Abstract

The past fifteen years are characterized by an exponential growth of the Web both in the number of Web sites available and in the number of their users. This growth generated huge quantities of data related to the users interaction with the Web sites, recorded in Web log files. Moreover, the Web sites owners expressed the need to better understand their visitors in order to better serve them.

The Web Use Mining (WUM) is a rather recent research field and it corresponds to the process of knowledge discovery from databases (KDD) applied to the Web usage data. It comprises three main stages: the preprocessing of raw data, the discovery of schemas and the analysis (or interpretation) of results. A WUM process extracts behavioral patterns from the Web usage data and, if available, from the Web site information (structure and content) and on the Web site users (user profiles).

The quantity of the Web usage data to be analyzed and its low quality (in particular the absence of structure) are the principal problems in WUM. When applied to these data, the classic algorithms of data mining, generally, give disappointing results in terms of behaviors of the Web sites' users (e.g. obvious sequential patterns, stripped of interest).

In this thesis, we bring two significant contributions for a WUM process, both implemented in our toolbox, the AxisLogMiner. We propose a complete methodology for preprocessing the Web logs and a divisive general methodology with three approaches (as well as associated concrete methods) for the discovery of sequential patterns with a low support.

Our first contribution concerns the preprocessing of the Web usage data, which received less attention from the WUM research. The originality of the methodology for WUM preprocessing that we proposed consists in its *Intersites* aspect, essential to apprehend the behaviors of the users that navigate in a transparent way, for example, on several Web sites of the same organization. In addition to the integration of main existing work on this topic, we propose in our methodology four distinct steps: the data fusion, data cleaning, data structuration and data summarization. More precisely, we propose several heuristics for cleaning the Web robots, aggregated variables describing the sessions and the visits, as well as the recording of this data in a relational model. Several experiments were carried out, proving that our methodology allows a strong reduction (up to 10 times) of the initial number of requests and it offers richer logs, structured for the following stage of data mining.

Our second contribution aims at discovering from a large preprocessed log file the minority behaviors corresponding to the sequential patterns with low support. For that, we propose a general methodology aiming at dividing the preprocessed log file into a series of sub-logs. Based on this methodology, we designed three approaches for extracting sequential patterns with low support (the *Sequential*, *Iterative* and *Hierarchical* approaches). These approaches were implemented in hybrid concrete methods using algorithms of clustering and sequential pattern mining. Several experiments, carried out on logs collected from academic sites, enabled us to discover interesting sequential patterns having a very low support (down to 0.009%), while their discovery by a traditional algorithms was impossible.

Finally, we propose a toolbox the *AxisLogMiner*, which supports our preprocessing methodology and, currently, two of the hybrid methods for the discovery of sequential patterns in WUM. This toolbox was used to preprocess several log files and also to experiment on our methods implemented for extracting sequential patterns with low support.

Keywords: *Web Usage Mining, Web logs, WUM Methodology, WUM Preprocessing, Inter-sites WUM, Web Mining, Data Mining, Sequential Pattern Mining, Low Support, Clustering, Divisive Methodology, WUM toolbox, Apriori-GST, AxisLogMiner*