Web Usage Mining through Associative Models

Paolo Giudici  
*University of Pavia, Italy*

Paola Cerchiello  
*University of Pavia, Italy*

**INTRODUCTION**

The aim of this contribution is to show how the information, concerning the order in which the pages of a Web site are visited, can be profitably used to predict the visit behaviour at the site. Usually every click corresponds to the visualization of a Web page. Thus, a Web clickstream defines the sequence of the Web pages requested by a user. Such a sequence identifies a user session.

Typically, a Web usage mining analysis only concentrates on the part of each user session concerning the access at one specific site. The set of the pages seen in a user session, on a determinate site, is usually referred to with the term server session or, more simply, visit.

Our objective is to show how associative models can be used to understand the most likely paths of navigation in a Web site, with the aim of predicting, possibly online, which pages will be seen, having seen a specific path of pages in the past. Such analysis can be very useful to understand, for instance, what is the probability of seeing a page of interest (such as the buying page in an e-commerce site) coming from a specified page. Or what is the probability of entering or (exiting) the Web site from any particular page.

The two most successful association models for Web usage mining are: sequence rules, which belong to the class of local data mining methods known as association rules; and Markov chain models, which can be seen, on the other hand, as (global) predictive data mining methods.

**BACKGROUND**

We now describe what a sequence rule is. For more details the reader can consult a recent text on data mining, such as Han & Kamber (2001), Witten & Frank (1999) or, from a more statistical viewpoint, Hand et al. (2001), Hastie et al. (2001) and Giudici (2003).

An association rule is a statement between two sets of binary variables (itemsets) A and B, that can be written in the form $A \rightarrow B$, to be interpreted as a logical statement: if A, then B. If the rule is ordered in time we have a sequence rule and, in this case A precedes B.

In Web clickstream analysis, a sequence rule is typically indirect: namely, between the visit of page A and the visit of page B other pages can be seen. On the other hand, in a direct sequence rule A and B are seen consecutively.

A sequence rule model is, essentially, an algorithm that searches for the most interesting rules in a database. The most common of such algorithms is the Apriori model, introduced by Agrawal et al. (1995). In order to find a set of rules, statistical measures of “interestingness” have to be specified. The measures more commonly used in Web mining to evaluate the importance of a sequence rule are the indexes of support and confidence.

The support is a relative frequency that indicates the percentage of the users that have visited in succession the two pages. In presence of a high number of visits, as it is usually the case, it is possible to state that the support for the rule approximates the probability a user session contains the two pages in sequence. Therefore, the confidence approximates the conditional probability that in a server session in which has been seen the page A is subsequently required page B.

While the support approximates the joint probability of seeing pages A and B, the confidence approximates the conditional probability that in a server session in which has been seen the page A is subsequently required page B.

The above referred to itemsets A and B containing one page each; however, each itemset can contain more than one page, and the previous definitions carry through. The order of a sequence is the total number of pages involved in the rule. For instance, the rules discussed previously are sequences of order two.

Therefore, the output of a sequence search algorithm (e.g., the a priori algorithm) can be visualised in terms of the sequence rules with the highest interestingness, as measured, for instance, by the support and confidence of the rules that are selected.
Related Content

Management of Data Streams for Large-Scale Data Mining
[www.igi-global.com/chapter/management-data-streams-large-scale/7789?camid=4v1a](www.igi-global.com/chapter/management-data-streams-large-scale/7789?camid=4v1a)

Discovering an Effective Measure in Data Mining
[www.igi-global.com/chapter/discovering-effective-measure-data-mining/10624?camid=4v1a](www.igi-global.com/chapter/discovering-effective-measure-data-mining/10624?camid=4v1a)

Projected Clustering for Biological Data Analysis
[www.igi-global.com/chapter/projected-clustering-biological-data-analysis/11035?camid=4v1a](www.igi-global.com/chapter/projected-clustering-biological-data-analysis/11035?camid=4v1a)

Data Mining for the Chemical Process Industry
[www.igi-global.com/chapter/data-mining-chemical-process-industry/10860?camid=4v1a](www.igi-global.com/chapter/data-mining-chemical-process-industry/10860?camid=4v1a)