Model-Based Cluster Analysis for Web Users Sessions

George Pallis, Lefteris Angelis, and Athena Vakali

Department of Informatics,
Aristotle University of Thessaloniki,
54124, Thessaloniki, Greece
gpallis@ccf.auth.gr
{lef, avakali}@csd.auth.gr

Abstract. One of the main issues in Web usage mining is the discovery of patterns in the navigational behavior of Web users. Standard approaches, such as clustering of users' sessions and discovering association rules or frequent navigational paths, do not generally allow to characterize or quantify the unobservable factors that lead to common navigational patterns. Therefore, it is necessary to develop techniques that can discover hidden and useful relationships among users as well as between users and Web objects. *Correspondence Analysis* (CO-AN) is particularly useful in this context, since it can uncover meaningful associations among users and pages. We present a model-based cluster analysis for Web users sessions including a novel visualization and interpretation approach which is based on CO-AN.

1 Introduction

The explosive growth of the Web and the increased number of users have led more and more organizations to put their information on the Web and provide sophisticated Web-based services such as distance education, on-line shopping etc. However, the continuous growth in the size and use of the Internet is increasing the difficulties in managing the information. Thus, an urgent need exists for developing new techniques in order to improve the Web performance.

In this context, cluster analysis can be considered as one of the most important aspects in the Web mining process for discovering meaningful groups (interesting distributions or patterns over the considered data sets) as well as interpreting and visualizing the key behaviors exhibited by the users in each cluster. The clustering problem is about partitioning a given data set into clusters (groups) such that the data points in the same cluster are more similar to each other than points in different clusters.

Here, we focus on clustering Web users based on their navigation behavior. Specifically, a Web user may visit a Web site from time to time and spend arbitrary amount of time between consecutive visits. To deal with the unpredictable nature of Web browsing, a new concept, *session*, was introduced as the unit of interaction between a user and a Web server. By clustering the users navigation sessions, the Web developer may understand the browsing behavior of users better and may provide more suitable and customized services to the users. In particular, the knowledge discovered from these sessions will certainly contribute in the construction and maintenance of real-time intelligent Web servers that are able to dynamically adapt their designs to satisfy the future users' needs.

M.-S. Hacid et al. (Eds.): ISMIS 2005, LNAI 3488, pp. 219–227, 2005.

Therefore, understanding how users navigate a Web site is an essential step for Web sites developers to customize content (generating pages on a user's previous activities) and to consider creative caching and prefetching schemes to deliver content as quickly as possible.

The main problem with clustering algorithms is that it is difficult to assess the quality of the clusters returned and interpret these results by extracting useful inferences for the users' navigation behavior. Therefore, in most applications the resulting clustering scheme needs some evaluation regarding its validity [10]. Evaluating and assessing the results of a clustering algorithm is the main challenge of cluster validity. Up to now, several clustering validation approaches have been proposed in the literature [4, 8, 9, 10, 11].

The purpose of this paper is to present a comprehensive methodology including clustering at users' sessions, evaluation of clustering results and interpretation of the results. All these steps use advanced statistical methods which help not only to evaluate the clustering scheme but also to discover useful associations among clusters.

In [11], we introduced a probabilistic validation algorithm for model-based clustering, which is based on the χ^2 statistic. Here, we further proceed into presenting a detailed analysis for model-based clustering scheme[1] on a busy Web server and propose an efficient interpretation and visualization technique in order to extract useful conclusions for the underlying clusters. The main novelty in our work is that we find the equilibrium distribution of Web users clusters, each considered as a Markov chain (which represents the probability of a user to access a particular Web page in infinite number of clicks, independently of its previous pages that he has visited) and then we apply on them a correspondence analysis in order to further interpret the clustering results, uncovering meaningful associations among users and Web pages. More specifically, the main technical contributions of this work can be summarized as follows:

- We present a detailed framework for model-based cluster analysis for Web users sessions, studying the equilibrium distribution of each cluster;
- We introduce a visualization method for interpreting the equilibrium distribution of clusters based on correspondence analysis;
- The methods described were tested on a real data set collected from a busy Web server (msn.com).

To the best of our knowledge, there are not previous research works dealing with the interpretation and visualization of model-based clustering schemes using the concept of correspondence analysis [7]. Existing works on model-based clustering largely concentrate on a specific model or application, without providing a visualization method. The main reason is the high complexity which have these implementations [1]. A notable exception is the work in [2] where an interesting visualization tool was presented.

