

A system based on a modified version of the FCM algorithm for profiling Web users from access log

Paolo Corsini, Laura De Dosso, Beatrice Lazzerini, Francesco Marcelloni

Dipartimento di Ingegneria dell'Informazione

via Diotisalvi, 2 -56123 Pisa

ITALY

e-mail: {p.corsini, l.dedosso, b.lazzerini, f.marcelloni}@iet.unipi.it

Abstract

In this paper, we present a system based on an appropriately targeted version of the well-known fuzzy C-means (FCM) algorithm to determine a small number of profiles of typical Web site users from the Web access log. These profiles can be extremely useful, for instance, to customize the Web site, or to send personalized advertisements. After filtering the access log, for instance, by eliminating occasional users, the FCM algorithm clusters the users of the Web site into groups characterized by a set of common interests and represented by a prototype, which defines the profile of the group typical member. To show the effectiveness of our system, we describe how the profiles determined by the FCM algorithm are a concise representation of the association rules discovered applying the well-known A-priori algorithm to the raw access log data.

Keywords: Web mining, user profile, fuzzy c-means, association rules.

1 Introduction

The rapid development of the World Wide Web as a medium for commerce and information dissemination has generated a growing interest in tools able to cluster the users into different groups and generating common user profiles from the Web access log. The identification of these profiles can be extremely useful, for instance, to E-commerce companies to send targeted advertisements, to guide

the user navigation and to define their market strategy. The identification of Web user profiles has been investigated in the recent literature by using different techniques [6][8].

In this work, we present a systematic approach to determine a small number of profiles of typical Web site users from the Web access log. We assume that Web pages of the site have been prearranged into a number of different classes, depending on the specific topic which is principally dealt with in the pages. This assumption is not a limitation as most Web commercial portals use such organization. Each user is, therefore, represented by the number of accesses to each class (or topic, in the following). The set of users is firstly filtered to remove possible noise, such as occasional users. Then, the fuzzy C-means (FCM) algorithm [2] with an appropriate distance function is applied to the filtered data to find out a small number of clusters. The optimal number of these clusters is determined by using the Xie-Beni index [7]. The prototype of each cluster summarizes the navigation preferences of the users strongly belonging to the cluster, thus identifying the profile of its typical members. The membership of each user to a cluster can be interpreted as the affinity degree of the user with the profile.

We applied our system to Web access log data collected by a commercial web portal during an observation period of 30 days and containing 1,249,426 users with accesses to 38 different topics. After the filtering of the raw data removed over 70% of the users, 21 profiles were determined as optimal summarizing representation of the users' interests.

To validate the results of our system, we applied the well-known A-priori algorithm proposed by Agrawal and Srikant [1] to determine a set of association rules between topics. The support and

confidence of each rule were evaluated based on the number of users. We show that the profiles determined by the FCM algorithm are a concise representation of the association rules with the highest supports and confidences.

2 The Profiling System

Let M be the number of topics. Each user u_i can be represented as a point $u_i = [u_{i,1}, \dots, u_{i,M}]$ in the space \mathcal{R}^M , where $u_{i,j}$ is the number of accesses of user u_i to the topic j during the observation time. Users are arrayed into an $N \times M$ matrix, where rows and columns represent, respectively, users and topics. Since Web portals are typically visited by a large amount of users, the number of rows is of the order of millions. Further, as a user is generally interested in a few topics, the matrix is very sparse. These two characteristics contribute to make the profiling process hard. In the experiments shown in this paper, the number N of users is 1,249,426, the number M of topics is 38, the total number of accesses is 13,961,483. Further, the distribution of the number of accesses among the various topics shows a large variability ranging from 2,910 to 2,558,102.

Our system consists of two modules in cascade. The first module, denoted FILTER, exploits some considerations on the Web user behavior to reduce noise and possibly decrease the number of users and the number of topics. The second module, denoted PROFILER, adopts the well-known fuzzy C-means algorithm, modified by using an appropriate distance rather than the classical Euclidean distance, to cluster the filtered user matrix and discover a set of profiles of typical users. In the following, we examine each module in detail.

3 The FILTER module

The FILTER module reduces noisy information from the access log by applying the following four steps in sequence:

3.1 Removing occasional users

Users who have visited a very few pages of the portal cannot be considered as sound samples of the body of users. Indeed, if the number is proportionally relevant with respect to the total number of users, these occasional users could

significantly affect the profiling process, thus leading the system to identify profiles which do not correspond to typical users. In our system, a user is judged to be occasional whether he/she has accessed a number of pages lower than a fixed threshold α_1 . In the experiments, we set α_1 to 4. Using this threshold, we removed approximately the 50% of the users, with a 7% reduction of the total number of accesses.

