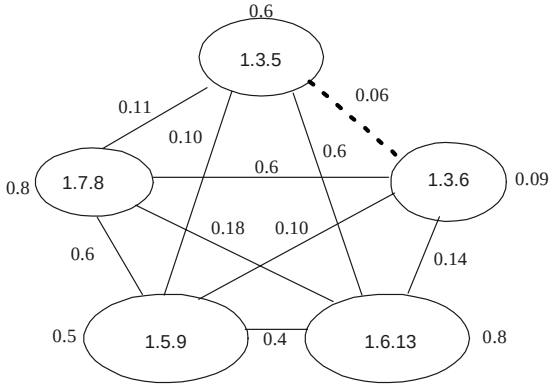

Fig. 2. An example of a graph for cluster mining.

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 its parent and grandparents in the thematic hierarchy. Thus, the category "1.6.12.122.258" is also mapped onto the following categories: "1", "1.6", "1.6.12", "1.6.12.122". 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 its children. The underlying assumption for the update of the weights is that if a certain category exists in the data, then its parent categories should also be examined for the construction of the community model. In this manner, even if a category (or a pair of categories) have a low occurrence (co-occurrence) frequency, their parents may have a sufficiently high frequency to be included in a community model. This enhancement allows the algorithm to start from a particular category and ascend the topic hierarchy accordingly. The result is the construction of a topic tree, even if only a few nodes of the tree exist in the usage data.

The 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 a feature *i* in the binary feature vector *j*, and there are *N* vectors, the weight of w_i for that vertice is calculated as follows:

$$w_i = \frac{\sum_{j=1}^{N} a_{ij}}{N}. \quad (1)$$

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

$$w_{ik} = \frac{\sum_{j=1}^{N} a_{ik}^{j}}{N}. \quad (2)$$

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:

$$w'_p = w_p + \sum_{i} w_i \quad (3)$$

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 the maximal cliques in the graph of categories* [2].

3.3 Post-Processing and Model Evaluation

The discovered patterns are topic trees, representing 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 directory is shown in Figure 3, where the community directory is "superimposed" onto the hierarchy of categories. Grey boxes represent the categories that belong to a particular community, while white boxes represent the rest of the categories in the Web directory. Each category has been labeled using the most frequent terms of the Web pages that belong to this category. The categories "12.2", "18.79" and "18.85" appear in the community model, due to the frequency of their children. Furthermore, some of their children, e.g. "18.79.5" and "18.79.6" (the spotted boxes) may also not be sufficiently frequent to appear in the model. Nevertheless, they force their parent category, i.e., "18.79" into the model.

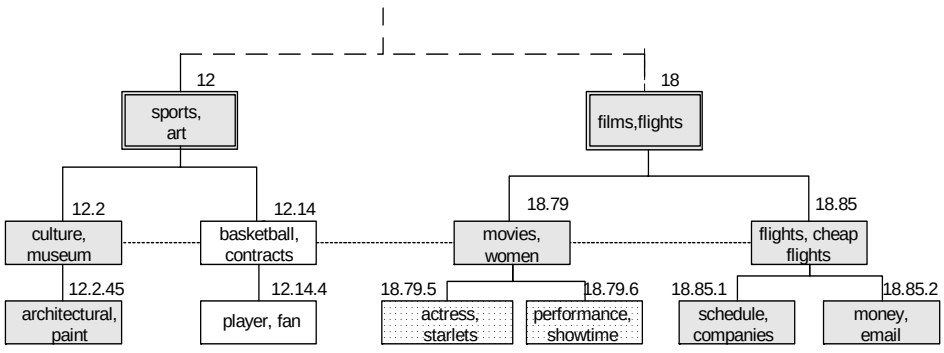

Fig. 3. An example of a Web community directory.