The rest of the paper is organized as follows. In Section 2, we present in detail the framework for model-based cluster analysis for grouping the Web users' sessions. The experimental results are given in Section 3. Finally, we conclude the paper and give some remarks for future research.

[1] Each cluster is represented by a probability model and the sessions are grouped according to the order in which users request Web pages.

2 The Clustering Procedure

The procedure for clustering Web users sessions is not a straightforward task but it consists of 4 steps. The first step is the preprocessing of Web log files[2] in order to mine the Web users' sessions. Then, we assign the sessions into the appropriate clusters. The third step is to validate the clusters which have been created in the previous step. Finally, the last step of the clustering procedure is to visualize and interpret the clusters. In the next paragraphs, we present in detail the above procedure.

2.1 Web Data Preprocessing

Web log data are undergone a certain pre-processing, such as invalid data cleaning and session identification. Data cleaning removes the records which do not include useful information for the users' navigation behavior, such as graphics, javascripts etc. The remaining page requests are categorized into different categories. The process of grouping the Web pages into categories is a usual practice, since it improves the data management and in addition eliminates the complexity of the underlying problem (since the number of page categories is smaller than the number of Web pages in a Web site) [2, 11]. In particular, the individual pages are grouped into semantically similar groups (as determined by the Web site administrator).

Moreover, we use heuristic methods to identify the Web access sessions, based on IP and time-outs [3]. We consider that we have an ordered set of traces with respect to the IPs. Therefore, a new session is created when a new IP address is encountered or if the visiting page time does not exceed 30 minutes for the same IP address.

2.2 The Clustering Algorithm

The way that users navigate in a Web site depends on several factors such as user's interest, site structure etc. In this framework, we assume that a user arrives at the Web site in a particular time and is assigned to one of the underlying clusters with some probability. In the next paragraphs, we present the model-based approach which we follow in order to cluster the Web users' sessions:

- **Representation of clusters:** Each cluster consists of Web users' sessions. To model heterogeneity of them we use a mixture of first-order Markov chains, where each cluster in a mixture represents a behavior described by a single Markov chain. Specifically, Markov models can be viewed as stochastic generalizations of finite-state automata, when both transitions between states and generation of output symbols are governed by probability distributions. In our framework, a useful representation of such a model is the transition matrix of size V×V (where V is the number of categories which are denoted by A_1, A_2, ..., A_V) describing the probability that a user will go to a certain page category given (s)he is viewing the current page category and a vector of V initial probabilities describing how likely is that a user will begin his/her navigation session in a given page category.

[2] Web log files provide information about activities performed by a user from the moment the user enters a Web site to the moment the same user leaves it.

- **Clustering users' sessions:** Each cluster has a data-generating model with different estimate parameters for each one. Therefore, this model can be well defined, if only we estimate the parameters of each model component, the probability distribution used to assign users to the various clusters and the number of components. We cluster users' sessions by learning a mixture of first-order Markov models using the EM algorithm. The EM algorithm originates from [5] and in [2] a method for employing on EM on users' sessions is proposed. It alternates two steps:
  - **The expectation E-step**: Given a set of parameter estimates the E-step calculates the conditional expectation of the complete-data log likelihood given the observed data and the parameter estimates.
  - **The maximization M-step**: Given a complete-data log likelihood, the M-step finds the parameter estimates to maximize the complete-data log likelihood from the E-step.

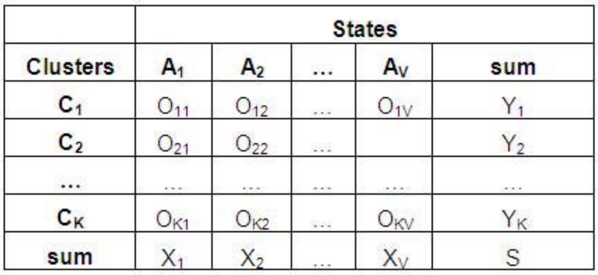
  The two steps are iterated until convergence (i.e. a local optimal solution is reached). Concerning the complexity of the EM algorithm, it depends on the complexity of the E and M steps at each iteration. For example, in our case (Markov mixtures) the complexity is linear in the sum of the lengths of all sessions. Note, that for more complex mixture models the complexity can be higher.
- **Number of clusters:** The number of clusters may be determined by using several probabilistic criteria, such as BIC (Bayesian Information Criterion), bayesian approximations, or bootstrap methods [1, 6]. Formally, we assume that there are K clusters C_1, C_2, ..., C_K and each of them is generated from its own probability distribution. Once the model is specified, we use the EM algorithm and probabilistic out-of-sample evaluation to determine the best number of clusters. A model is fitted on a subsample of sessions (the so-called training data set) and then scored on the remaining data (the so-called testing data set). Thus, we get an objective measure of how well each model fits the data. The model with the minimum out-of-sample predictive log score is selected.

