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.

An important point that must be made about sequence rules is that they are typically indirect: that is, the sequence $A \rightarrow B$ means that A has been seen before B, but not necessarily immediately before. Other pages might have been seen in between. From an interpretational viewpoint we underline that indirect rules are not much usel; direct rules, for which $A \rightarrow B$ means that B is seen consecutively to A are much more interpretable.

MAIN THRUST

A graphical model [see, for example, Edwards (1995), Jensen (1996), Lauritzen (1996), Cox & Wermuth (1996), Whittaker (1996) and Cowell et al. (1999)] is a family of probability distributions incorporating the conditional independence assumptions represented by a graph. It is specified via a graph that depicts the local relations among the variables (that are represented with nodes). Undirected graphs give rise to symmetric graphical models (such as graphical loglinear models and graphical Gaussian models). Directed acyclic graphs (DAGs) give rise to recursive graphical models, which are used in probabilistic expert systems.

It is known that recursive graphical models consist of a powerful tool for predictive data mining, because of their fundamental assumption of casual dependency between variables. On the other hand, symmetric graphical models can be considered as an important and valid tool in the preliminary phase of analysis because they can show the main relevant association, useful to construct a subsequent recursive model.

Both models have been used and compared with association rules in Web usage mining [see, e.g., Heckerman et al. (2000); Blanc & Giudici (2002) and, for a review, Giudici (2003)]. Although results are comparable, we remark that graphical models are usually built from contingency tables and, therefore, cannot simply take order into account.

We now consider a different model for the analysis of Web usage transactional dataset. Association rules (of which sequence rules are a special instance) are an instance of a local model: they take into account only a portion of the data, that is, that which satisfies the rule being examined. We now consider a global model, for which association patterns are discovered on the basis of the whole dataset. A global model suited to analyse the Web clickstream data is the Markov chain model. Precisely, here we consider discrete Markov chains.

The idea is to introduce dependence between time-specific variables. In each session, to each time point i, here corresponding to the i-th click, it corresponds a discrete random variable, with as many modalities as the number of pages (these are named states of the chain). The observed i-th page in the session is the observed realisation of the Markov chain, at time i, for that session. Time can go from i=1 to i=T, and T can be any finite number. Note that a session can stop well before T: in this case the last page seen is said an absorbing state (end_session for our data).

A Markov chain model establishes a probabilistic dependence between what is seen before time i, and what will be seen at time i. In particular, a first-order Markov chain, which is the model we consider here, establishes that what is seen at time i depends only on what is seen at time i-1.

This short-memory dependence can be assessed by a transition matrix that establishes what is the probability of going from any one page to any other page in one step only. For example, with 36 pages there are 36 X 36 probabilities of this kind.

The conditional probabilities in the transition matrix can be estimated on the basis of the available conditional frequencies. If we add the assumption that the transition matrix is constant in time, as we shall do, we can use the frequencies of any two adjacent pairs of time-ordered clicks to estimate the conditional probabilities.

Note the analogy of Markov chains with direct sequences. It can be shown that a first order Markov chain is a model for direct sequences of order two; a second-order Markov model is a model for direct sequences of order three, and so on. The difference is that the Markov chain model is a global and not a local model. This is mainly reflected in the fact that Markov chains consider all pages and not only those with a high support. Furthermore, the Markov model is a probabilistic model and, as such, allows inferential results to be obtained.

For space purposes, we now briefly consider some of the results that can be obtained from the application of Markov chain models. For more details, see Giudici (2003).

For instance, we can evaluate where it is most likely to enter the site. To obtain this we have to consider the transition probabilities of the start_session row. We can also consider the most likely exit pages. To obtain this we have to consider the transition probabilities of the end_session column. We can also build up several graphical structures that correspond to paths, with an associated occurrence probability. For example, from the transition matrix we can establish a path that connects nodes through the most likely transitions.