Mining usage profiles from access data using fuzzy clustering

G. CASTELLANO, A. M. FANELLI, M. A. TORSELLO
Department of Computer Science
University of Bari
Via Orabona, 4 – 70126 Bari
ITALY

Abstract: - In this work, we present an approach to clustering Web site users into different groups and generating common user profiles. These profiles are intended to be used to make recommendations by suggesting interesting links to the user. By using a fuzzy clustering algorithm, we enable generation of overlapping clusters that can capture the uncertainty among Web user's navigation behavior. Preliminary experimental results are presented to show the clusters generated by mining the access log data of a web site.

Key-Words: - Web mining, fuzzy clustering, access log, personalization, user profile

1 Introduction

The last years have been characterized by an exponential growth both of the number of online available Web applications and of the number of their users. This growth has generated huge quantities of data related to user interactions with the Web sites, stored by the servers in user access log files. On the other hand, the degree of personalization that a Web site is able to offer in presenting its services to users represents an important attribute contributing to the site's success. Hence, the need for a Web site that understands the interests of its users is becoming a fundamental issue. If properly exploited, log files can reveal useful information about user preferences. Therefore data mining, intended as knowledge discovery process from large database, has naturally found application on Web data, leading to the so-called *Web Mining* [17], [9], [15], [5], [11].

Three principal areas can be identified in Web Mining:

- *Web Content Mining* which focuses on the information available in the web pages;
- *Web Structure Mining* which searches the information resources in the structure of web sites;
- *Web Usage Mining* which deals with the knowledge extraction from server log files in order to derive useful patterns of user access.

Recently, several research activities have especially investigated *Web Usage Mining* techniques and a lot of works have been published on these topics [10], [5], [2], [14], [4], [1], [8]. A variety of traditional machine learning methods have been used for pattern discovery in Web Usage Mining [7], [11], [13], [16], [20]. Among these, unsupervised methods, especially clustering, seem to be the most appropriate to group users with common browsing behavior. A wide range of applications can benefit from the knowledge discovered by the clustering process, from real-time content personalization to dynamic link suggestion.

In the choice of the clustering method for Web Usage Mining, one important constraint to be considered is the possibility to obtain overlapping clusters, so that a user can belong to more than one group. Another key feature to be addressed in the developing Web Usage Mining techniques is vagueness and imprecision inherent Web usage data [6]. To deal with the ambiguity and the uncertainty underlying Web interaction data, fuzzy reasoning appears to be an effective tool.

In this paper, we explore the possibility of using fuzzy clustering to mine usage profiles from web log data. In particular, we use the fuzzy C-Means algorithm to categorize user sessions in order to derive groups of users which exhibit similar access patterns. The obtained clusters represent user profiles which can be exploited to implement different personalization functions, such as dynamic suggestion of links to Web pages retained interesting for the user.

The rest of the paper is organized as follows: Section 2 describes the steps of log data preprocessing aimed to identify user sessions, namely data cleaning and user identification. Section 3 deals with the employment of fuzzy clustering to derive user profiles by categorizing user sessions previously determined. Finally, Section 4 presents simulation results and draws final conclusions.

2 Log data preprocessing

When users visit a Web site, the Web server stores the information about their accesses in a log file. Each record of a log file represents a page request executed from a Web user. In particular, it typically contains the following information (fig. 1): user's IP address, date and time of the access, URL of the requested page, request protocol, a code indicating the status of the request, size of the page (if the request is successful).

```
66.249.65.243 - - [26/Mar/2006:07:10:44 +0200]
"GET/gallery2/main.php?g2_view=core.DownloadItem&g
2_itemId=5875 HTTP/1.1" 200 1745276
```

Figure 1: An example of record in a log file.

The aim of the preprocessing step is to identify user sessions starting from the information contained in the access log file. Since sessions encode the navigational behaviour of the users, their identification covers an important role for the success of a personalization system that is based on user profiles expressed as session categories.

In order to generate user profiles by a web mining technique, access log data need to be transformed into an appropriate format that meets the requirements of the mining task. In our approach, data preprocessing involves two main steps: data cleaning and session identification.

2.1 Data cleaning

The first step of log data preprocessing consists in removing useless requests from log files. In particular, data cleaning removes redundant references such as images, sound files, multiple frames, and dynamic pages that have the same template. We eliminate the irrelevant items by checking the suffix of the URL requests. Hence, all log entries with filename suffixes such as gif, jpeg, jpg, and map are removed. Moreover, data cleaning process identifies Web robots and removes their requests. These operations allow to not only remove uninteresting sessions but also to simplify the mining task that will follow.

