Mining User Communities in Digital Libraries

Christos Papatheodorou[1] Sarantos Kapidakis[1]

Michalis Sfakakis[2] Alexandra Vassiliou[3]

[1] Department of Archive and Library Sciences, Ionian University,
Eleftherias Sq., Paleo Anaktoro, Corfu 49100, Greece
{papatheodor, sarantos}@ionio.gr
Phone:+30-26610-874{19, 13} Fax: :+30-26610-87436

[2] National Documentation Centre,
48 Vas. Constantinou, Athens 11635, Greece
msfaka@ekt.gr
Phone:+30-210-7273961

[3] Library, National Centre for Scientific Research "Demokritos",
Aghia Paraskevi, Athens 15310, Greece
alex@lib.demokritos.gr
Phone:+30-210-6503287 Fax: :+30-210-6522965

Abstract. The interest in the analysis of library user behavior has been increasing rapidly, since the advent of digital libraries and Internet. In this context, we analyze the queries posed to a digital library and recorded into the Z39.50 session log files and we construct communities of users with common interests, using data mining techniques. One of our main concerns is the construction of meaningful communities that can be used for improving information access. Analysis of the results brings to surface some of the important properties of the task, suggesting the feasibility of a common methodology.

1 Introduction

Information services aim to satisfy the needs of their users in a way that ensures precision and effectiveness. Many of these services use intelligent information retrieval and filtering techniques to personalize and customize their content to the users interests and preferences (Basu et al. 1998; Billsus and Pazzani 2000; Anderson and Horvitz 2002). Several information providers exploit user modeling techniques to understand and evaluate the usage of their services. The study of user behavior has become a crucial point in a number of digital library projects (Peterson Bishop 1995; Payette and Rieger 1997; Esposito et al. 1998) and specifies a number of critical factors during the design and development process of a digital library (Van House et al. 1996).

Data mining offers powerful techniques for discovering nontrivial and useful patterns in voluminous datasets. Many of such techniques have been applied to information services and especially to the Web offering personalization and improving information access (Chakrabarti et al. 1999). The application of such techniques in library

systems revealed the bibliomining domain, which aims to upgrade almost all the decision making processes concerning library and information management.

The goal of this paper is to show how the administrator of a digital library can analyze the behavior of the users and extract information, which is useful for improving information access. In particular, we are interested in formulating community models, which represent patterns of usage of digital libraries and can be associated with different types of user. Our main concern is the association of the digital library queries. Similar queries, recorded into Z39.50 session log files, are grouped into clusters. The clusters map to user community models, which represent the users demands and querying habits. Such models could be used in a query expansion process contributing to efficient retrieval. Generally, the query analysis can be beneficial to both the digital library and its users in many ways:

Service optimization. It can help the administrators to re-organize the digital library content, authorities and user interfaces making them more suitable for different user groups.

Decision support. It can help the administrator to form an effective query expansion strategy for the digital library.

Personalization. It can help the users identify information of interest to them by recommending similar subjects.

Our work is motivated by the need to improve the querying mechanisms provided by the digital library of the Hellenic National Documentation Centre (NDC) (http://theses.ndc.gr) which is one of the most significant in Greece and consists of many collections that are unique world wide, having internationally interesting content. The digital library of NDC is targeted to a number of diverse types of user groups (e.g. students, researchers, professionals, librarians, etc.), mainly in Greece, from a variety of scientific domains.

In the following section we describe the problem of creating user communities and we present the followed methodology. Then we describe the digital collections of NDC, their characteristics, the targeted user groups they refer to and the functionality of the available operations by the system. In section 4 we evaluate our methodology to two different collections of the NDC. Finally we present a number of interesting issues derived from this work for further research.

2 Problem statement and Methodology

User Communities correspond to groups of users who exhibit common behavior in their interaction with an information system (Orwant, 1995). This type of user model has only recently been paid some attention, primarily in the context of collaborative filtering, which aims to personalize an information service without having to analyze the content of the service (Breese et al., 1998; Pennock et al., 2000).

Our objective is to propose a methodology for the:

– Discovery of communities of users who express their needs using common vocabulary (mainly in controlled terms, such as subject headings or authors, where terms are selected from a big global authority file), by clustering the logged queries. Each community is described by a subset of the digital library authority file.

- Recommendation of particular terms for query expansion. The suggested terms correspond to relevant search terms used by the community users.

We consider the community discovery problem as a typical data mining task. Hence, the stages for creating user communities from Z39.50 server logs are the same as those of any other data mining task: data collection and pre-processing, pattern discovery and knowledge post-processing.