Having generated the community models, we need to decide on their desired properties, in order to evaluate them. For this purpose, we use ideas from existing work on community modeling and in particular the measure of *distinctiveness* [15]. 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 of a set of models M by the ratio between the number of distinct categories that are covered and the number of models in M. Thus, if J the number of models in M, A_j the categories used in the j-th model, and A' the different categories appearing at least in one model, distinctiveness is given by equation 4.

$$Distinctiveness(M) = \frac{|A'|}{\sum_{j} |A_j|}. \qquad (4)$$

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 A' that are used in a set of community models is also of interest as it shows the extent to which there is a focus on a subset of categories by the users. These two measures are used in the experimental results presented in the following section.

4. 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. We analyzed log files consisting of 781,069 records, and the results are presented here.

In the stage of pre-processing, data cleaning has been performed and the remaining data has been characterized using the hierarchical agglomerative clustering mentioned in section 3.1. The process resulted in the creation of 998 distinct categories. Based on these characterized data, we constructed 2,253 user sessions, using a time-interval of 60 minutes as a threshold on the "silence" period between two consecutive requests from the same IP. After mapping the Web pages of the sessions to the categories of the hierarchy, we translated the sessions into binary vectors and analyzed them by the CDM algorithm, in order to identify community models, in the form of topic 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 distinct categories, while varying the connectivity threshold. Figures 4 and 5 present the results of this process.

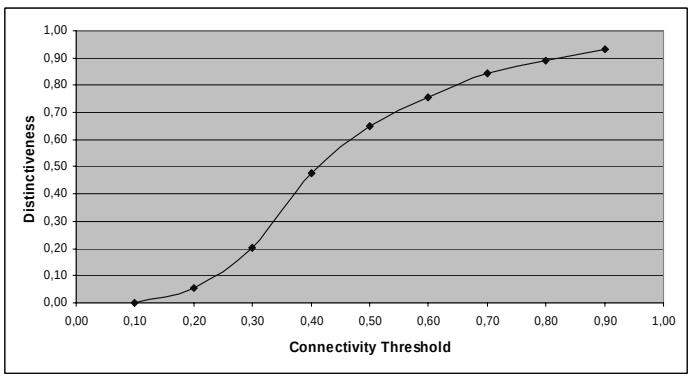

Fig. 4. Distinctiveness as a function of the connectivity threshold.

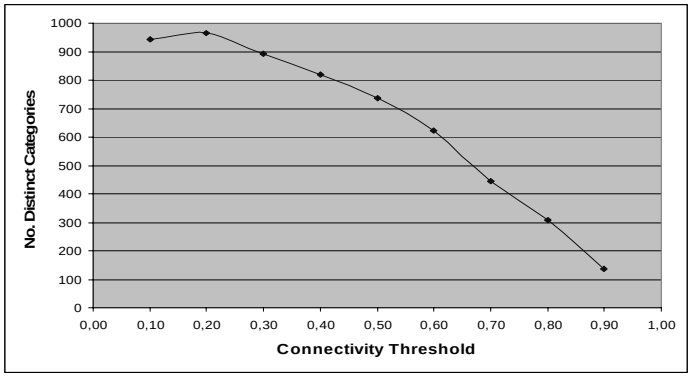

Fig. 5. Number of distinct categories as a function of the connectivity threshold.

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 0.7. This effect is justified by the decrease in the number of distinct categories, as shown in Figure 5. Nevertheless, more than half of the categories have frequency of occurrence greater than 600 (threshold 0.6), while at that threshold value the level of distinctiveness exceeds 0.7, i.e. 70% of the categories that appear in the model are distinct. These figures provide an indication of the behavior and the effectiveness of the community modeling algorithm. At the same time, they assist in selecting an appropriate value for the connectivity threshold and a corresponding set of community models. The results are encouraging for the exploitation of the proposed methodology.

5. Conclusions and Future Work

This paper has presented a novel methodology for the personalization of Web Directories with the aid of document clustering and Web usage mining. The concept of a Web community directory has been described, corresponding 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 an existing directory, and ascend its hierarchical structure. The initial directory is generated by a document clustering algorithm, based on the content of the pages appearing in an access log.

