Large-Scale Mining of Usage Data on Web Sites

Georgios Paliouras,* Christos Papatheodorou,+ Vangelis Karkaletsis,* Panayotis Tzitziras,+ Constantine D. Spyropoulos,*

* Institute of Informatics and Telecommunications,
+ Division of Applied Technologies,
National Centre for Scientific Research (NCSR) "Demokritos",
15310, Aghia Paraskevi, Attikis, GREECE
*e-mail: {paliourg, vangelis, costass}@iit.demokritos.gr
+e-mail: {papatheodor, tzitziras}@lib.demokritos.gr

Abstract

In this paper we present an approach to the discovery of trends in the usage of large Web-based information systems. This approach is based on the empirical analysis of the users′ interaction with the system and the construction of user groups with common interests (user communities). The empirical analysis is achieved with the use of cluster mining, a technique that process data collected from the users' interaction with the Web site. Our main concern is the construction of meaningful communities, which can be used for improving the structure of the site as well as for making suggestions to the users at a personal level. Our case study on a site providing information for researchers in Chemistry shows that the proposed method provides effective mining of large usage databases.

Introduction

As the Web is expanding at an increasingly fast rate, embracing a large number of services, the issue of efficient information access is becoming a crucial factor in the design of Web sites. However, the manner in which a user accesses the information available on a Web site is heavily dependent on the user's needs, interests, knowledge and prejudices. As a result, the structure of the site should reflect the requirements of its users.

The first step towards providing efficient information access in a site is to understand its usage. This can be done by monitoring the daily usage of the site and analyzing the collected data. Commonly, the data that is collected by the administrators of various sites consists of general-purpose statistical figures, such as the number of users who access a particular page within certain periods of the day. This information can be useful in drawing a few general conclusions on the usage of a site, but does not facilitate the adaptation of the site to the needs of the users.

In this paper we examine an alternative, more personalized approach to the collection and analysis of usage data. This approach is based on the analysis of access logs, which record the date and time each page is accessed, as well as the IP number of the visitor. We organize the access-log information in sessions, each of them providing a navigational pattern, associating a set of pages in the site, and then we construct *community models*. A community (Orwant, 1995) is a group of users with common navigational behavior and the community model describes the common features in the behavior of the users. The construction of communities is done with the use of the Cluster Mining algorithm (Perkowitz & Etzioni, 1998).

The work presented in this paper builds upon previous work of ours on the construction of user communities from usage data on various information services on the Internet (Paliouras et al. 1998, 1999a, 1999b). The main difference with our previous work is the scale of the problem. Previously, we had only examined a small Web site, consisting of 41 pages (Paliouras et al. 1999b), while now we are looking at a site with a few thousand pages and very high hit rate. The increase in scale introduces a number of important issues, concerning feature selection, scalability of the clustering algorithms and interpretation of the results. We are addressing each of these issues individually, introducing new ideas to our method for constructing community models, with the aid of clustering algorithms.

The first issue that we address is that of data engineering, which includes selecting the right representation for the training data and reducing the dimensionality of the problem. Regarding the representation of the data, we have seen in our previous work (Paliouras et al. 1999b) that representing access sessions by means of transitions between pages produces interesting navigation patterns for the community models. This representation is used again here, but we combine it with a simpler representation, where access sessions are represented as bags of pages.

Another issue, in which we pay substantial attention, is the characterization of the community models, i.e., the construction of meaningful communities that are useful for the system administrator. Ideally we would like to be able to construct a prototypical model for each community, which is representative of the participating users and significantly different from the users of other communities.

Such *community descriptions* can be used to:

- introduce structure or re-organise the existing structure of an information service,
- make suggestions at a personal level to the users within a specific community,
- support the expansion strategy for the service, etc.

Section 2 of the paper presents the algorithm that is used for the construction of communities and discusses the problem of constructing meaningful communities. Section 3 presents the experimental setting and results and section 4 summarises the presented work, introducing our plans for the future.

Learning User Communities

Learning from navigation patterns

The most effective way to learn about the use of a Web-based information service and draw conclusions, which may help to improve it, is through the direct analysis of the users' interaction with it (i.e., user queries and/or navigation patterns). A number of interesting attempts to achieve such analysis have been done lately, in the context of analysing usage data on the Web, using machine learning methods.

Machine learning has mainly been used for acquiring models of individual users interacting with an information system, e.g. (Bloedorn, Mani and MacMillan, 1996; Chiu, 1997; Raskutti and Beitz, 1996). In such situations, the use of the system by an individual is monitored and the collected data are used to construct the model of the user, i.e., his/her individual requirements.