2.2 Session identification

A user session can be defined as a limited set of pages accessed by the same user within a particular visit. According to [12], [18], we identify a user session as the set of accesses originating from the same IP address within a predefined time period. If the time between page requests exceeds a certain limit, we assume that the user is starting a new session. Here, we use 30 minutes as a default timeout. Supposing that the Web site is composed of n pages, each URL is assigned to a unique number j=1…n. Thus, a user session is represented by an n-dimensional vector where the j-th element expresses the degree of interest of the user for the j-th Web page. The degree of interest for a Web page can be defined in different ways. In this work, we consider the degree of interest to a URL as strictly related to the frequency of accesses to the page (number of accesses to that page / total number of accesses during the session) and to the time the user spends on the page. Formally, the i-th user session is represented by a vector $\mathbf{s}^{(i)} = \left(s_1^{(i)}, s_2^{(i)}, \ldots, s_n^{(i)}\right)$ with the property:

$$s_j^{(i)} = \begin{cases} f_j^{(i)} t_j^{(i)} & \text{if the user visits the } j\text{-th URL} \\ 0 & \text{otherwise} \end{cases}$$

for j=1…n, where $f_j^{(i)}$ and $t_j^{(i)}$ indicate, respectively, the access frequency and the time spent by the user on the j-th page during the i-th session. Summarizing, after the preprocessing phase, a collection of N sessions $\mathbf{s}^{(i)}$ is identified from the log data.

3 Session categorization by fuzzy clustering

Once user sessions have been identified, a clustering process is applied in order to group similar sessions in the same category. Each session category includes users exhibiting a common browsing behavior and hence similar interests. Hence, the identified session categories represent the different user profiles that will be successively exploited for suggesting links to pages considered interesting for a current user. Figure 2 illusrates the general scheme of the proposed model for mining usage profiles, showing the different involved steps starting from the log data preprocessing to the final clustering process.

Clustering is a well-known process of unsupervised learning. In literature, a lot of works have focused on clustering Web data using different types of clustering algorithms [19]. One important criteria to be considered in the choice of the clustering method is the possibility of creating overlapping clusters. This is a fundamental facet in Web personalization, where the ambiguity of the navigational data requires that a user may belong to more than one category or profile. Fuzzy clustering turns out to be a good candidate method to handle ambiguity in the data, since it enables the creation of overlapping clusters and introduces a degree of item-membership in each cluster. In this work, the well-known Fuzzy C-Means (FCM) clustering algorithm [3] is applied in order to group user sessions in overlapping categories which represent user profiles. As a result, FCM provides:

- C cluster prototypes represented as vectors $\mathbf{p}^{(c)} = \left(p_1^{(c)}, p_2^{(c)}, \ldots, p_n^{(c)}\right)$ with c=1,…,C;
- a fuzzy partition matrix $\mathbf{M} = \left[m_{ic}\right]_{\substack{i=1,\ldots,N \\ c=1,\ldots,C}}$ where each value $m_{i,c}$ represents the membership degree of the i-th session to the c-th cluster.

Briefly, the FCM algorithm is based on the minimization of the following objective function:

$$\mathrm{F}_\alpha = \sum_{i=1}^{N} \sum_{c=1}^{C} m_{ic}^{\alpha} \left\| s_i - p_c \right\|^2, \; 1 \le \alpha < \infty$$

where α is any real number greater than 1, m_{ic} is the degree of membership of the session s_i in the cluster c, p_c is the center of the c-cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. The algorithm is composed of the following