Data Collection and pre-processing

The pre-processing aims to enable them to be used as input to the next stage of pattern discovery. The Z39.50 protocol organizes the user queries into sessions (containing one or more queries) and for each log file record that refer to searches, we keep the following data: session code, date and time of the query, origin IP address, code of the collection accessed and the search terms of the access under consideration.

We prepare the training dataset, i.e. a table for which each column (feature) corresponds to a search term and each row (objects) corresponds to a session. Each row is a binary vector and each vector cell indicates the existence of a search term in a session.

Pattern discovery

Given the training data in the appropriate form, interesting patterns are extracted with the use of machine learning techniques, such as clustering, classification, association rule discovery etc. We have opted for the use of unsupervised learning through clustering, due to the requirement for the formulation of user groups with no prior decisions. We propose two simple and efficient graph theory based clustering algorithms, which search for either all the cliques or all the connected components of a graph that represents the sessions' characteristic features (i.e. search terms).

The process starts by constructing a graph $G(A,E)$. The set of vertices A corresponds to the search terms used in the user queries. The set of edges E corresponds to search terms co-occurrence as observed in the user queries. For instance, in the NDC digital library that we examine, if a session contains the search terms "Museums-Administration" and "Archaeology" an edge is added between the relevant vertices. The weights on the vertices and the edges are computed as the frequencies of the users' choices and their combinations respectively. Edge frequencies are normalized by dividing them with the maximum of the frequencies of the two vertices that they connect. The effect of the normalization is to remove the bias for characteristics that appear very often in all users.

The connectivity of the resulting graph G is usually high. For this reason we make use of a *connectivity threshold*, aiming to the reduction of the number of edges in the graph. The connectivity threshold represents the minimum weight allowed for the edge existence. When this threshold is high the graph will be sparse and the derived clusters will have small size and high quality. When the threshold is lower the derived clusters will be larger.

After the edge reduction the clique mining method accepts as clusters, i.e. user communities all the existing cliques, which are: all the subgraphs of the graph G in

which every pair of nodes has an edge between them. Despite the large complexity of the clique-finding problem, the implemented algorithm (Bron and Kerbosch 1973) is very fast.

Alternatively, we could use the connected components discovery method, which parses the graph G and searches for all the subgraphs in which every pair of nodes is connected by a path of edges between them. The derived components represent disjoint groups of search terms and each of them corresponds to a user community.

The main advantages for using the connected components mining method are mainly its efficiency (the algorithm complexity is almost linear on the number of sessions) as well as its ability to discover patterns in very sparse data sets (Perkowitz and Etzioni 1998). As it will be shown in section 4, a sparse graph that contains a number of connected components may be too sparse to contain any sizeable cliques. However the cliques mining method provides more cohesive and coherent clusters. One more difference between the two clustering methods is that the clique mining produces overlapping clusters, while the connected components are, by default, disjoint clusters.

Pattern post-processing and evaluation

There are several issues concerning the expressiveness of the clusters produced by the proposed algorithms, as well as the recommendation of the appropriate subject headings to the digital library users. The main questions that have to be answered are:

- Which clusters constitute actually user communities
- Which subject terms provide meaningful descriptions of the user communities (i.e. community models) and could be recommended in a query expansion process.

In the case of clique mining, we evaluate the generated clusters by varying the connectivity threshold, and measuring the following two properties:

Coverage: the proportion of search terms participating in the clustering, since due to the connectivity threshold not all the search terms are members of the generated clusters.

Overlap: the extent of overlap between the clusters. This is the average value of the size of the intersection of any pair of clusters divided by the size of their union (Jaccard co-efficient).

In the case of discovering connected components there is no overlap between them, so the criterion for selecting clusters as meaningful user communities is the coverage of the search terms. Moreover, the components that are exactly the same with user sessions are dropped, as they do not form user communities but simply user sessions.

The terms describing each user community come from frequently used search terms, which exist in the digital library authority files. These terms characterize each community, i.e. constitute the *community model*. The terms that make up community model could be recommended to the users as alternative or additional search terms in a query expansion mechanism.

3 Description of the usage data

NDC digital library hosts many collections covering various scientific domains. All the collections are structured using the UNIMARC format with almost the same level of quality attributes specificity, completeness of fields, syntactic correctness, but a different level of the quality consistency attribute as described in every category. From their 300,000 metadata records there are links giving online access to 14,000 digitized documents composed of 2,000,000 scanned pages and few other object formats.