We are concerned with a higher level of generalization of the users' interests: the construction of user communities. One approach to the construction of user communities is by generalizing the properties of user models. This approach requires the application of unsupervised learning techniques to the data, i.e., the users' characteristic features, which in our case are the visited Web pages.

Unsupervised learning tasks have been approached by a variety of methods, ranging from statistical clustering techniques to neural networks and symbolic machine learning. In this work, we have opted for the statistical learning methods. The statistical learning algorithm used here is a variant of the *cluster mining* algorithm used in PageGather (Perkowitz & Etzioni, 1998). The cluster mining algorithm in PageGather is applied to Web access trails, i.e., it translates the access trails into a graph and searches for highly connected subgraphs.

Cluster Mining

The cluster mining algorithm that we use here discovers patterns of common behaviour, by looking for all cliques in a graph that represents the users' characteristic features. We start by constructing a weighted graph $G(V,E,W_V,W_E)$. The set of vertices V corresponds to the *users' characteristic features*. The set of edges E corresponds to the combination of the users' characteristics as they are observed in their interaction with the system. For instance, if the Web site is a library on Chemistry, and the user visits pages concerning "Organic Chemistry" and "Polymers" we create an edge between the relevant vertices (Figure 1). The weights on the vertices W_V and the edges W_E are computed as the frequencies of the users' interests and their combinations respectively.

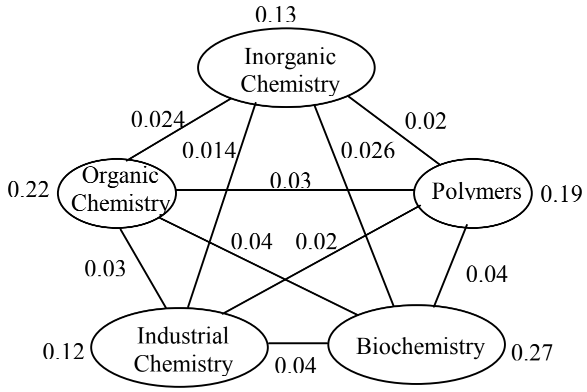

Figure 1. Non-normalised graph.

Edge frequencies are normalised by dividing them with the maximum of the frequencies of the two vertices that they connect. The effect of normalisation is to remove the bias for characteristics that appear very often in all users. According to the previous example, the resulting normalised graph is given in Figure 2.

The connectivity of the resulting graph is usually very high. For this reason we make use of a *connectivity threshold* aiming to reduce the edges of the graph. In our example in Figure 2, if the threshold equals 0.1 the edge ("Inorganic", "Biochemistry") is dropped.

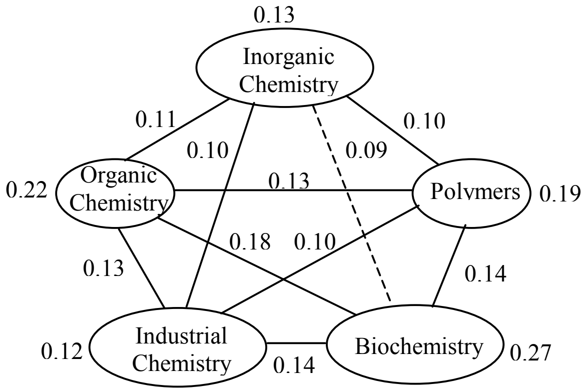

Figure 2. Normalised graph.

Despite the large theoretical complexity of the clique-finding problem, in practice the algorithm that has been implemented (Bron & Kerbosch, 1973) is fast.[1] The efficiency of the algorithm allowed a full investigation of the effect of the connectivity threshold.

The algorithm that we use differs in two ways from the cluster mining algorithm used in the PageGather system: (a) PageGather does not normalise the weights W_E and (b) it restricts its search to cliques of size k and to connected components.

Cluster mining does not attempt to form independent user groups, but the generated clusters group together characteristic features of the users directly. If needed, a user can be associated with the clique(s) that best match the user's behaviour. Alternatively each user can be associated probabilistically with each of the cliques. As mentioned above, the focus of our work is on the discovery of the behavioural patterns of user communities. For this reason, no attempt is made to match individual users with the cliques generated by the cluster mining algorithm.

