Variable Length Markov Chains for Web Usage Mining

José Borges
School of Engineering, University of Porto, Portugal

Mark Levene
Birkbeck, University of London, UK

INTRODUCTION

Web usage mining is usually defined as the discipline that concentrates on developing techniques that model and study users' Web navigation behavior by means of analyzing data obtained from user interactions with Web resources; see (Mobasher, 2006; Liu, 2007) for recent reviews on web usage mining. When users access Web resources they leave a trace behind that is stored in log files, such traces are called *clickstream* records. Clickstream records can be preprocessed into time-ordered sessions of sequential clicks (Spiliopoulou et al., 2003), where a *user session* represents a *trail* the user followed through the Web space. The process of session reconstruction is called *sessionizing*.

Understanding user Web navigation behavior is a fundamental step in providing guidelines on how to improve users' Web experience. In this context, a model able to represent usage data can be used to induce frequent navigation patterns, to predict future user navigation intentions, and to provide a platform for adapting Web pages according to user specific information needs (Anand et al., 2005; Eirinaki et al., 2007). Techniques using association rules (Herlocker et al., 2004) or clustering methods (Mobasher et al., 2002) have been used in this context. Given a set of transactions clustering techniques can be used, for example, to find user segments, and association rule techniques can be used, for example, to find important relationships among pages based on the users navigational patterns. These methods have the limitation that the ordering of page views is not taken into consideration in the modeling of user sessions (Liu, 2007). Two methods that take into account the page view ordering are: tree based methods (Chen et al., 2003) used for prefetching Web resources, and Markov models (Borges et al., 2000; Deshpande et al., 2004) used for link prediction. Moreover, recent studies have been conducted on the use of visualization techniques for discovering navigational trends from usage data (Chen et al., 2007a; Chen et al., 2007b).

BACKGROUND

In (Mobasher, 2006) a review of Web usage mining methods was given and Markov models were discussed as one of the techniques used for the analysis of navigational patterns. In fact, *Markov models* provide an effective way of representing Web usage data, since they are based on a well established theory and provide a compact way of representing clickstream records. Markov models provide the means for predicting a user's next link choice based on his previous navigation trail (Dongshan et al., 2002; Deshpande et al., 2004), and as a platform for inducing user *frequent trails* (Borges et al., 2000).

In a first-order Markov model each Web page is represented by a state. In addition, a *transition probability* between two states represents the estimated probability (according to past usage) of viewing the two connected states in a sequence. Each user session is individually processed to count the number of times each page was visited and the number of times each pair of pages was viewed in a sequence, that is, the number of times a transition was followed. The model is built incrementally by processing the entire collection of user sessions. The transition probability between every pair of connected pages is estimated by the proportion of times the corresponding link was followed after viewing the anchor page. The ratio between the number of times a page was viewed and the total number of page views gives the probability estimate for a user choosing the corresponding page from the set of all pages in the site; we call this probability the *initial probability* of a state. The probability estimate of a trail is given by

the product of the initial probability of the first state in the trail and the probability of the traversed links, that is, the transition probabilities.

We note that, as an alternative for modeling a page as a state it is possible to group pages into categories, for example based on their content. In such a scenario each state corresponds to a Web page category, and state transitions model the user's navigation through page categories. In addition, in the case of Web sites composed of pages dynamically built from database queries, each state will correspond to a query. In this case the user's navigation through the content requests is being modeled rather than the navigation through the content that has been viewed. Finally, for Web sites in which page content changes frequently, mechanism that deal with concept drift, such as the use of a sliding window (Koychev, 2007), can be utilized in order to take into account the change of users' behavior over time.

MAIN FOCUS

Building a Variable Length Markov Chain

A first-order Markov model has the limitation of taking into account only the last viewed page when providing the next link choice prediction probability. Thus, it assumes that user navigation options are influenced only by the current page being viewed. To tackle this limitation a method based on building a sequence of higher-order Markov models and a technique to choose the best model to use in each case (Deshpande et al., 2004) has been proposed. However, we stress that the amount of navigation history a user takes into account when deciding which page to visit next, varies from site to site or even from page to page within a given site. Thus, a method that produces a single model representing the variable length history of pages is potentially valuable for studying user Web navigation behavior.

A Variable Length Markov Chain (VLMC) is a Markov model extension that allows variable length navigation history to be captured within a single model (Bejerano, 2004), and a method that transforms a first-order Markov model into a VLMC was presented in (Borges et al., 2005). The method makes use of a clustering-based *cloning* operation that duplicates states having in-paths inducing distinct conditional probabilities for subsequent link choices.

An n-order VLMC model is obtained by upgrading the corresponding previous-order model; for example, a second-order VLMC is obtained by upgrading a first-order model. The method evaluates one state at a time, measuring the accuracy with which n-order conditional probabilities are represented by the (n-1)-order model. From the input data, the n-order conditional probabilities are induced and compared to the (n-1)-order conditional probabilities represented by the model. If a state is shown to be inaccurate, its accuracy in representing the outlinks' transition probabilities can be increased by separating the in-paths to the state that correspond to different conditional probabilities. Thus, a state that is not accurately representing n-order conditional probabilities is cloned; the number of clones is determined by the required precision (set by a parameter) and a clustering technique that groups in-paths corresponding to similar conditional probabilities is used. As a result we obtain a model in which a transition probability between two states takes into account the path users followed to reach the anchor state prior to choosing the outlink corresponding to the transition.

We note that, when evaluating the accuracy of a state in a n-order model all n-length paths to that state are evaluated, thus, long sessions are pre-processed into (n+1)-grams whose corresponding conditional probabilities are evaluated in sequence. Sessions holding cycles are dealt with in the same way.

In Figure 1 we give a brief example to illustrate the method used to construct a first-order model from a collection of sessions and the resulting second-order model. On the left we present a collection of sessions and the corresponding frequency of occurrence, and, in the middle, we present the first-order model resulting from the sessions. Next to a link we show the number of times the link was followed and the corresponding estimated transition probability. According to the input data, the probability of the next link choice when viewing page D depends on the preceding page, thus, state D is set to be cloned. A clustering method identifies the estimated conditional probabilities from A and B to be closer, and therefore they are assigned to the same state; the conditional probabilities from C are kept separate. Transition probabilities from D and D' are more accurate in the resulting second-order model and in case higher accuracy is need more clones will be created.