# A Web Personalization System based on Web Usage Mining Techniques[*]

Massimiliano Albanese
malbanes@unina.it

Antonio Picariello
picus@unina.it

Carlo and Lucio Sansone
{carlosan,sansone}@unina.it

Dipartimento di Informatica e Sistemistica
Università di Napoli Federico II
via Claudio, 21 80125 Napoli, Italy

## ABSTRACT

In the past few years, web usage mining techniques have grown rapidly together with the explosive growth of the web, both in the research and commercial areas. In this work we present a *Web mining* strategy for *Web personalization* based on a novel pattern recognition strategy which analyzes and classifies both static and dynamic features. The results of experiments on the data from a large commercial web site are presented to show the effectiveness of the proposed system.

## Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*; I.5.3 [**Pattern Recognition**]: Clustering—*Algorithms*

## General Terms

Algorithms, Experimentation

## Keywords

Clustering, web personalization, web usage mining

## 1. INTRODUCTION

It is well known that the world wide web may be considered as a huge and global information center. A *web site* usually contains great amounts of information distributed through hundreds of pages. Without proper guidance, a visitor often wanders aimlessly without visiting important pages, loses interest and leaves the site sooner than expected. This consideration is at the basis of the great interest about web information mining both in the academic and the industrial world. Usually, three types of data have to be managed in a web site: *content*, *structure* and *log* data. *Content data* consist of whatever is in a web page; *structure data* refer to the organization of the content; *usage data* are the usage patterns of web sites. The application of the data mining techniques to these different data sets is at the basis of the three different research directions in the field of web mining: *web content mining*, *web structure mining* and *web usage mining* [5]. In this paper, we are interested in the *web usage mining* domain, which is usually described as *the process of customizing the content and the structure of web sites* in order to provide users with the information they are interested in, without asking for it explicitly [3, 4]. Various personalization schemes have been suggested in the literature. The novelty of our strategy for personalizing the content of a web site is that we address all the following issues: i) a two-phase classification approach is used rather than a single-phase one; ii) both user-provided data and browsing patterns are taken into account; iii) both users and contents are classified.

## 2. THE USAGE MINING STRATEGY

In this section we describe our novel *web usage mining* strategy. It consists of two phases: in the first one a pattern analysis and classification is performed by means of an unsupervised clustering algorithm, using the registration information provided by the users. In the second one a reclassification is iteratively repeated until a suitable convergence is reached. Reclassification is used to overcome the inaccuracy of the registration information and it is accomplished by the *log analysis* and *content management* modules, based on the users' navigational behavior. We use an unsupervised clustering procedure for partitioning the feature space built upon the user-provided data into a certain number of clusters (each one representing a class) that group together users appearing to be similar. In order to choose the optimal number of clusters, we maximize the generalization capability of the system as defined in [2]. We propose the use of *Autoclass C* [1], a fuzzy unsupervised clustering algorithm based on the Bayesian theory. Each cluster is described through a likelihood function depending on some parameters. Given the number of classes, the *Autoclass C Search* module estimates such parameters on the training data and finds the partition of the feature space that maximizes the log-likelihood value. Once the optimal number of clusters has been chosen, the classification is performed by the *prediction* module of *Autoclass C*. By using the Bayesian rule and the likelihood function of each class, it attributes a user to that class which exhibits the maximum *a posteriori* probability. If a new user registers itself to the web site, it is classified according to the same scheme. Eventually, if a user explicitly changes the data in its registration form, it is classified again using the *Autoclass C prediction* module. The reclassification phase is based on the interaction of each user with the web site. We assume that the interaction can be performed in three different ways: *queries* containing some keywords, *searches* among directories, *navigation* of some pages. All the material on the web site is managed by the *content management* module of the system, which associates each resource (a keyword, a directory, a news headline or an article) to a specific content category. On the other hand the

[*] This work has been supported by the Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR) in the framework of the FIRB Project "Middleware for advanced services over large-scale, wired-wireless distributed systems (WEB-MINDS)"

*WWW2004*, May 17–22, 2004, New York, New York, USA.
ACM 1-58113-912-8/04/0005.

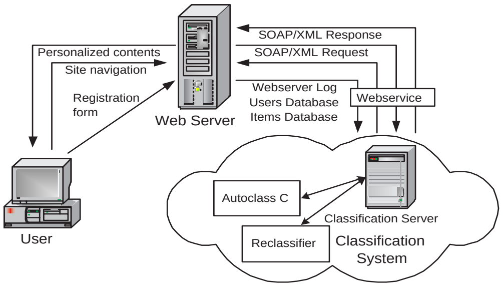

**Figure 1: System architecture**

*log analysis module* records all the activities of the users. In order to use these information for reclassifying users we need to attribute each category to a specific user class. This can be accomplished by considering the first classification performed by *Autoclass C* and counting the number $N_i$ of times in which the users of the i-class requested resources belonging to a specific category, over a time interval $T$. Each category is then attributed to the class that maximizes $N_i$. This way of classifying the content categories can suffer the inaccuracy of the first classification. However, if the time interval $T$ is wide enough and the percentage of correctly classified users is acceptable (say, greater than $50\%$), also the classification of the categories can be considered reliable. Now, a reclassification can be performed, by considering the resources that each user requested in a predefined time interval (*reclassification period*). If the majority of the requested contents belong to a class different from the initial one, the user is *reclassified*. The whole reclassification process will lead to convergence if, after a suitable number of reclassifications, the number of reclassified users goes to zero.