We have tested this methodology by applying it on access logs collected at the proxy servers of an ISP and have provided initial results, indicative of the behavior of the mining algorithm. Proxy server logs 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 a Web directory, as constructed by document clustering. In this manner, the corresponding community models take the form of thematic hierarchies.

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 new issues arise. Various components of the methodology could be replaced by a number of alternatives.

For instance, other mining methods could be adapted to the task of discovering community directories and compared to the algorithm presented here. Similarly, different methods of constructing the initial thematic hierarchy could be examined. Finally, additional evaluation is required, in order to test the robustness of the mining algorithm to a changing environment and the usability of the resulting community directories.

Acknowledgements

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".

REFERENCES

1. Anderson, C. R. and Eric Horvitz.: Web Montage: A Dynamic Personalized Start Page. In Proceedings of the 11th WWW Conference 2002, (2002)
2. Bron, C., Kerbosch, J.: Algorithm 457---finding all cliques of an undirected graph. Communications of the ACM, 16, 9, 575-577, (1973)
3. Chen, H., Dumais, S. T.: Bringing order to the web: automatically categorizing search results. In Proceedings of CHI'00, Human Factors in Computing Systems, 145-152, (2000)
4. Cooley, R.: Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data. Ph.D. Thesis, University of Minnesota, (2000).
5. Excite, http://www.excite.com
6. Heer, J., Chi, Ed H.: Identification of Web User Traffic Composition using Multi-Modal Clustering and Information Scent. In Proceedings of the Workshop on Web Mining, SIAM Conference on Data Mining, 51-58. (2001)
7. Kamdar, T., Joshi, A.: On Creating Adaptive Web Sites using WebLog Mining. TR-CS-00-05. Department of Computer Science and Electrical Engineering University of Maryland, Baltimore County, (2000)
8. Li W-S., Vu, Q., Chang, E., Agrawal, D., Hara, Y., Takano, H.: PowerBookmarks: A System for Personalizable Web Information Organization, Sharing, and Management. In Proceedings of the 8th WWW Conference, (1999)
9. Mladenic, D.: Turning Yahoo into an Automatic Web-Page Classifier. In Proceedings of the 13th European Conference on Artificial Intelligence, 473-474, (1998)
10. Mobasher, B., Cooley, R., Srivastava, J.: Automatic personalization based on Web usage mining. TR-99010, Department of Computer Science. DePaul University, (1999)
11. Mobasher, B., Cooley, R., Srivastava, J.: Creating Adaptive Web Sites Through Usage-Based Clustering of URLs. In Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop, *(*1999)
12. Mobasher, B., H. Dai, T. Luo, Y. Sung, J. Zhu.: Integrating Web Usage and Content Mining for More Effective Personalization. In Proceedings of the International Conference on E-Commerce and Web Technologies. Greenwich, UK, 165-176, (2000)
13. Ngu, D. S. W., Wu, X.: 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)
14. Open Directory Project (ODP). http://dmoz.org
15. Paliouras, G., Papatheodorou, C., Karkaletsis, V, Spyropoulos, C.D.: Discovering User Communities on the Internet using Unsupervised Machine Learning Techniques., Interacting with Computers Journal, 14,6, 761-791, (2002)

16. 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
17. Spiliopoulou, N., Faulstich, L. C.: WUM: A Web Utilization Miner. In International Workshop on the Web and Databases. Valencia, Spain, (1998)
18. Srivastava, J., Cooley, R., Deshpande, M., Tan, P. T.: Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. In SIGKDD Explorations, 1, 2, (2000)
19. Yahoo! http://www.yahoo.com
20. Yan, T. W., Jacobsen, M., Garcia-Molina, H., Dayal, U.: From User Access Patterns to Dynamic Hypertext Linking. In Proceedings of the 5th WWW Conference, Paris, France, (1996)
21. Zhao, Y., Karypis, G.: Evaluation of hierarchical clustering algorithms for document datasets. In CIKM, (2002)