The web-based retrieval system that we monitored is implementing a Z39.50 client and connects to a Z39.50 server. The users start their sessions by selecting and connecting to a collection. After connecting to a collection, a user may express his search request or browse specific access points and then retrieves the documents.

For our experiment, two categories of digital library collections were chosen, including two collections each, based on the completeness of their metadata descriptions as well as the different ways that the users access them - Table 1 presents the access point usage in the above NDC collections:

1. Hellenic Archaeological Records – grARGOS (*C1*) and International Archaeological Records – intARGOS (*C2*). They consist of archaeological records, including library material with diverse types of data and a good consistency level, targeted to a specific scientific user group (e.g. researchers on Archaeology) and have the largest authority files among all collections. However only the 5% of the users pose queries through the "Subject heading" access point, while more than 30% use the "Author" access point. This is due to the nature of the collection and the high degree of specialization of the users and because the authors of the works, included in these collections, are known by their direct relation with their subjects, places and findings – often coming from excavations. For example, Evans is directly related with the palace of King Minoa at Knossos, Crete. This explains how the "Author" access point has in fact a thematic character in ARGOS user sessions.
2. Hellenic School Libraries (*C3*) and Hellenic Public Libraries Union Catalogue (*C4*). They are union catalogs for library material from many domains, consisting of general-purpose records, and a very good consistency level, targeted to librarians, and have the smallest authority files. The users access them using all the main access points uniformly. The high percentage of the usage of "Subject heading" access point is explained because the target group of these collections are librarians, i.e. an homogeneous user group experienced in searching and retrieving information.

Table 1. Summary of Access Point usage per Collection

Access Point	Total	C1	C2	C3	C4
Any	50.5%	39%	30%	30%	35%
Author	19.3%	28%	36%	23%	26%
Title	16%	26%	26%	21%	18%
Subject Heading	9.5%	4%	5%	22%	15%

The operations performed by the users on the content of the two selected collection categories were logged for a period of thirty-two months (September 2000 till April 2003), keeping the sessions that used the "Authors" and "Subject heading" access points for the two collection categories. Table 2 shows the numbers of the logged sessions, queries and search terms per collection, during the mentioned period.

Table 2. Logged sessions, queries and search terms per Collection and Access Point

	ARGOS		Public-School Libraries	
	Authors	Subject heading	Authors	Subject heading
Sessions	8,172	1,141	4,965	2,148
Queries	15,069	1,742	10,566	5,622
Search terms	10,020	1,434	6,114	2,336

4 Experimental Results

In the following paragraph we present the results from the application of the clique mining method on the datasets concerning the "Author" access point and then the results from the application of the connected components mining method on the datasets concerning the "Subject heading" access point.

"Author" access point analysis - Cliques

The processing of the data sets from both collection categories concerning the "Author" access point resulted to cliques for various values of the connectivity threshold. Depending on the value of the connectivity threshold the coverage of the clusters and the overlap varied. Table 3 presents the results along those two dimensions. The column named "Number of cliques" indicates the amount of cliques that correspond to a particular connectivity threshold. As expected, the coverage for small threshold values is large due to the large number of cliques. A similar behavior follows the overlap. For both collections around the threshold value 0.5, about the ten percent of the search terms appear in the cliques (coverage equals to 0.12 for ARGOS and 0.14 for Public-School Libraries), while there is little overlap between the cliques (for ARGOS equals to 0.0009 and for Public-School Libraries equals to 0.0007). This observation suggests the selection of this threshold value for the formation of the desired representative vocabulary. At this connectivity threshold value, 69 clusters are generated for the ARGOS collection category and 74 for Public-School Libraries.

Table 3. The results of the clique mining algorithm on access point "Author"

Connectivity Threshold	ARGOS			Public-School Libraries		
	Number of Cliques	Coverage	Overlap	Number of Cliques	Coverage	Overlap
0.3	407	0.64	0.001	454	0.69	0.001
0.4	266	0.44	0.001	289	0.46	0.001
0.5	69	0.12	0.0009	74	0.14	0.0007
0.7	41	0.08	0.0004	30	0.06	0.00

Table 4 presents an indicative example of two user community models per each collection category.