Meaningful Communities

In contrast to the common clustering methods, such as COBWEB (Fisher, 1987) and Autoclass (Hanson, Stutz and Cheeseman, 1991), the clusters generated by cluster mining group together characteristic features of the users directly. Each clique discovered by the cluster mining algorithm, is already a navigational pattern. This is an important advantage of this mining approach.

In order to examine the expressiveness of the communities produced by the cluster mining algorithm, we varied the *connectivity threshold*, and measured the following two properties of the generated community descriptions:

Coverage: the proportion of features covered by the descriptions. Some of the features will not be covered, because their frequency will not have increased sufficiently. In order for this to happen in the proposed method, which generates *all* cliques in a graph, we need to ignore singleton cliques.

Overlap: the extent of overlap between the constructed descriptions. This is measured simply as the ratio between the total number of features in the description and the number of distinct features that are covered.

Case Study

Data Engineering

For this experiment, we used the access logs of the site "Information Retrieval in Chemistry" (http://macedonia.chem.demokritos.gr), which consists of a few thousand pages with a high hit rate. The log files consisted of 137,150 Web-server calls (log file entries) and covered the period between January and August 1999. Each log entry recorded a visitor's access date and time, its computer IP address and domain name, and the target page (URL).

In order to construct a training set for the clustering algorithm, the data in the log files passed through two stages of pre-processing:

1. Access sessions were extracted.
2. The paths recorded in the access sessions were translated into feature vectors.

Extracting access sessions from log files is a less deterministic process than one initially would imagine. This process involves the following stages:

1. Grouping the logs by date and IP address.
2. Selecting a time-frame within which two hits from the same IP address can be considered to belong to the same access session.
3. Grouping the pages accessed by the same IP address within the selected time-frame to form a session.

In order to select the appropriate time-frame, we generated the frequency distribution of the page transitions in minutes. According to this distribution, transitions from one page to another, made with a time interval longer than one hour, had very low frequency. Thus, a sensible definition of the *access session* is a sequence of page transitions for the same IP address, where each transition is done at a time interval smaller than one hour. Based on this definition, our log files consisted of 11,893 access sessions.

Concerning the translation of access sessions to feature vectors, we examined two alternative approaches. In the first approach each feature in the feature vector represented the absence or presence of a particular page of the Web site in the session. In the second approach, we used transitions between pages, rather than individual pages as the basic path components. There were 1,027 pages in the site that were visited at least once. Clearly the number of all possible transitions between these pages is prohibitively large. Even the number of different transitions that appear in the log files is very large. Thus, we needed a method to reduce the number of features in both experiments. This reduction was achieved by examining the frequency distributions of the pages and transitions from one page to another. The two distributions were highly skewed, i.e., there was a small number of very frequent pages and transitions. Thus, we decided on a cut-off frequency of 30 for pages and 20 for transitions, which were the points where the corresponding distributions were becoming flat. Additionally we removed all transitions from a page to itself. As a result, 229 pages and 251 transitions survived this selection and were used to form the binary feature vectors. We also tried a method that uses Mutual Information, as a criterion for selecting features for unsupervised learning (Sahami, 1997). More than 90% of the features selected by this method, were within the high-frequency range that we selected. However, some of the features that were eliminated were clearly important. For

[1] It generates all cliques (approx. 200) of a large graph (239 vertices), with an average clique size of 100 vertices, in about 5 mins on a common SparcServer.

this reason, we opted for the simple frequency-threshold approach.

The cluster mining algorithm was applied to both representations of the data. In the first representation, the binary vector corresponds to a sequence of pages that the user has visited in one session. The cluster mining algorithm constructs groups of pages, which often co-occur in access sessions. In the second representation, each binary vector corresponds to a sequence of page transitions. In this case, cluster mining provides clusters of co-occurring page transitions. Some of the desired results in this case study are: paths that are commonly followed by different groups of people, pages that people often visit in the same session through different paths, etc. Such information is valuable for the adaptation of the site to the interests of the users.

Results

Depending on the value of the connectivity threshold, the average size and the number of cliques generated by the cluster mining algorithm varied. Figures 3 and 4 display the results of those experiments for both representations of the data, i.e., pages and transitions.

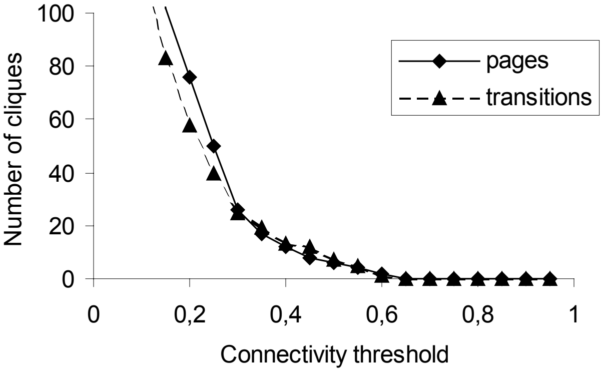

Figure 3. Number of cliques for different values of the connectivity threshold.

