Mining Generalized Web Data for Discovering Usage Patterns

Doru Tanasa
INRIA Sophia Antipolis, France

Florent Masseglia
INRIA Sophia Antipolis, France

Brigitte Trousse
INRIA Sophia Antipolis, France

INTRODUCTION

Web Usage Mining (WUM) includes all the Data Mining techniques used to analyze the behavior of a Web site's users (Cooley, Mobasher & Srivastava, 1999, Spiliopoulou, Faulstich & Winkler, 1999, Mobasher, Dai, Luo & Nakagawa, 2002). Based mainly on the data stored into the access log files, these methods allow the discovery of frequent behaviors. In particular, the extraction of sequential patterns (Agrawal, & Srikant, 1995) is well suited to the context of Web logs analysis, given the chronological nature of their records. On a Web portal, one could discover for example that "25% of the users navigated on the site in a particular order, by consulting first the homepage then the page with an article about the bird flu, then the Dow Jones index evolution to finally return on the homepage before consulting their personal e-mail as a subscriber". In theory, this analysis allows us to find frequent behaviors rather easily. However, reality shows that the diversity of the Web pages and behaviors makes this approach delicate. Indeed, it is often necessary to set minimum thresholds of frequency (i.e. minimum support) of about 1% or 2% before revealing these behaviors. Such low supports combined with significant characteristics of access log files (e.g. huge number of records) are generally the cause of failures or limitations for the existent techniques employed in Web usage analysis.

A solution for this problem consists in clustering the pages by topic, in the form of a taxonomy for example, in order to obtain a more general behavior. Considering again the previous example, one could have obtained: "70% of the users navigate on the Web site in a particular order, while consulting the home page then a page of news, then a page on financial indexes, then return on the homepage before consulting a service of communication offered by the Web portal". A page on the financial indexes can relate to the Dow Jones as well as the FTSE 100 or the NIKKEI (and in a similar way: the e-mail or the chat are services of communication, the bird flu belongs to the news section, etc.). Moreover, the fact of grouping these pages under the "financial indexes" term has a direct impact by increasing the support of such behaviors and thus their readability, their relevance and significance.

The drawback of using a taxonomy comes from the time and energy necessary to its definition and maintenance. In this chapter, we propose solutions to facilitate (or guide as much as possible) the automatic creation of this taxonomy allowing a WUM process to return more effective and relevant results. These solutions include a prior clustering of the pages depending on the way they are reached by the users. We will show the relevance of our approach in terms of efficiency and effectiveness when extracting the results.

BACKGROUND

The structure of a log file is formally described in Definition 7 (at the end of this chapter). This data structure can be easily transformed to the one used by sequential pattern mining algorithms. A record in a log file contains, among other data, the client IP, the date and time of the request, and the Web resource requested. To extract frequent behaviors from such a log file, for each user session in the log file, we first have to: transform the ID-Session into a client number (ID), the date and time into a time number, and the URL into an item number. Table 1 gives a file example obtained after that pre-

Table 1. File obtained after a pre-processing step

Client \ Date	d1	d2	d3	d4	d5
1	a	c	d	b	c
2	a	c	b	f	c
3	a	g	c	b	c

processing. To each client corresponds a series of times and the URL requested by the client at each time. For instance, the client 2 requested the URL "f" at time *d4*. The goal is thus, according to definition 2 and by means of a data mining step, to find the sequential patterns in the file that can be considered as frequent. The result may be, for instance, *<(a)(c)(b)(c)>* (with the file illustrated in table 1 and a minimum support given by the user: 100%). Such a result, once mapped back into URLs, strengthens the discovery of a frequent behavior, common to *n* users (with *n* the threshold given for the data mining process) and also gives the sequence of events composing that behavior.

The main interest in employing sequential patterns for Web usage mining aims at taking into account the time-dimension of the data as in the papers described thereafter.

The WUM tool (Web Utilisation Miner) proposed in (Spiliopoulou, Faulstich & Winkler, 1999) allows the discovery of navigation patterns which are interesting either from the statistical point of view or through their structure. The extraction of sequential patterns proposed by WUM is based on the frequency of the patterns considered.

Unfortunately, the amount of data (in terms of different items—pages—as well as sequences) is an issue for the techniques of sequential pattern mining. The solution would consist in lowering the minimum support used, but in this case the algorithms are not able to succeed. A proposal for solving this issue was made by the authors of (Masseglia, Tanasa & Trousse, 2004), who were interested in extracting sequential patterns with low support on the basis that high values for the minimum support often generate obvious patterns. The authors proposed to divide the problem in a recursive way in order to proceed to a phase of data mining on each sub-problem.

Another solution is to reduce the number of items by using a generalization of URLs. In (Fu, Sandhu & Shih, 2000) the authors use a syntactic generalization of URLs with a different type of analysis (clustering). Before applying a clustering, the syntactic topics of a level greater than two are replaced by their syntactic topics of a lower level.

Finally, there are other methods (Srikant & Agrawal, 1996) that can be used to extract sequential patterns by taking into account a generalization, but in other domain than the Web. Nevertheless, the automatic construction of generalizations (in the form of classes) of Web pages and for the purposes of a Web usage analysis was not studied yet. We propose in the following section such a method which is based on characteristics of Web usage data.

GWUM: MOTIVATIONS AND GENERAL PRINCIPLE

We present here our main motivations and the way we perform an usage-driven generalization of Web pages. As we mentioned in the introduction, the generalization of the items is a key factor during the extraction of sequential patterns. To understand the advantage of our work compared to a classical technique of sequential pattern mining, we propose the following example.

Let us consider the recordings from INRIA Sophia-Antipolis' Web access log (in the preprocessed form) as illustrated in Table 2. We can see there that the user C1 at date D1 made a request for the URL "homepage_DT" which is Doru Tanasa's homepage, then at date D2 he made a request for the publications page of Doru Tanasa and, finally, a request for INRIA's homepage. Similarly, user C2 made a request for Sergiu Chelcea's homepage at Date D1, and so on.

When extracting sequential patterns with a minimum support of 100%, no patterns will be found in this log (there is no item supported by 100% of the sequences). To find a frequent pattern, it will be necessary to lower the minimum support down to 50%, which allows the extraction of the following behaviors:

Table 2. Accesses to the Web site grouped by client (User)

Date / Client	D1	D2	D3
C1	homepage_DT	publications_DT	homepage_Inria
C2	homepage_SC	publications_SC	logiciels_AxIS
C3	homepage_DT	publications_AxIS	publications_DT
C4	homepage_AxIS	homepage_SC	publications_SC