Table 4. A sample of user community models on access point "Author"

Com-munity	ARGOS Connectivity threshold = 0.5	PUBLIC-SCHOOL LIBRARIES Connectivity threshold = 0.5
1	Kunz, Hagu, Orszag, Kubcza	Tomboulides, Reeve, Slott, Pinar
2	Brancacci, Dubella, Cinuzz, Marciol, Adrian, Valde	Harrison, Hutchinson, Morgan, Aitke, Littlewod, Rinvolucre, Bygat

"Subject heading" access point analysis - Connected Components

The clique-mining algorithm has been applied on the datasets produced by the logs for the two access points and for each collection category. However for both collection categories, the graph generated from the "Subject heading" access point dataset was too sparse. Even with no connectivity threshold, the algorithm could not produce any sizeable cliques. Therefore for the "Subject heading" access point dataset we finally used a different method, the connected components mining method.
The component mining algorithm has been applied on the two datasets, that refer to the "Subject heading" access point, producing connected components of variant sizes presented by Table 5. The column named "Component size" indicates the number of the included search terms in each component, while the column named "Component number" corresponds to the amount of the generated connected components of a particular size.

Table 5. Components size and number per Collection on access point "Subject heading"

Component size	Component number	
	ARGOS	Public-School Libraries
1	50	56
2	21	31
3	11	9
4	13	5
5	6	4
6	3	2
7	2	2
8	1	
9	1	
14	1	
16	1	
22	1	
93	1	
960		1

A significant observation on the datasets is that most sessions include just one query in both collections. Therefore most derived components are of small size. In particular, the components having size more than 8 search terms are unique in both collections. Furthermore, an interesting phenomenon is the existence of one giant component, in both case studies. For ARGOS the giant component consisted of 93

search terms, while in the Public-School Libraries the giant component consisted of 967 search terms. These components do not seem to be homogeneous and meaningful. They are the union of sessions consisting of search terms with high appearance frequencies but they also include almost arbitrary terms with low occurrence frequencies, present in the same sessions. However, it is important to identify these special clusters and not treat them as normal cohesive communities.

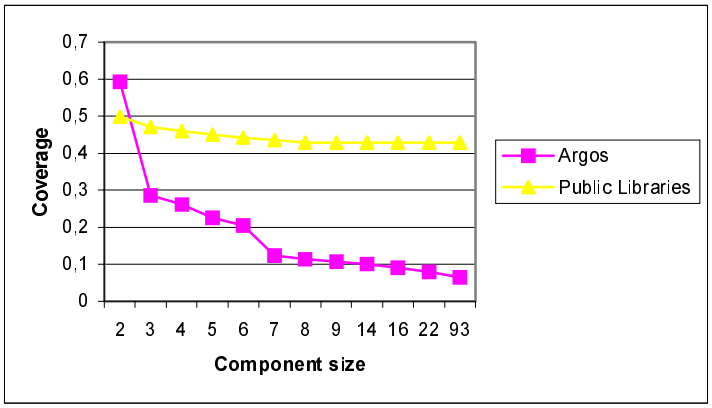

Figure 1. The coverage curves on access point "Subject heading" for the two collections

Concerning the search terms coverage we observe that it decreases as the size of the components increases. This behavior is to be expected since there are very few large components, which include only a small part of the search terms. Figure 1 shows the coverage decrease with respect to the component size. The Public-School Libraries curve is less steep than the ARGOS due to the usage of the "Subject heading" access point. As mentioned above, Public-School Libraries users are librarians and search the collections using the "Subject heading" access point more than average users do. So, the derived components complement each other and include more than the 40% of the search terms. On the other side ARGOS users are archaeologists and prefer the "Author" access point. Therefore ARGOS user sessions consist of few queries with very different search terms.

Table 6. A sample of user community models on access point "Subject heading"

Community	ARGOS Component size >= 3 Number of communities = 41	Public-School Libraries Component size >= 4 Number of communities = 14
1	Peloponnese, Medieval Peloponnese, Medieval Greece, Paleologos	Special education, Disabilities, Handicap, Children Disabilities, Higher Education, Education history
2	Asia Minor, Smyrna, Balkan Peninsula, Asia Minor –Economic conditions, Asia Minor -- history, Smyrna disaster 1922, Smyrna -- Economic conditions, Smyrna Fire 1922	Literature, Greek literature, Classical literature, Classical literature - - critics, Classical philology European literature

According to Figure 1, around the component size with value 3 (coverage value 28.5%) the ARGOS coverage curve becomes less steep, while the Public-School Li-

braries coverage curve becomes less steep around the component size of value 3 (coverage value 45.9%). This observation suggests the selection of this size value for the formation of the desired user community models. Hence our analysis leads to the selection of components with size greater than or equal to 3 search terms for the ARGOS collection, and more than 4 search terms for the Public-School Libraries Collection. Table 6 presents an indicative example of two user community models per each collection. The derived community models are disjoint sets of subject headings and constitute specific vocabularies, which describe the communities. These subject headings could be recommended and used by the digital library in a query expansion process.