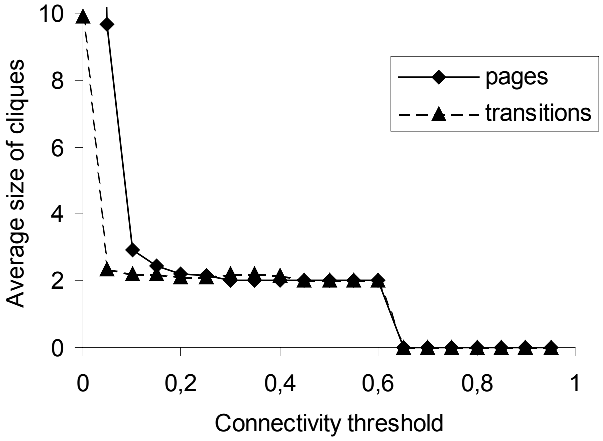

Figure 4. Average clique size for different values of the connectivity threshold.

For small values of the threshold, the graph is highly connected and contains many large cliques. It is really above the threshold value 0.2, that the number of the cliques drops to manageable levels. At that level, almost all cliques are pairs of pages or transitions. The first unexpected observation in these data is the close proximity of the curves for the two different representations. One explanation for this can be given by considering the organization of the Web site. The site is organized roughly as a tree, which is very wide, but shallow. This means, that there are many pages at the same conceptual level, corresponding to different areas of chemistry. Given that the visitor, can easily move from any of these pages to any other, the concept of navigational patterns between pages at the same conceptual level becomes very weak. As a result, frequent transition sequences become equivalent to frequent sets of pages, corresponding both to the characteristic interests of different communities. The second interesting observation is the sharp fall in the average size of the cliques for both representations. The result of this phenomenon is that the associations found between the pages are really co-occurrence patterns, rather than substantial groups of pages. This observation justifies the choice of a flat hierarchy for such a large Web site.

In addition to the number of cliques and their average size, we examined the coverage of the cliques and the overlap between them. Figures 5 and 6 present the results along those two dimensions. As expected, the overlap for small threshold values is very large, due to the large number of very large cliques. However, it falls sharply, following the sharp fall in the size of the cliques. The coverage of the cliques falls at a much lower pace. The result of this is that around the threshold value 0.2, about half of the pages and the transitions appear in the cliques, while there is little overlap between the cliques. In other words, the pairs of pages and transitions that are grouped together by the algorithm at that threshold level seem to be quite distinct and as a result they cover a large proportion of the original features. This observation suggests the

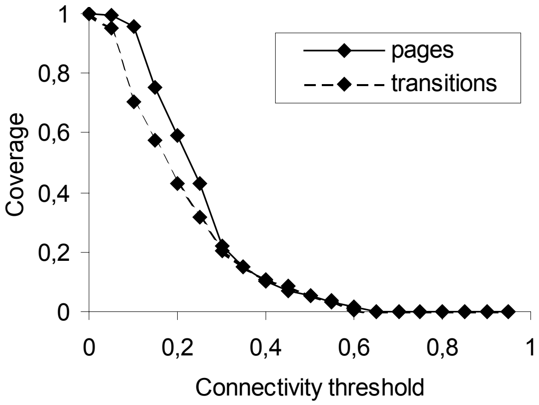

Figure 5. Coverage of cliques for different values of the connectivity threshold.

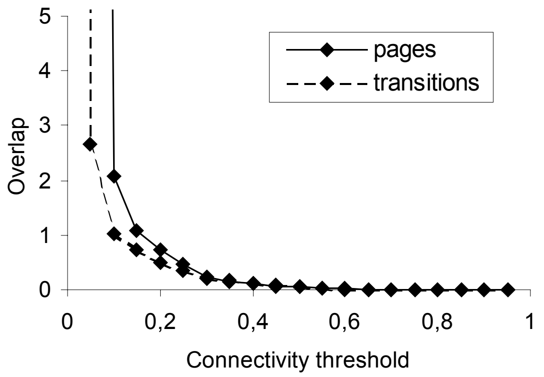

Figure 6. Overlap between cliques for different values of the connectivity threshold.

selection of this threshold value, for the formation of the communities, which then need to be examined further by the administrator of the site.