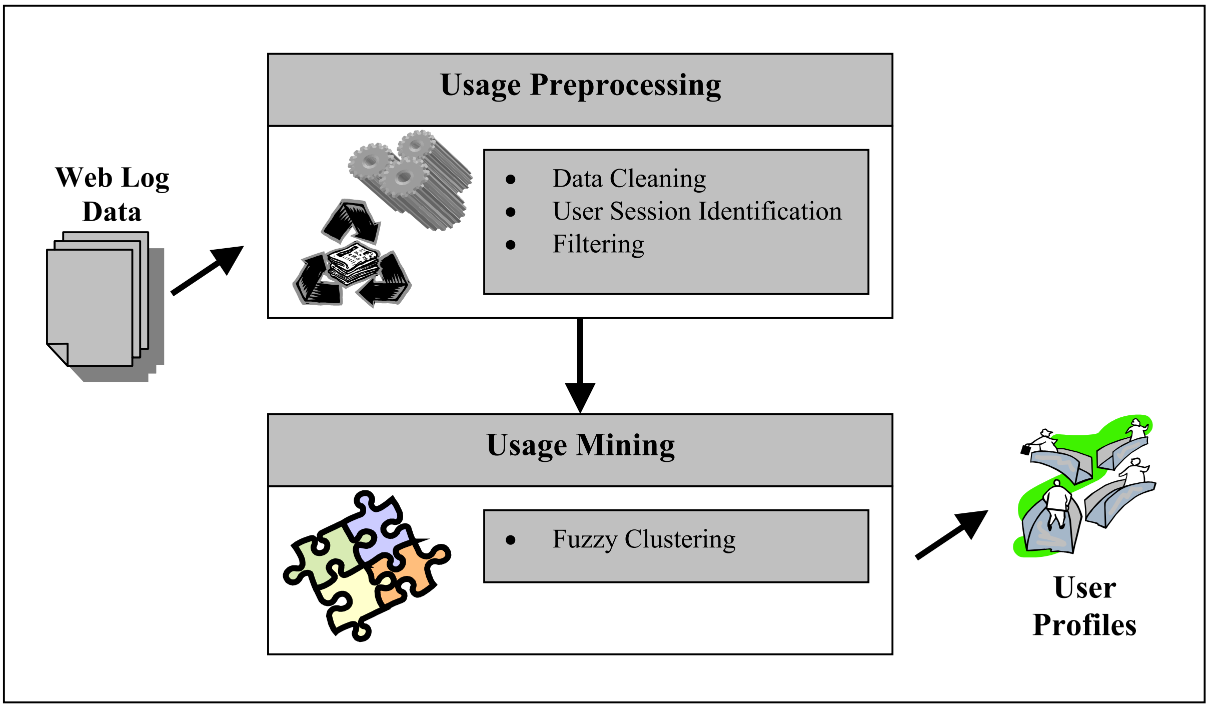

Figure 2: The general scheme of the proposed approach for mining usage profiles.

steps:

```
1. Initialize $\mathbf{M} = [m_{ic}]_{\substack{i=1,..,N \\ c=1,..,C}}$ matrix, $\mathbf{M}^{(0)}$
2. At k-th step: calculate the centers vectors
   $p^k = (p_c)_{c=1,..,C}$ with $\mathbf{M}^{(k)}$
   $p_c = \dfrac{\sum_{i=1}^{N} m_{ic}^{\alpha} s_i}{\sum_{i=1}^{N} m_{ic}^{\alpha}}$
3. Update $\mathbf{M}^{(k)}$, $\mathbf{M}^{(k+1)}$
   $m_{ic} = \dfrac{1}{\sum_{k=1}^{C} \left( \dfrac{\|s_i - p_c\|}{\|s_i - p_k\|} \right)^{\frac{2}{\alpha-1}}}$
4. If $\|M^{(k+1)} - M^{(k)}\| < \varepsilon$, with $0 < \varepsilon < 1$, STOP;
   otherwise return to step 2.
```

Summarizing, the clustering phase mines a collection of C session categories from session data, representing profiles of users that have accessed to the Web site under analysis.

4 Results and Conclusions

To test the proposed approach for mining usage profiles, a preliminary simulation was performed. A sample Web site was considered in order to carry out the experiments. During the log data preprocessing step, we applied a filtering process to select the mostly visited Web pages. For the sake of brevity, we indicate the selected pages through filtering process by the letters A, B, C, D, E, F, G, H, I and L. In our experiments, we considered the server log files containing user accesses to the sample Web site covering a time period of two weeks. Starting from these data, a total number of 62 user sessions were identified. Next, the FCM algorithm was applied in order to obtain clusters of users with similar navigational behavior corresponding to the user profiles. Carrying out different tests, we determined the best number of user profiles setting the number of clusters $C = 6$. Indeed, we observed that setting an higher number of clusters (i.e. $C = 8$ or $C = 10$) we obtained various prototype vectors with similar values. This demonstrated that a lower number of clusters was enough to model all the existing profiles. On the other hand, setting a number below 6, we risked to leave out interesting user profiles.

Table 1: The extracted user profiles.

User Profiles	Prototype Vectors	Common Access Pages
1	A=0.86, H=0.82	{A, H}
2	B=0.86, I= 0.80	{B, I}
3	D= 0.88, G= 0.85, L= 0.83	{D, G, L}
4	A= 0.82, I= 0.80	{A, I}
5	C= 0.87, F= 0.84, I= 0.76	{C, F, I}
6	E= 0.89, L= 0.84	{E, L}

Table 1 shows the information about the six obtained user profiles. In particular, for each user profile labeled with numbers from 1 to 6, the pages with highest rate of interest are indicated. In the last column, common access pages characterizing the user profile are reported. It can be noted that some pages (e.g. page A) belong to more than one cluster, thus showing the importance of producing overlapping clusters.