3.2 For each user, removing occasional accesses to topics in which the user is not really interested

During the navigation on the portal pages, users can access inadvertently topics which they are not really interested in. Obviously, we expect that the number of accesses to these topics is a modest percentage of the total number of accesses. To remove the occasionality from the typical behavior of the user, we set to zero the number of accesses to a topic when it is less than a fixed percentage α_2 of the total number of accesses by the user. The number of occasional accesses which are set to zero is not lost, but is collected into a virtual topic, denoted Other. This topic will be used in step 4 of the FILTER. In the experiments, setting α_2 to 5%, only 2% of the accesses are considered occasional.

3.3 Removing topics of poor interest

Some topics could be accessed by a very small number of users during the observation period. This occurs, for instance, for those topics such as Holidays which are interesting for the users only in some periods of the year. If the percentage of users which have visited pages of the topic is low, the topic will characterize no profile. We recall that the profiles will be determined so as to represent the behavior of typical users. Thus, we remove the topics which have not been accessed by a number of users larger than a fixed threshold α_3 . In the experiments, we set α_3 to 0.005%. No topic was discarded with this threshold.

3.4 Removing focused users and undecided users

Profiles of typical users are often used to decide market strategies or place targeted advertisements. To this aim, profiles characterized by only one topic, that is, profiles which represent focused users, or profiles characterized by too many topics, that is, profiles which represent undecided users, may not be interesting and, worst, might "hide" more commercially interesting profiles. To avoid these undesirable results, we remove users with accesses to only one topic and users with the number of accesses to the virtual topic Other larger than a fixed percentage α_5 of the total number of accesses of the

user. In the experiments, α_5 was set to 70%. The removal of focused and undecided users further reduces the number of users of approximately the 23% and 1% of the initial number of users, respectively. These reductions lead us to conclude that a large amount of the portal users focus their accesses on only one topic. On the other hand, a few users are characterized by an undecided behavior.

Analyzing the results produced by the FILTER module applied to the test Web access log, we can conclude that the filtering process strongly reduces the number of users (from 1,249,426 to 335,053) and the total number of accesses (from 13,961,483 to 5,830,874). Obviously, this reduction speeds up the execution of the clustering algorithm used to determine the profiles.

Figures 1 and 2 show the distribution of the users among the 38 topics before and after the filtering process. It can be noted that the relative ratio between bars of the histogram in Fig. 1 is approximately maintained in Fig. 2. This confirms that the parameters used in the filtering process allow reducing the number of users without altering their distribution among the topics.

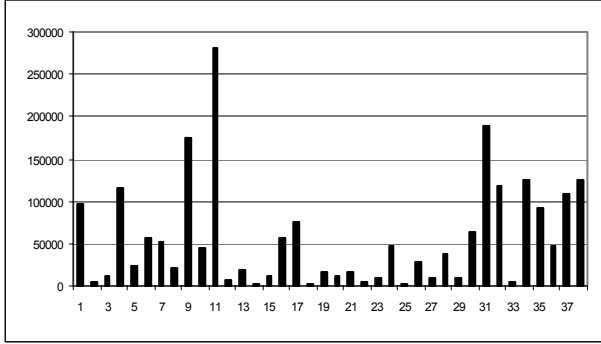


Figure 1: Distribution of the users before the filtering process

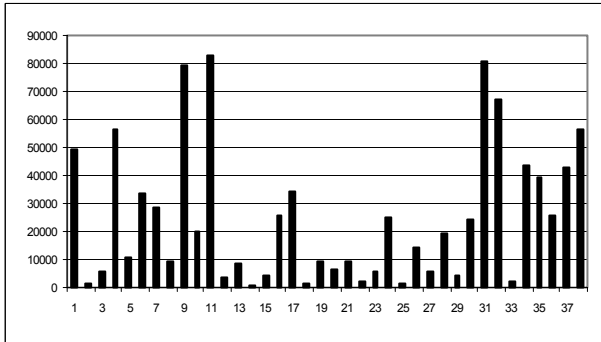


Figure 2: Distribution of the users after the filtering process

4 The PROFILER module

Let $\hat{U} = [\hat{u}_i]$ be the vector of the \hat{N} users survived after the filtering process. Each user can be represented as a vector $\hat{u}_i = [\hat{u}_{i,1}, \dots, \hat{u}_{i,\hat{M}}]$ in the space $\mathfrak{R}^{\hat{M}}$ of the \hat{M} topics which have not been eliminated in step 3 of the FILTER module. The coordinates of each vector correspond to the number of accesses to each topic. We observe that, in the profiling perspective, the behavior of a user is more accurately described by the relative orientation of the vector rather than its magnitude. Indeed, two users who access the same topics with the same proportion of the total number of accesses, though a different number of times, can be considered as samples of a same behavioral profile.

