Annales UMCS Informatica AI 7 (2007) 153-161

Annales UMCS
Informatica
Lublin-Polonia
Sectio AI

http://www.annales.umcs.lublin.pl/

Data mining techniques for portal participants profiling

Danuta Zakrzewska[*], Justyna Kapka

Institute of Computer Science, Technical University of Łódź,
Wólczańska 215, 93-005 Łódź, Poland

Abstract

Recently, a large number of virtual learning communities appeared in the Web, however, keeping them up occurs to be problematic and information accessibility is one of the factors that may influence their sustainability. This feature may be achieved by dividing users into groups according to their information needs and by adapting properly the portal contents. In the paper application for data mining algorithms, for finding patterns together with different groups of preferences is considered. We base our research on the data contained in log files. Combination of sequential pattern mining and clustering techniques is proposed. We describe the data preparation process. The experiments conducted for real data log files are discussed.

1. Introduction

Together with the Web development there appeared a lot of virtual learning communities centered around Internet portals. At the beginning they attract the members, but their sustainability becomes a big problem and acceptance of technology may play a significant role in it. In [1], Teo, Chan, Wei & Zhang introduced extended TAM (Technology Acceptance Model) of sustainability of virtual learning communities. They indicated that information content, the type of communication channels provided and information organization are important success factors. User centered adaptation ability of a portal, which takes into account all of these factors, may help achieve information accessibility satisfying users' expectations.

In the paper we propose using data mining techniques for portal participants segmentation. We base our research on the data contained in log files and according to users' behavior, which may be presented in a form of sequential patterns, we group users with similar information interests and the same navigation paths. It may help us determine patterns and make the portal friendly for all big groups of users. In our investigations we connect sequential pattern

[*] Corresponding author: *e-mail address*: dzakrz@ics.p.lodz.pl

mining with the clustering method, introduced in [2]. Our approach, by combining two different techniques allows to build navigation patterns and discover the groups with similar information preferences at the same time. Finding sequential patterns helps in discovering all the frequent subsequences from all sequences of pages visited by users. Clustering algorithm, in turn, allows to segment almost all the portal participants and organize the information content, that will fulfill their needs.

The paper is organized as follows: after short description of relevant research, we describe the method of profiles building as well as the data preparation process. We discuss the efficiency of the techniques applied. Next we present and analyze some exemplary results for certain log files. Finally, we give some conclusions concerning the applied techniques and the future research.

2. Relevant research

Lately many authors have examined techniques for categorization of web users according to their information needs, by extracting knowledge from logging data (see [3-6]). As the main goals of the research in this area, there should be mentioned building adaptive Web sites or Web recommender systems. The investigations concerning mining for identification of users' needs were also conducted for the information systems designing (see [7]) as well as the workflow process building (see [8] for example). According to the research experience so far, logs occurred to be a good source for mining user preferences in all considered cases.

There are two main directions in log data mining: finding navigational patterns and dividing users according to their preferences. There exist different approaches for analyzing log data, as main techniques there should be mentioned association rules, sequential patterns and clustering. The first two groups of methods are usually applied for finding navigational patterns of users, while cluster analysis is used for data segmentation and finding groups of users with similar preferences. The wide review of all the methods and recent development in the area, are presented in [9,10] and [11]. Most authors focus their research on applying one methodology, however, some of personalization systems use different methods separately (see [6,12]). Our approach, by combining two different techniques: sequential pattern mining and cluster analysis, allows to build navigation patterns and discover the groups with similar interests at the same time. The first technique contrary to association rules, which do not take into account the element of time, allows for determining event sequences and it seems to be more suitable for finding such patterns as users navigation paths. Additionally, sequential pattern mining may help discover all the frequent subsequences from all sequences of pages visited by users (compare [3,5]), while clustering algorithm allows to segment almost all the portal participants and organize the information content, that will fulfil their needs. To this effect the special clustering algorithm is used, based on the POPC (Pattern Oriented Partial Clustering) algorithm described in [2].