Summarizing, as supported by preliminary experimental results, the proposed approach can successfully discover user profiles from Web usage data. Further work is in progress to test the proposed approach to other Web sites having a high daily number of accesses.

References:

[1] A. Abraham, Business Intelligence from Web Usage Mining, *Journal of Information & Knowledge Management*, Vol. 2, No. 4, 2003, pp. 375-390.

[2] S. Araya, M. Silva, R. Weber, A methodology for web usage mining and its application to target group identification, *Fuzzy Sets and Systems*, 148, 2004, pp. 139–152.

[3] J.C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*, Plenum Press, New York, 1981.

[4] Y. H. Cho, J. K. Kim, Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce, *Expert Systems with Applications*, 26, 2004, pp. 233–246.

[5] F. M. Facca and P. L. Lanzi, Mining interesting knowledge from weblogs: a survey, *Data & Knowledge Engineering*, 53, 2005, pp. 225–241.

[6] E. Frias-Martinez, G. Magoulas, S. Chen, and R. Macredie, Modeling human behavior in user-adaptive systems: Recent advances using soft computing techniques, *Expert Systems with Applications,* 29, 2005, pp. 320-329.

[7] M. N. Garofalakis, R. Rastogi, S. Seshadri, K. Shim, Data minino and the web: past, present and future, *In Proc. of the second international workshop on web information and data management*, ACM, 1999.

[8] M . Kitsuregawa, M. Toyoda, I. Pramudiono, Web community mining and web log mining: commodity cluster based execution, *In Proceedings of the 13th Australasian Database Conference (ADC(02),* Melbourne, Australia, 5, 2002, pp. 3–10.

[9] R. Kosala, and H. Blockeel, WebMining Research: A Survey, *SIGKDDExplorations*, Vol.2, No.1, 2000, pp. 1-15.

[10] F. Masseglia, P. Poncelet, R. Cicchetti, An efficient algorithm for web usage mining, *J. Networking Inf. Syst. (NIS)*, 2(5-6), 1999, pp. 571–603.

[11] B. Mobasher, R. Cooley, J. Srivastava, Automatic personalization based onWeb usage mining, TR-99010, Department of Computer Science. DePaul University, 1999.

[12] O. Nasraoui, World Wide Web Personalization, In J. Wang (ed), *Encyclopedia of Data Mining and Data Warehousing*, Idea Group, 2005.

[13] O. Nasraoui, R. Krishnapuram, A Joshi, Relational clustering based on a new robust estimator with applications to web mining, *In Proc. of the International Conf. North American Fuzzy Info. Proc. Society (NAFIPS 99),* New York, 1999, pp. 705-709.

[14] D. Pierrakos, G. Paliouras, C. Papatheodorou, and C. D. Spyropoulos, Web usage mining as a tool for personalization: a survey. *User Modeling and User-Adapted Interaction*, Vol. 13, No. 4, 2003, pp. 311-372.

[15] K. P. Sankar, T. Varun, M. Pabitra, Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions, *IEEE Transaction on Neural Networks*, Vol. 13, No. 5, 2002, pp. 1163-1177.

[16] M. Spiliopoulou, L. C. Faulstich, K.Wilkler, A data miner analyzing the navigational behavior of Web users, *In Proceedings of theWorkshop on Machine Learning in User Modeling of the ACAI99*, Chania, Greece, 1999, pp. 54-64.

[17] J. Srivastava, R. Cooley, M. Deshpande, PN. Tan, Web usage mining: discovery and applications of usage patterns from web data. *SIGKDD Explorations,* Vol. 1, No. 2, 2000, pp.12–23.

[18] B.S. Suryavanshi, N. Shiri, and S.P. Mudur, A Fuzzy Hybrid Collaborative Filtering Technique for Web Personalization, in *Proc. of 3rd Workshop on Intelligent Techniques for Web Personalisation (ITWP'05)*, in conjunction with the *19th International Joint Conference on Artificial Intelligence (IJCAI05)*, Edinburg, Scotland, UK, 2005.

[19] A. Vakali, J. Pokorný and T. Dalamagas, An Overview of Web Data Clustering Practices, *EDBT Workshops*, 2004, pp. 597-606

[20] X. Wang, A. Abraham, K. A. Smith, Intelligent web traffic mining and analysis, *Journal of Network and Computer Applications*, 28, 2005, pp.147–165.