5 Conclusions

Efficient and effective access to on-line information becomes increasingly critical as the amount of services for delivering information over the Internet is continuously increasing. The objective of this study was to analyze the behavior of the users of a digital library and to provide useful information to its administrator, in order to improve the provided services. The approach that we have adopted was to construct user communities, corresponding to groups of users with similar behavior. The end result was a behavioral pattern for each community, which provides much richer information to the administrator than the commonly used statistical figures about the usage of a service.

One of the important issues that have arisen in this work was the necessary engineering of the data collected. Query-based information retrieval services need various language engineering tools and domain-specific classification hierarchies, in order to reduce the dimensionality of the problem and generate the final data set.

Another important issue is the choice of the clustering method, which will construct the communities. Many graph theoretic methods have been used for the identification of user communities (Imafuji and Kitsuregawa 2002, Paliouras et al. 2002). The advantages of our approach are its simplicity and its computational efficiency. Component mining method generates disjoint communities, which map each user uniquely to a community. This might be desirable in services where each user community needs to be treated separately and data are sparse. However, it will not be desirable in many services, where users naturally belong to more than one community. In such cases the clique-mining algorithm is more suitable. Further research is also worthwhile, to evaluate other clustering algorithms.

Concluding we could say that the discovery of behavioral patterns for user communities, with the use of clustering methods, is feasible and can provide very valuable information to the administrator of a digital library.

References

Anderson, C.R. and E. Horvitz. 2002. Web Montage: A dynamic personalized start page. In *Proceedings 11th Intl. World Wide Web Conference*. ACM Press.

Basu, C., H. Hirsh and W. Cohen 1998. Recommendation as Classification: Using Social and Content-Based Information in Recommendation. In *Proceedings 15th National Conference in Artificial Intelligence (AAAI'98)*.

Billsus, D., and M. Pazzani. 2000. User Modeling for Adaptive News Access. *User Modeling and User-Adapted Interaction* 10:147-180.

Breese J.S., D. Heckerman and C. Kadie. 1998. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In *Proceedings 14th Conference on Uncertainty in Artificial Intelligence (UAI'98)*.

Bron, G. and J. Kerbosch. 1973. Finding all cliques of an undirected graph. *Communications of the ACM* 16:575-577.

Chakrabarti, S., M. van der Berg and B. Dom. 1999. Focused crawling: A new approach to topic-specific web resource discovery. In *Proceedings. 8th Intl. World Wide Web Conference*, 545-562.

Esposito, F., D. Malerba, G. Semeraro, N. Fanizzi and S. Ferilli. 1998. Adding Machine Learning and Knowledge Intensive Techniques to a Digital Library Service. *International Journal on Digital Libraries*, 2:3-19.

Imafuji, N. and M. Kitsuregawa. 2002. Effects of Maximum Flow Algorithm on Identifying Web Community. In *4th ACM CIKM International Workshop on Web Information and Data Management (WIDM'02)*.

Orwant, J. 1995. Heterogeneous Learning in the Doppelgänger User Modeling System. *User Modelling and User-Adapted Interaction*. 4:107-130.

Paliouras, G., C. Papatheodorou, V. Karkaletsis and C.D. Spyropoulos. 2002. Discovering User Communities on the Internet using Unsupervised Machine Learning Techniques. *Interacting with Computers Journal*. 14:761-791.

Payette, S.D. and O.Y. Rieger 1997. Z39.50 The User's Perspective. *D-Lib Magazine*, April 1997.

Pennock, D., E. Horvitz, S. Lawrence and C. Lee Giles. 2000. Collaborative filtering by personality diagnosis: A hybrid memory- and model-based approach. In *Proceedings 16th Conference on Uncertainty in Artificial Intelligence (UAI-2000)*, 473-480.

Perkowitz, M. and O. Etzioni. 1998. Adaptive Web Sites: Automatically synthesizing Web pages. In *Proceedings 15th National Conference in Artificial Intelligence (AAAI 98)*.

Peterson Bishop, A. 1995. Working toward an understanding of digital library use: a report on the user research efforts of the NSF/ARPA/NASA DLI projects. *D-Lib Magazine* October 1995.

Sfakakis, M. and S. Kapidakis. 2002. User Behavior Tendencies on Data Collections in a Digital Library, In *Proceedings 6th European Conference on Digital Libraries, (ECDL 2002)*, LNCS 2458, 550-559. Springer-Verlag.

Van House, N.A. et. al. 1996. User centered iterative design for digital libraries: the Cypress experience. *D-Lib Magazine* February 1996.