2.3 Cluster Validation

An important issue in cluster analysis is the evaluation of clustering results to find the partitioning that best fits the underlying data. Towards this direction, we propose an efficient validation technique for model-based clustering approaches, where each cluster is represented by an ergodic Markov chain. By the term "ergodic", we mean a Markov chain that has the following two properties: 1) each node can reach any other node (all states intercommunicate), and 2) the chain is not periodic (all states have period one).

In order to validate the clustering scheme, we consider the equilibrium distribution of each cluster produced by the algorithm. These distributions represent the probabilities of a user to access each state in infinite number of states independently of its initial state. Then, the validation is performed by testing the homogeneity of the equilibrium distribution by the χ^2 test. More specifically, χ^2 testing is used to test the homogeneity among multiple clusters with probabilistic distributions by constructing a contingency table. This statistic is used to assess evidence that two or more distributions are dissimilar. Considering that in model-based approach the clusters represent a probabilistic distribution, we can directly apply the test of homogeneity by fitting the state frequen-

Table 1. A Contingency Table for Chi-square Testing



	States				
Clusters	**A_1**	**A_2**	**...**	**A_V**	**sum**
C_1	O_{11}	O_{12}	...	O_{1V}	Y_1
C_2	O_{21}	O_{22}	...		Y_2
...	...	...	...	...	...
C_K	O_{K1}	O_{K2}	...	O_{KV}	Y_K
sum	X_1	X_2	...	X_V	S

cies in the cluster into the contingency table, which rejects the fact that our modeling simplifies the testing.

In our framework, the states represent the page categories. Table 1 is the contingency table for testing. A contingency table test (or test of independence) is one that tests the hypothesis that the data are cross-classified in independent ways. In particular, O_{ij} stands for the frequency of A_j state in cluster C_i. O_{ij} is computed by multiplying the relative frequency of A_j state with the number of sessions that belong to cluster C_i. X_i is the sum of all the O_{ij} in i-th column and Y_j is the sum of all the O_{ij} in j-th raw. In this framework, we want to test the following hypothesis (for all the states and clusters of the underlying model):

Null Hypothesis (Ho): The distributions of the states in each cluster are all the same.
Testing: The following equation computes the χ^2 statistic:

$$\chi^2(C_1, C_2, ..., C_K)= \sum_{i=1}^{K}\sum_{j=1}^{V} \frac{(O_{ij}-Y_i\times\frac{X_j}{S})^2}{Y_i\times\frac{X_j}{S}}$$

A large value of the χ^2 criterion shows that the equilibrium distributions for each cluster are significantly different, which in turn is an indication of the heterogeneity among clusters. Therefore, we should know a critical χ^2 value that is the boundary of the area of hypothesis' s rejection in a contingency table test. In order to find this critical value, we should define the level of significance (α) and the degrees of freedom (df). In statistics, it is known that a χ^2 has asymptotically a χ^2 distribution with (K-1)×(V-1) df. Therefore, if the value of χ^2 distribution is greater than a critical value, such as $\chi^2_{(K-1)\times(V-1);\alpha}$, we can reject the Ho at the α level of significance. Otherwise, there is no much evidence to reject Ho.

2.4 Cluster Interpretation and Visualization

Interpreting the navigation behaviors exhibited by the Web users in each cluster is important for a number of tasks, such as managing the Web site, identifying malicious visitors, and targeted advertising. It also helps to understand the navigation patterns of different user groups and therefore helps in organizing the Web site to better suit the users' needs. Furthermore, interpreting the results of clusters contributes to identify and provide customized services and recommendations to Web users.

Here, we introduce a model-based clustering interpretation approach by analyzing the contingency tables, that have been occurred in validation process, as described in the previous Section. In particular, an analysis of these tables includes examining row and column discrete variables and testing for independence via the χ^2 statistic. However, the number of variables can be quite large, and the χ^2 test does not reveal the dependence structure. Thus, in order to analyze the clusters in a more efficient way, we propose to use the correspondence analysis method.

Correspondence analysis (CO-AN) is a standard multi-variate statistical analysis method aiming to analyze and visualize simple two-way and multi-way contingency tables containing some measure of correspondence between the rows and columns. Thus, one of the goals of CO-AN is to describe the relationships between the categories for each variable, as well as the relationship between the variables. In this context, CO-AN can be used in order to interpret and visualize the Web users' navigation behaviors.