3. Profiles building

The presented profiles building system is composed of three main steps. In the preprocessing phase data preparation consists in identification of users and their navigation paths. As the result of second step we get all sequential patterns. At the final stage, clusters containing sequential patterns and users are obtained. The first element is to help choose the way of portal information organization , the second one allows to assign users into properly profiling portal contents. The overview of system architecture is presented on Fig. 1.

3.1. Data preparation

We consider log data, written on the Web server, in *Common Logfile Format* with records in the form ([9]):

Remotehost rfc931 authuser date “request” status bytes.

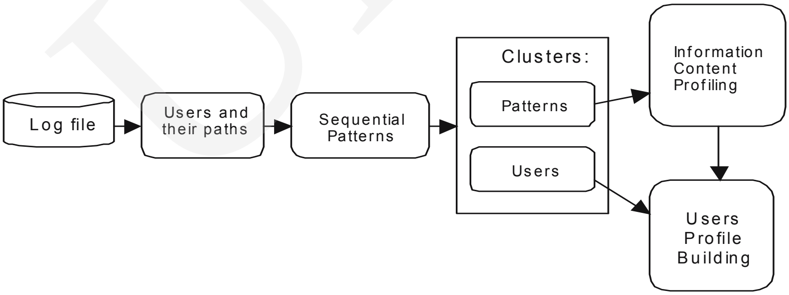

Fig. 1. System architecture: from log file to profile building

The first three fields determine the users’ names, however, only the third one signifies the authorized name, the other two are connected with remote computers and may not characterize users properly. The fourth attribute means the access date. The other field important for our considerations is *“request”*, which describes the exact request line as it comes from the users. The attribute *“status”* identifies error (success) code and the last one *“bytes”* means the number of bytes transferred from the Web server to the user.

In the first step of data preparation, the records with the error “*status”* are removed. There are only two attributes taken into account while considering reminding records: “*authuser”* and *“request”*. All the records, cleaned of other attributes, are sorted by the first one, and thus we obtain users with their paths in the prepared data file.

3.2. Sequential pattern mining

The main idea of this step is based on the algorithm presented in [13], in which sequential patterns are represented by maximal sequences among all sequences with a certain minimal support. At this stage the system finds sequential patterns in the preprocessed data. In our approach we seek all the maximal sequences (that are not contained in any other sequences), in the set of all the sequences of each user. Each maximal sequence found by the algorithm, determines the pattern, that will be used in the cluster analysis.

Let S be the set of all users' sequences, maxlength – the length of the longest sequence. The algorithm, in pseudo-code, can be presented as follows:

```
For (k=maxlength; k>1; k--) do
    Foreach k-sequence $s_k$ do
        Remove all subsequences $s_k$ from S.
```

The obtained maximal sequences identify users' navigational paths and may be used in the process of information content profiling. After finding all sequential patterns, the data are prepared for clustering technique.

3.3. Clustering

We base our clustering method on the algorithm POPC introduced in [2]. Having maximal sequential patterns for all users, we assign them to a certain number of clusters, taking into account patterns similarity. Likewise in [2] we consider two sequential patterns as similar if they contain identical subsequences or there exists a connecting path in the other sequence. Fig. 2. presents an example of a connecting path, in that case all the sequences ($url_3 \rightarrow url_5 \rightarrow url_7 \rightarrow url_{10}$, $url_5 \rightarrow url_6 \rightarrow url_7 \rightarrow url_{10} \rightarrow url_{14}$, $url_6 \rightarrow url_{10} \rightarrow url_{13} \rightarrow url_{14}$) would be put into the same cluster.

Hierarchical agglomerative algorithm is used, starting with every sequential pattern as a separate cluster. The distance between clusters is counted by using the Jaccard's coefficient [14]. Clusters with the biggest value of this coefficient are merged until certain number of them is obtained.

Two versions of the algorithm are implemented, the first one by using similarity matrix where distance values are remembered with clusters and the second one by calculating distance values while building clusters from the beginning. The last version is presented below.

