Soft Computing manuscript No.
(will be inserted by the editor)

José Borges · Mark Levene

Testing the predictive power of variable history web usage

the date of receipt and acceptance should be inserted later

Abstract We present two methods for testing the predictive power of a variable length Markov chain induced from a collection of user web navigation sessions. The collection of sessions is split into a training and a test set. The first method uses a χ^2 statistical test to measure the significance of the distance between the distribution of the probabilities assigned to the test trails by a Markov model build from the full collection of sessions and a model built from the training set. The statistical test measures the ability of the model to generalise its predictions to the unseen sessions from the test set. The second method evaluates the model ability to predict the last page of a navigation session based on the preceding pages viewed by recording the mean absolute error of the rank of the last occurring page among the predictions provided by the model. Experimental results conducted on both real and random data sets are reported and the results show that in most cases a second-order model is able to capture sufficient history to predict the next link choice with high accuracy.

Keywords Web usage mining · Web navigation · Variable length Markov chain

1 Introduction

Web mining has been defined as the research field focused on studying the application of data mining techniques to web data. More specifically, the field of research focused on developing techniques to model and study user web navigation data has been called *web usage mining*, Mobasher (2004).

José Borges
School of Engineering, University of Porto, R. Dr. Roberto Frias, 4200 - Porto, Portugal E-mail: jlborges@fe.up.pt

Mark Levene
School of Computer Science and Information Systems, Birkbeck, University of London, Malet Street, London WC1E 7HX, U.K. E-mail: mark@dcs.bbk.ac.uk

When users visit a web site, data representing their navigation experience is recorded in server log files. A log file consists of an ordered sequence of web page requests from which it is possible to accurately infer user navigation sessions, see Spiliopoulou et al (2003). A user navigation session, also referred to as a *trail*, is usually defined as sequence of web page views by a single user during a single visit to the web site within a fixed time frame. A collection of user sessions within a site is referred to as the site's usage data.

Techniques have been proposed for mining user web navigation patterns from usage data, see for example Sarukkai (2000). The analysis of such patterns helps us understand the user behaviour when visiting the site, providing guidelines to improve web site design. On the other hand, models for usage data have been proposed aiming at prediction of the next link choice of a user when viewing a given page, see for example Anderson et al (2002). Accurately predicting the next link choice is a step towards web site personalization, Mobasher (2004).

Several models have been proposed for modelling user web data. In Schechter et al (1998) the authors utilised a tree-based data structure to represent the collection of paths inferred from the log data in order to predict the next page accessed. Dongshan and Junyi (2002) proposed a hybrid-order tree-like Markov model to predict web page access, which provides good scalability and high coverage. Chen et al (2003) use a PPM (Prediction by Partial Match) tree that restricts the roots to popular nodes. Assuming that most user sessions start with popular pages, they reduce the model complexity by eliminating branches having a non-popular page as root. Deshpande and Karypis (2004) proposed a technique that builds $k^{th} - order$ Markov models and combines them to include the highest order model covering each state. Techniques are then devised to eliminate states and reduce the model complexity.

In Borges and Levene (2000) we have proposed a first-order Markov model for a collection of user navigation sessions. More recently we have proposed a method that makes use of the cloning operation, Levene and Loizou

(2003), in a way that enables the model to represent higher-order conditional probabilities, Borges and Levene (2005a,b).

In this work, we propose to evaluate the prediction power of a model that takes into account a variable length history in estimating the probability of the next link choice, given a user navigation trail. We present a new way for testing the predictive power of a variable length Markov chain using a χ^2 statistical test. We present a statistical test to assess the model's ability to generalise its predictions for unseen navigation sessions from a test set of trails, by comparing the probabilities estimates of test trails, computed from a model built without these trails, with the probabilities computed from a model built from the full set of trails.

In addition, we present a method to evaluate the model's ability to predict the next link choice on a navigation trail given the preceding navigation history. We first measure the rank of the followed link among the link suggestions given by the model. Then, if the followed link has the top rank we have a perfect hit, and the further away the link is from the top rank the worse the prediction is. To measure the rank deviation we use the *mean absolute error*, which is a standard evaluation metric.

Results from extensive experiments conducted with three real data sets and one randomly generated data set are presented. The results show that there is a statistically significant gain in accuracy when moving from a first-order model to a second-order model. Even higher-order models enable us, in general, to achieve improved accuracy but the gain in accuracy is not statistically significant.

In Section 2 we review the variable length Markov chain concept, in Section 3 we present the proposed method for testing the predictive power of a variable length Markov chain, and in Section 4 we present the method for assessing the prediction of the rank of a link. Section 5 presents the results of our extensive experimental evaluation, and in Section 6 we give our concluding remarks.

2 Variable length Markov chains

A *Variable Length Markov Chain* (simply VLMC), Bejerano (2004), is a stochastic model, which extends the notion of a fixed-order Markov chain, Kemeny and Snell (1960), by allowing a variable length history of states to be captured within the model. In the context of web usage mining using VLMCs address two known problems of fixed-order Markov chains:

i) A first-order (or, more generally, a lower-order) Markov chain is, in most cases, less accurate than a higher-order counterpart, Borges and Levene (2000); Jespersen et al (2003).

ii) A (non-variable) higher-order Markov chain leads to an exponentially larger state space than its lower-order counterpart, Borges and Levene (2000); Deshpande and Karypis (2004).

VLMCs address these two problems by modelling the higher-order dependencies between states of the Markov chain only in cases when they are available and statistically significant. In this way the stochastic model is more accurate on the one hand, addressing (i), and more compact on the other hand, addressing (ii).