The rows and the columns of the contingency table as described in Table 1 represent distributions of the categorical variables. In order to measure their similarity and depict them geometrically, CO-AN uses a distance based on the χ^2 statistic. The object of CO-AN is to explain the total variation in the underlying correspondence table. In essence, the correspondence map, that is occurred by this analysis, is a graphical tool which helps the researcher easily to notice relationships within this table. When interpreting a correspondence map it is often helpful to refer back to the original correspondence table. Therefore, CO-AN can be proved a very useful tool for Web site developers since it could be used to graphically provide a meaningful interpretation of clusters as well as, the relationship among users' navigation behaviors. For instance, they might find which Web page categories are associated with each others, with respect to users' sessions. Then, the Web developer may arrange the structure of the Web site so that these pages are interlinked together.

CO-AN may also have several applications in commercial Web sites [1]. CO-AN of Web users sessions through e-commerce Web sites can provide valuable insights into customer behavior and provide clues about whether improvements in site design might be useful. Moreover, CO-AN method may be used to recommend new products to Web site visitors, based on their browsing behavior. It may also be used to understand what factors influence the way customers make purchases on a Web site.

3 Experimental Evaluation

3.1 Data Set

The methods described were tested on a real data set. In particular, the data set comes from Internet Information Server (IIS) logs for msnbc.com and news-related portions of msn.com[3].

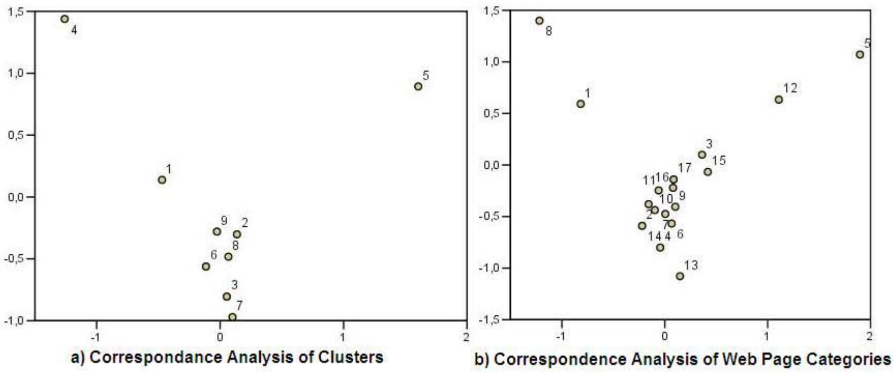
In our experiments, there are 17 categories, which are presented in detail in Figure 2. Each category includes a number of URLs and the data set consists of approximately 6,000 users' sessions, with an average of 5,7 page views per session. The number of URLs per category ranges from 10 to 5000. Then, we select some of the sessions as (80% of

[3] Msnbc.com anonymous web data: http://kdd.ics.uci.edu/databases/msnbc/msnbc.html.

the total data) training data set and the rest as testing data set in order to determine the number of clusters.

3.2 Cluster Analysis: Validation and Interpretation

The first task is to define an optimal value for the number of clusters. As we mentioned in the previous Section, we choose the number of clusters minimizing the out-of-sample predictive score. We tested several out-of-sample log-likelihoods for varying number of clusters and found that the minimum value is achieved for the choice of 9 clusters. Then, we cluster the users' navigation sessions as described in Section 3. Based on our data set, the result is χ^2= 7040,3 with 128 df. Considering that the $\chi^2_{128;0,005}$=127.81, we conclude that the resulted clusters are an effective choice.



a) Correspondance Analysis of Clusters

b) Correspondence Analysis of Web Page Categories

Fig. 1. Correspondence Analysis of Web Page Categories

Once the clustering algorithm has been described, it is now time to obtain some graphs and interpret their behavior. In this context, Figure 1a illustrates which clusters have common characteristics with each other. As can be seen by this Figure, each cluster is represented by a point. Thus, we observe that the users identified as belonging to clusters 2 and 8 have some common characteristics (i.e. their distances are sufficiently small) while the users belonging to clusters 4 and 5 seem to be different (i.e. the points representing these clusters are isolated).

CO-AN can also be used to graphically display the relationship among Web page categories. As we described above, Web requests are not recorded at the finest level of detail (at the level of URL) but they are rather recorded at the level of page category (as determined by the Web site administrator). Figure 1b illustrates the associations between Web page categories that have occurred using CO-AN. From this Figure, we observe that Web users who visit Web pages about "weather" they do not visit Web pages about "opinion" (the points representing these categories are isolated). In this framework, we also find that the Web pages about "bulletin board service (bbs)" are usually directly associated with Web pages about "travel".