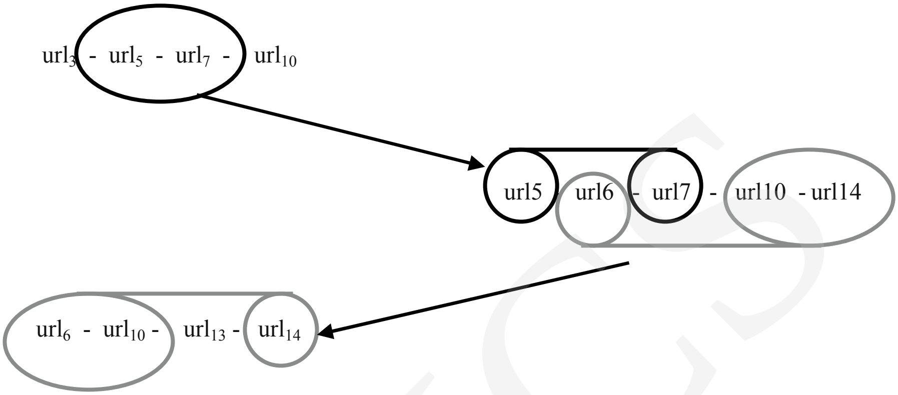

Fig. 2. An example of a connecting path

Let C_1 be the initial set of clusters, n the demanded number of clusters, D the set of the sequences and s_j – sequence from D. Let us denote by c_a, c_b the clusters considered for merging operation and *Jaccard*(c_a,c_b) the proper Jaccard's coefficient, then the clustering algorithm can be written as follows:

```
$C_1 = \{s_j: s_j \in D \wedge s_j$ contains pattern $p_i$ }.
k=1; while (|$C_k$|> n ) do
{foreach ($c_a$,$c_b$)∈ $C_k$ such, that Jaccard($c_a$,$c_b$)>0) do
{ if Jaccard($c_a$,$c_b$) is maximal then remember($c_a$,$c_b$);}
merge clusters($c_a$,$c_b$);
delete remembered pair of clusters;
add new cluster $C_{k+1}$ to the set of clusters; k++;}
Result = $C_k$;
```

Analysis of the obtained clusters containing similar navigational paths in the form of sequences, allows for designing different information contents of portals, that may satisfy preferences of users assigned to clusters. However, users are usually connected to more than one sequential pattern and it may happen that they are allocated into more than one cluster, or on the contrary they may not be assigned at all. Especially the first feature of the algorithm may cause some problems while information content organizing, which will be discussed in the final part of the paper.

The example of sequential pattern and clustering stages of the system is presented below:

Example 1

Let us consider a data set with navigation paths:
ulr1→ulr2→ ulr3→ ulr4→ ulr5,

ulr6→ ulr7→ ulr8,
ulr2→ ulr3→ ulr5,
ulr4→ ulr5→ ulr6→ ulr7→ ulr8→ ulr9,
ulr7→ ulr9,
ulr3→ ulr5→ ulr6→ ulr8,
ulr2,
ulr3,
ulr4→ ulr5,

Three maximal sequential patterns were found by the system:

ulr4→ulr5→ ulr6→ ulr7→ ulr8→ ulr9,
ulr1→ ulr2→ ulr3→ ulr4→ ulr5,
ulr3→ ulr5→ ulr6→ ulr8.

While assigning them into two clusters we obtain (it is easy to notice that there exists a connecting path between elements of Cluster 1):

Cluster 1: ulr4→ ulr5→ ulr6→ ulr7→ ulr8→ ulr9,
ulr3→ ulr5→ ulr6→ ulr8.
Cluster 2: ulr1→ ulr2→ ulr3→ ulr4→ ulr5.

Additionally to the sequential patterns that mean navigational paths, users that are connected with these patterns, will be assigned to the proper clusters.