Conclusions

Efficient and effective access to on-line information becomes increasingly critical as the amount of information that becomes on-line increases at an overwhelming pace. Web-based information systems constitute a vital component of this new style of information exchange and user modelling technology can facilitate considerably the access to them. This is a much more informative analysis than the simple Web usage statistics that are commonly used by system administrators. The results obtained by a Web usage analysis, can be used to modify the structure of a Web site, in reaction to the interests of the visitors and/or make the site adaptive to different types of visitors.

We have examined two different representations of the access sessions, as simple bags of pages and as sets of page transitions. Higher-order representations, i.e., longer sequences of page transitions, may be interesting, but are likely to increase the dimensionality and reduce radically the density of the training data. This can have a seriously negative effect on the ability of the learning algorithms to generalize.

Finally, the work presented here could be extended in several ways. We are currently comparing the cluster mining method with clustering methods, including Autoclass (Hanson, Stutz and Cheeseman, 1991) and the neural clustering module in the IBM Intelligent Miner^{TM} package. The same Web site is used as a testbed for this comparison. Our longer-term plans focus on the next step to the approach presented here, i.e., the use of our results for the personalization of a Web site. The results presented here show that the discovery of behavioural patterns for user communities, with the use of cluster mining, is feasible even for large sites and can provide very valuable information to the Web-site administrator.

Acknowledgments

We would like to thank the coordinator of the team "Information Retrieval in Chemistry", Dr. E. Varveri, as well as the members of the team Mr. A. Varveris and Mr. P. Telonis, for the dataset used in our case study.

References

Bloedorn, E.; Mani, I.; and MacMillan, T.R. 1996. Machine Learning of User Profiles: Representational Issues. In *Proceedings of the National Conference on Artificial Intelligence*, 433-438: AAAI Press.

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

Chiu, P. 1997. Using C4.5 as an Induction Engine for Agent Modelling: An experiment of Optimisation. In *Proceedings of the International Conference on User Modelling,* Workshop on Machine Learning for User Modelling.

Fisher, D. 1987. Knowledge Acquisition via Incremental Conceptual Clustering, *Machine Learning,* 2:139-172.

Hanson, R., Stutz, J., Cheeseman, P. 1991. Bayesian Classification Theory, Technical Report FIA-90-12-7-01, NASA Ames Research Center, Artificial Intelligence Branch.

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

Paliouras, G.; Papatheodorou, C.; Karkaletsis, V.; Spyropoulos, C.D.; and Malaveta, V. 1998. Learning User Communities for Improving the Services of Information Providers. In *Proceedings of the Second European Conference on Digital Libraries*, 367-384. Heraklion, Greece: Lecture Notes in Computer Science, n. 1513, Springer-Verlag.

Paliouras, G.; Karkaletsis, V.; Papatheodorou, C.; and Spyropoulos, C.D. 1999a. Exploiting Learning Techniques for the Acquisition of User Stereotypes and Communities. In *Proceedings of the Seventh International Conference on User Modeling*, 169-178. Banff, Canada: CISM Courses and Lectures, n. 407, Springer-Verlag.

Paliouras, G.; Papatheodorou, C.; Karkaletsis, V.; Spyropoulos, C.D.; and Tzitziras, P. 1999b. From Web Usage Statistics to Web Usage Analysis. In *Proceedings of the IEEE Conference on Systems Man and Cybernetics*, II-159-164. Tokyo, Japan: IEEE Press.

Perkowitz, M.; and Etzioni, O. 1998. Adaptive Web Sites: Automatically synthesizing Web pages. In *Proceedings of the Fifteenth National Conference in Artificial Intelligence*. Madison, Wisconsin, MW: AAAI Press.

Perkowitz M.; and Etzioni O. 1999. Adaptive Web Sites: Conceptual Cluster Mining. In Proceedings of the Sixteenth International Joint Conference in Artificial

Intelligence, 264-269. Stockholm, Sweden: International Joint Conferences on Artificial Intelligence, Inc.

Raskutti, B.; and Beitz, A. 1996. Acquiring User Preferences for Information Filtering in Interactive Multi-Media Services. In *Proceedings of the Pacific Rim International Conference on Artificial Intelligence*, 47-58.

Sahami, M. 1998. Using Machine Learning to Improve Information Access, *Ph.D. Thesis*, Department of Computer Science, Stanford University.