## 3. EXPERIMENTS

Figure 1 shows at a glance the overall architecture of the system. We have used an experimental commercial web site called *pariare.com*, usually visited by hundreds of users a day, which gives information about entertainment in the metropolitan area of Naples (*www.pariare.com*). The system has been tested for a period of six weeks. During this time the percentage of reclassified users has been tracked together with the percentages of transitions from a class to another one. The users, already registered to the web site when the experimentation started, have been initially classified using *Autoclass C*. First of all, we have determined the optimal number of clusters for classifying the users. The initial data set was made up of $2682$ users. It was divided into a training and a testing set respectively made up of $1282$ and $1400$ users. The features used to classify the users are: (1) age; (2) sex; (3) category of places in which users prefer to go; (4) number of times per week in which users go out; (5) preferred day of the week to go out; (5) the *Pariapoli* parameter (a measure of the degree of interest towards the virtual community of *Pariare*, evaluated as the normalized number of the information fields filled in the registration form); (7) type of entertainment users are looking for. Figure 2 shows how the percentage of reclassified users converges to zero, with respect to both the sample users and all users. Figure 3.a shows the distribution of the users among the classes produced by *Autoclass C* on the same day of the last reclassification, while figure 3.b shows the distribution of the users among the classes after the last reclassification. The comparison between the two distributions shows the benefits of adopting a classification strategy that takes into account both user-provided data and navigational behavior of the users: such a strategy can best fit the actual preferences of the users.

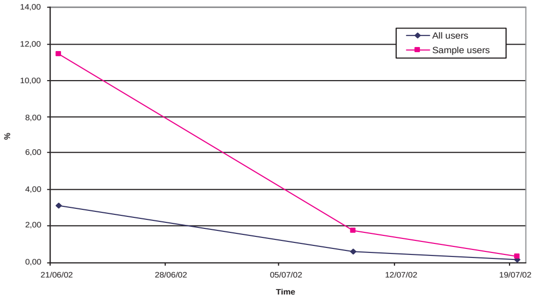

**Figure 2: Percentage of reclassified users vs time**

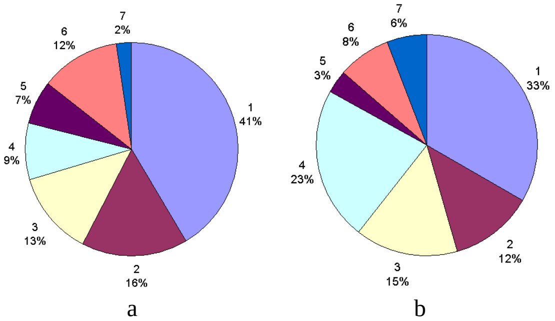

**Figure 3: Distribution of users among classes produced a) by *Autoclass C* at the time of the last reclassification; b) by the last run of the reclassification algorithm**

## 4. CONCLUSIONS

In this work we have introduced an interesting solution based on pattern recognition techniques, in order to classify a web user, based on its interaction with the web site. Several experiments have been performed and their results have been discussed. An off-line processing approach has been chosen for the classification task because it's easy to verify that the users do not change their preferences so frequently to justify the burden of an on-line processing.

We are planning to keep working on (a) a generalization of the proposed strategy in order to manage the structure of a web site; (b) the introduction of ontologies in order to automatically capture the web page contents from a semantical point of view; (c) a re-design of the system for a full compatibility with the web services standards; (d) the comparison of the proposed solution with the other ones presented in the literature and/or in the commercial areas.

## 5. REFERENCES

[1] P. Cheeseman and J. Stutz. Bayesian classification (autoclass): theory and results. *Advances in Knowledge Discovery and Data Mining, Eds. AAAI Press/MIT Press*, pages 61–83, 1996.

[2] C. De Stefano, C. Sansone, and M. Vento. Evaluating competitive learning strategies for handwritten character recognition. In *IEEE Int. Conf. on Systems, Man and Cybernetics Proceedings*, pages 759–764, Oct. 1994.

[3] M. Eirinaki and M. Vazirgiannis. Web mining for web personalization. *ACM TOIT.*, 3(1):2–27, Feb. 2003.

[4] M. Mulvenna, S. Anand, and A. Buchner. Personalization on the net using web mining. *CACM*, 43(8):123–125, Aug. 2000.

[5] F. Zhang and H. Chang. Research and development in web usage mining system–key issues and proposed solutions: a survey. In *First IEEE Int. Conf. on Machine Learning and Cybernetics Proceedings*, pages 986–990, Nov. 2002.