4. Experiments

Experiments were done for the real users log files. Performance of the algorithms was tested for different sizes of web log files and a different number of clusters on the computer with the processor AMD Athlon XP 1900+ and 256 MB RAM. Considered executive time comprises main steps of the algorithm not taking into account executive time of data preparation.

The size of log files has the significant influence on the number of maximal sequences generated by the first part of the algorithm, which can be seen in Table 1. The performance of the system depending on a log file size and the required number of clusters is shown in Fig. 3. Fig. 4, in turn, presents the dependence of run time on the number of required clusters and the log file size. It is easy to notice that the number of sequences and the run time are growing significantly together with the file size.

Figs. 3 and 4 show that the log file size (number of generated sequential patterns) has much more influence on the run time than the required number of clusters. Additionally, we can see in Fig. 3 that this dependence is bigger in the case of a smaller number of clusters assumed.

Table 1. Number of generated sequential patterns depending on the log file size

File size [MB]	Number of generated sequences
1	294
2	443
3	573
4	686

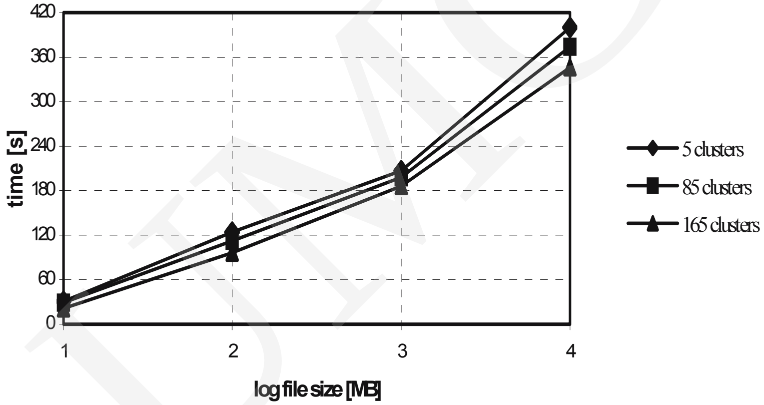

Fig. 3. Clustering time depending on the log file size and the number of clusters

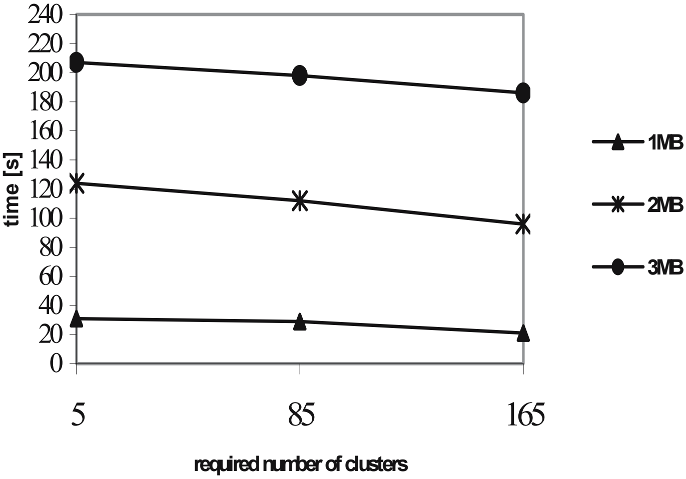

Fig. 4. Clustering time depending on the required number of clusters and the file size

Two versions of clustering algorithms were compared. Experiments showed that remembering the similarity matrix significantly improves efficiency of the algorithm, which can be seen in Fig. 5. The results obtained in both cases were the same.

Qualitative analysis of the obtained clusters by considering the samples also demonstrated the effectiveness of the system in finding portal users preferences, in almost all considered cases.