This observation leads us to state that the more two users are similar, the less the amplitude of the angle α formed by the corresponding vectors and, consequently, the higher the value of the cosine of α . Since the coordinates $\hat{u}_{k,j}$ of each vector \hat{u}_k vary on positive values, the cosine can assume only values in $[0, 1]$. Thus, we can define the dissimilarity $d(\hat{u}_k, \hat{u}_j)$ between two users \hat{u}_k and \hat{u}_j as:

$$d(\hat{u}_k, \hat{u}_j) = \sqrt{1 - \cos(\alpha)}$$

where $d(\hat{u}_k, \hat{u}_j)$ is called the cosine distance, and

$$\cos(\alpha) = \frac{\hat{u}_k \cdot \hat{u}_j}{\|\hat{u}_k\|_2 \|\hat{u}_j\|_2},$$

with $\|\cdot\|_2$ the Euclidean norm,

is the cosine of the angle formed by \hat{u}_k and \hat{u}_j . To speed up the computation of the cosine, we preliminarily normalize the users.

To cluster the users, we apply the version of the FCM algorithm proposed in [3]. Here, in place of the Euclidean distance, the dissimilarity measure between two users is computed as the cosine distance. Thus, the criterion function to be minimized becomes:

$$J(P, V) = \sum_{k=1}^N \sum_{i=1}^C (A_i(\hat{u}_k))^m \cdot d(\hat{u}_k, v_i)^2.$$

where $P = [A_1, \dots, A_C]$ is a fuzzy partition of the set \hat{U} of users, $A_i(\hat{u}_k)$ is the membership value of user \hat{u}_k to cluster A_i , $V = [v_1, \dots, v_C]$ are the C prototypes of the clusters in P , and m is the fuzzification constant. The optimal partition P is computed by

using an iterative method based on successive minimization of the functions $J(P, \cdot)$ and $J(\cdot, V)$. To minimize $J(P, \cdot)$, we apply the Lagrange multiplier method with the constraint $\sum_{i=1}^C A_i(\hat{u}_k) = 1$ and obtain the following formula:

$$A_i(\hat{u}_k) = \frac{1}{\sum_{j=1}^C \left(\frac{d(\hat{u}_k, v_j)}{d(\hat{u}_k, v_i)} \right)^{\frac{2}{m-1}}} \quad (1)$$

To minimize $J(\cdot, V)$, we apply again the Lagrange multiplier method with the constraint $\sum_{f=1}^{\hat{M}} v_{i,f}^2 = 1$ and get the following formula (see [3] for a demonstration):

$$v_{i,f} = \frac{\sum_{k=1}^N (A_i(\hat{u}_k))^m \hat{u}_{k,j}}{\sqrt{\sum_{t=1}^{\hat{M}} \left(\sum_{k=1}^N (A_i(\hat{u}_k))^m \hat{u}_{k,t} \right)^2}} \quad (2)$$

To determine the optimal number of clusters which partition the users, we executed the FCM with increasing values of the number C (from 16 to 32) of clusters and assessing the goodness of each resulting partition using the Xie-Beni index [7]. We plotted the Xie-Beni index versus C and chose, as optimal number of clusters, the value of C corresponding to the first distinctive local minimum [5]. We found out $C=21$ as optimal number of clusters. To speed up the execution of FCM and decrease the memory occupation, we adopted the implementation suggested in [4]. In the experiments, the execution time of FCM on a 2GHz Pentium IV with 1GB RAM and FreeBSD 4.5 as operating system was of the order of a few minutes, which is acceptable for this type of application. Due to the sparseness of the user matrix, we executed the FCM algorithm with the fuzzification coefficient m set to 1.15. Using an accuracy error equal to 0.01, we observed that the FCM converges after 15 ÷ 25 iterations.

Fig. 3 shows one of the profiles identified by the FCM algorithm. Here, only the topics with a considerable number of accesses are reported. The users who are represented by this profile are characterized by a strong interest in Football, a good interest in Sport, a modest interest in Cars and Motorcycles, Cinema and Music, and a scarce

interest in the other topics. The profile seems to identify users who navigate the Web portal in search of news to fill their spare time. We recall that a profile is a virtual user and is represented as a unit vector in the space $\mathfrak{R}^{\hat{M}}$ of the \hat{M} topics.