Therefore, a deep knowledge for the inside of each cluster can draw useful and meaningful inferences for the users' navigation behavior. More specifically, Figure 2 depicts

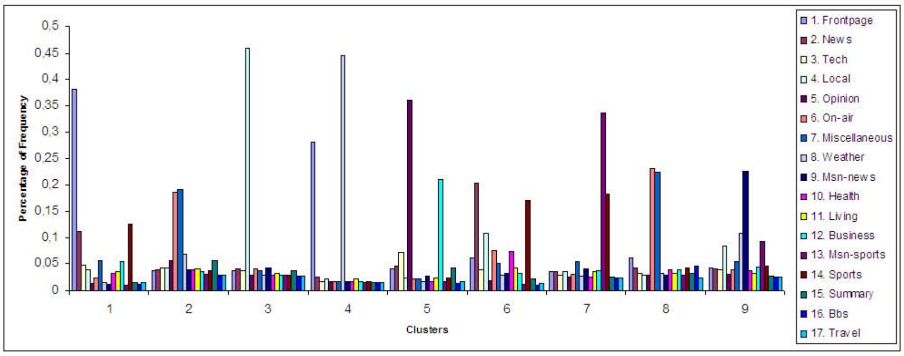


Fig. 2. The Percentage Frequency of Web Page Categories for each Cluster

the percentage frequency of requested Web page categories that have been observed for each cluster. For instance, in cluster 3, the majority of the requested Web page categories belongs to "local". On the other hand, the users in cluster 5 show high interest only for the categories which refer about "opinion" and "business". Furthermore, Figure 2 depicts that the users in clusters 2 and 8 have quite similar navigation behavior, since they usually visit Web pages about "on-air" and "miscellaneous".

4 Conclusions-Future Work

This paper presents a framework for model-based cluster analysis for Web users sessions. Taking into consideration that the Markov models may provide valuable information for users' navigation behavior which is often hidden, it is necessary to develop techniques that can discover hidden meaningful relationships among users as well as between users and Web objects. Towards this direction, statistical analysis helps to explore this hidden information in order to enhance the Web performance. Therefore, CO-AN is particularly useful in this context, since it can uncover semantic associations among users and pages.

For the future, we plan to apply the whole methodology described here in a Web log file in order to further analyze and interpret the results. Another goal is to develop an automatic mechanism in order to deliver the appropriate content to the interested users in a timely, scalable, and cost-effective manner.

References

1. Baldi P., Frasconi P., Smyth P. Modeling the Internet and the Web. Wiley Press,USA, 2003.
2. Cadez I. V., Heckerman D., Meek C., Smyth P., White S. Model-based Clustering and Visualization of Navigation Patterns on a Web Site. Journal of Data Mining and Knowledge Discovery, 7 (4), pp. 399-424, 2003.
3. Chen Z., Fu A., Tong F. Optimal Algorithms for Finding User Access Sessions from very Large Web Logs. World Wide Web: Internet and Information Systems, 6, pp. 259-279, 2003.
4. Chen K., Liu L. Validating and Refining Clusters via Visual Rendering. *In Proc. of the International Conference on Data Mining (ICDM 2003)*, Melbourne, Florida pp. 501-504, Dec. 2003.

5. Dempster A. P., Lsird N. M., Rubin D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. Statistics Society B, 39, pp. 1-22, 1977.
6. Fraley C., Raftery A. How Many Clusters? Which Clustering Method? Answers via Model-based Cluster Analysis. Computer Journal, 41, pp. 578-588, 1998.
7. Greenacre M.J. Correspondence Analysis in Practice. Academic Press, 1993.
8. Gunter S., Bunke H. Validation Indices for Graph Clustering. Pattern Recognition Letters 24 (8), pp. 1107-1113, 2003.
9. Halkidi M., Batistakis Y., Vazirgiannis M. On Clustering Validation Techniques. J. Intell. Inf. Syst. 17 (2-3), pp. 107-145, 2001.
10. Kohavi R. Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *In Proc. of the 14th International Joint Conference on A.I., Vol. 2*, Canada, 1995.
11. Pallis G., Angelis L., Vakali A., Pokorny J. A Probabilistic Validation Algorithm for Web Users' Clusters. *In Proc. of the IEEE International Conference on Systems, Man and Cybernetics (SMC 2004)*, Hague, Holland, 2004.