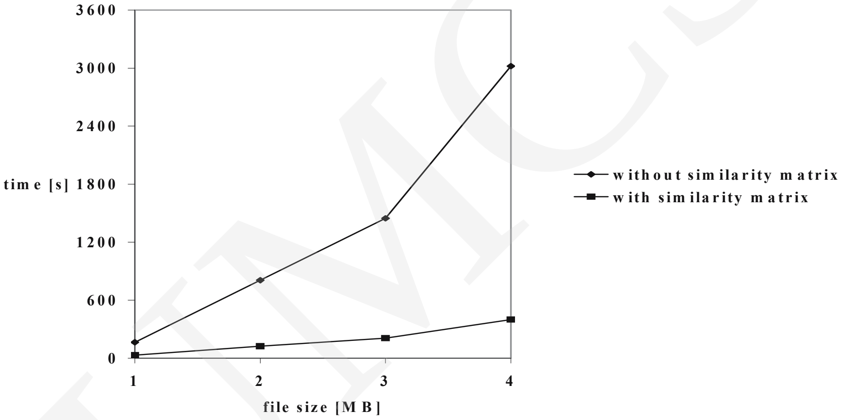

Fig. 5. Clustering time for two versions of algorithms

5. Conclusions

In the paper we consider combining sequential pattern mining approach with the cluster analysis into portal user profiles building for information content organizing. The special clustering technique proposed in [2], which assigns sequential patterns and user into clusters, allows to have the full image of users navigation paths and information preferences. The analysis of exemplary results showed that the method is effective, however, depends significantly on the log file size and the number of required clusters which in the case of fitting users interface cannot be very big.

The main disadvantage of this technique consists in the fact that users may be assigned into more than one cluster, which may be only solved by optional organization of information content and navigation. Users not assigned to any clusters may use any standard interface or be qualified into the biggest group of users.

The future research should be focused on the connection of the proposed system with the information content and organization design to enable the automation of the whole process.

References

[1] Teo H.H., Chan H.C., Wei K.K., Zhang Z., *Evaluating information and community adaptivity features for sustaining virtual learning communities*. International Journal of Human-Computer Studies, 59 (2003) 671.

[2] Morzy T., Wojciechowski M., Zakrzewicz M., *Web users clustering,* Proceedings of the 15th International Symposium on Computer and Information Sciences, Istanbul, Turkey (2000) 374.

[3] Zaïane O.R., Xin M., Han J., *Discovering web access patterns and trends by applying OLAP and data mining technology on web logs*. Proceedings of Advances in Digital Libraries Conference (ADL'98), Santa Barbara, CA (1998) 19.

[4] Pei J., Han J., Mortazavi-asl B., Zhu H., *Mining access patterns efficiently from web logs*. Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'00), Kyoto, Japan, Lecture Notes in Computer Science, 1805, Springer-Verlag, Berlin Heidelberg New York, (2000) 396.

[5] Srikant R., Yang Y., *Mining web logs to improve website organization*. World Wide Web, (2001) 430.

[6] Mobasher B., Cooley R., Srivastava J., *Automatic personalization based on web usage mining*. Communications of the ACM, 43 (2000) 142.

[7] Liu R.-L., Lin W.-J., *Mining for interactive identification of users' information needs*. Information Systems, 28 (2003) 815.

[8] van der Aalst W.M.P., van Dongen B.F., Herbst J., Maruster L., Schimm G., Weijters A.J.M.M., *Workflow mining: a survey of issues and approaches*. Data & Knowledge Engineering, 47 (2003) 237.

[9] Erinaki M., Vazigriannis M., *Web mining for web personalization*. ACM Transactions on Internet Technology, 3 (2003) 1.

[10] Pierrakos D., Pakouras G., Papatheodorou Ch., Spyropoulos C.D., *Web usage mining as a tool for personalization: a survey*. User Modelling and User Adapted Interaction, 13 (2003) 311.

[11] Facca F.M., Lanzi P,L., *Mining interesting knowledge from weblogs: a survey*. Data & Knowledge Engineering, 53 (2005) 225.

[12] www.mindlab.de.

[13] Agrawal R., Srikant R., *Mining sequential patterns*. Proceedings of the 11th International Conference on Data Engineering (ICDE'95), Taipei, Taiwan, (1995) 3.

[14] Han J., Kamber M., Data Mining: Concepts and Techniques, Morgan Kaufman Publishers (2001).