5 Validation

To validate the results achieved by our system, we applied the A-Priori algorithm to the raw access log data to discover association rules between topics [1]. We aim to verify if the relations between topics highlighted by the association rules with high support and confidence are contained in the profiles determined by our system. In fact, these relations summarize the behavior of typical users. An association rule is defined as an implication in the form $T_l \Rightarrow T_r$, where T_l and T_r are sets of topics. The implication expresses the fact that users, who have accessed the set of topics T_l , have also accessed the set T_r . The relevance and reliability of an association rule is determined by its support and its confidence. The support is defined as the ratio (expressed in percentage) between the users who have accessed all the topics in the set $T_l \cup T_r$ and the total number of users; the confidence coincides with the ratio (expressed in percentage) between the users who have accessed all the topics in the set $T_l \cup T_r$ and the users who have accessed all the topics in the set T_l .

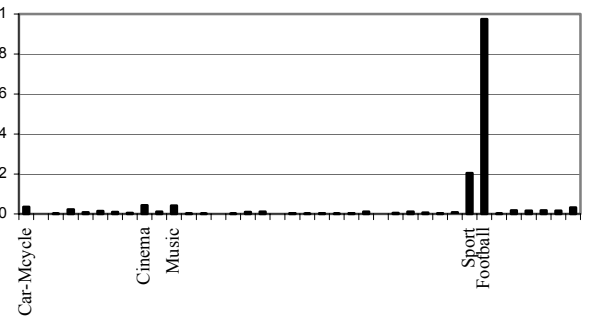


Figure 3: One of the 21 profiles

We executed the A-Priori algorithm in such a way as to discover association rules with support and confidence larger than 0.1% and 5%, respectively. To validate the results obtained by our system, we analyzed in detail the association rules as follows. For each profile determined by the system, we picked the topic (*prevalent topic*) with the highest value. For instance, in the profile in Fig. 3, we

picked topic Football. Then, we selected all association rules (*relevant association rules*) with the prevalent topic in the set T_j . We observed that all the significant topics of the profile were in the set T_r of the relevant association rules with the highest supports and confidences. As an example, Table 1 shows the set of the relevant association rules selected for topic Football with the highest supports and confidences. We can observe that the sets T_r of the rules contain all the significant topics of the profile in Fig. 3. This confirms that the relations highlighted in the profile are really the relations existing between the topics in the data set.

Table 1: Relevant association rules for Football.

Association Rules	Support	Confidence
Football => Sport	4.21%	44.62%
Football => Cinema	1.42%	15.11%
Football => Music	1.39%	14.76%
Football => Car-Motorcycle	1.35%	14.37%

6 Conclusions

In this paper, we have shown a system to determine a small number of profiles of typical Web site users from the Web access log. The main features of this system are an efficient filtering module, which reduces drastically the amount of raw access log data, and the FCM clustering algorithm with an appropriate definition of distance. To explain how the filtering process does not eliminate relevant information and the FCM works in an effective way, we have applied the A-Priori algorithm to the raw access log data to discover association rules between

topics. We have shown how the profiles determined by the system are a concise representation of the association rules with high support and confidence.

References

- [1] R. Agrawal, R. Srikant (1994). Fast Algorithms for Mining Association Rules. In *Proceedings of the 20th VLDB Conference*, pp. 478-499, Santiago, Chile.
- [2] J.C. Bezdek (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York.
- [3] F. Klawonn, A. Keller (1999), Fuzzy Clustering Based on Modified Distance Measures. In: D.J. Hand, J.N. Kok, M.R. Berthold (eds.): *Advances in Intelligent Data Analysis*, Springer, Berlin, pp. 291-301.
- [4] J.F. Kolen, T. Hutcheson (2002). Reducing the time complexity of the fuzzy C-means algorithm. *IEEE Transactions on Fuzzy Systems* vol. 10, no. 2, pp. 263-267.
- [5] M. Setnes, H. Roubos (2000). GA-fuzzy modeling and classification: Complexity and performance. *IEEE Transactions on Fuzzy Systems*, vol. 8, no. 5, pp. 509-522.
- [6] K.A. Smith, A. Ng (2003). Web page clustering using a self-organizing map of user navigation patterns. *Decision Support Systems*, vol. 35, pp. 245-256.
- [7] X.L. Xie, G. Beni (1991). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841-847.
- [8] Y. Xie, V.V. Proha (2001). Web User Clustering from Access-Log Using Belief Function. In *Proceedings of K-Cap'01*, pp. 202-208, October 22-23, Victoria, Canada.