To formalise VLMCs in the context of web usage mining, let $T = s_1, s_2, \ldots s_m$, be a trail of web pages through a given web site, i.e a sequence of web pages that were viewed by a user in the course of a navigation session within the site. Using the chain rule, the probability of T is given by

$$P(T) = P(s_1) \prod_{i=2}^{m} P(s_i \mid s_1, \ldots, s_{i-1}). \quad (1)$$

In practice we do not have full information of the conditional probabilities on the right-hand side of (1), and therefore we must estimate these with the knowledge we have available from the past activities of users navigating through the site. Our estimate of the conditional probability, is given by

$$P(s_i \mid s_1, \ldots, s_{i-1}) \approx P_M(s_i \mid s_{i-d_i}, \ldots, s_{i-1}) =$$

$$= \frac{P_M(s_{i-d_i}, \ldots, s_{i-1}, s_i)}{P_M(s_{i-d_i}, \ldots, s_{i-1})}, \quad (2)$$

where $s_{i-d_i}, \ldots, s_{i-1}$ is the *maximal* suffix of the trail $s_1, s_2, \ldots, s_i$, that we have at our disposal from the navigation sessions for estimating this conditional probability, and P_M represents the transition probability distribution of the d_i-order Markov chain for this conditional statement; we call d_i the *context* of the trail.

In the extreme case, when $d_i = 0$, we estimate the left-hand side of (2) by $P_M(s_i)$, i.e. the proportion of times users visited s_i; in this case the probability of a user being at s_i is independent of the user's history. (If there were no visits to s_i, then the best we can do is a uniform random guess of the web page the user will be at next.) When $d_i = 1$ we have the conventional (first-order) Markov chain, which is often a good approximation of the conditional probability, taking into account only the previous page a user has visited. The advantage of using VLMCs is that we get the best accuracy we can manage according to the cumulative available information at our disposal, which increases over time.

For a trail $T = s_1, s_2, \ldots, s_m$, we let $C(T)$ denote the frequency count of T and the estimate of the conditional probability (2) using the frequency counts, be

$$P_M(s_i \mid s_{i-d}, s_2, \ldots, s_{i-1}) \approx \frac{C(s_{i-d_i}, \ldots, s_{i-1}, s_i)}{C(s_{i-d_i}, \ldots, s_{i-1})}, \quad (3)$$

where, as before, d_i is the largest order of the Markov chain at our disposal for estimating the conditional probability $P(s_i \mid s_1, \ldots, s_{i-1})$.

The test for the predictive power of a VLMC that we propose in the next section can be used with any algorithm for constructing a model of user navigation sessions, as long as the frequency counts, as in (3), can be extracted from the model to compute the conditional probabilities. For the purpose of this paper we utilise the dynamic higher-order Markov model, proposed in Borges and Levene (2005a), to build a VLMC from the user navigation sessions as reconstructed from a web server log file. We now briefly describe the basic features of this model.

In the dynamic higher-order Markov model we build a VLMC by starting from a first-order model and adding states to the model whenever there is a significant difference between a lower order probability of order i and a higher-order probability of order $i+1$.

Initially a first-order model is constructed from a collection of trails representing the navigation sessions of a user (or a group of users) within a web site. Each state is labelled by the web page visited and a transition between states has associated with it the number of times the corresponding hyperlink was traversed. Transition counts are used to estimate the transition probabilities. For every session, there is one artificial start state S and one artificial final state, F. In Table 1 we show an example of a collection of sessions and in Figure 1 the corresponding first-order model.

In the first-order model, the probability attached to each out-link represents a first-order transition probability. In order to assess the second-order accuracy of a state, each first-order probability attached to an out-link from the state is compared to the corresponding second-order probability. When the first and second-order probabilities diverge, then a new state, called a *clone*, is added to the model, such that its in-links have the designated second-order probabilities. For the example in Figure 1, $P(A_3|A_2) = 0.375$ and according to the input data given in Table 1 $P(A_3|A_1, A_2) = 0.75$, therefore, a clone, A'_2 of the state A_2, is added to the model, and the link (A_1, A'_2) is re-directed to it. In addition, the out-link counts are updated so that its transition probabilities reflect the second-order probabilities. The modified second-order model is shown in Figure 2.

We note that the reason the state A_2 was cloned and its in-links distributed between itself and the clone, was that the out-link transition probabilities from state A_2 did not correspond to the second-order probabilities derived from the 3-gram counts of the user's navigation sessions. In the model given in Figure 2, the transitions from A_2 and its clone, A'_2, are now accurate with respect to the second-order probabilities given by the input data and thus no more cloning needs to take place. In Borges and Levene (2005b) the details of a generalisation of the method to third and higher-order transition probabilities are presented.

Session	Occurrences
A_1, A_2, A_3	3
A_1, A_2, A_4	1
A_5, A_2, A_4	3
A_5, A_2, A_6	1

Table 1 A collection of user navigation sessions

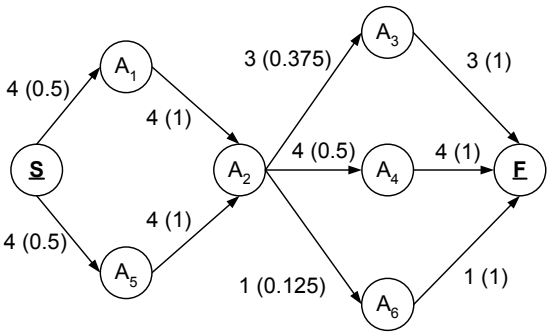

Fig. 1 The first-order model corresponding to the collection of sessions given in Table 1

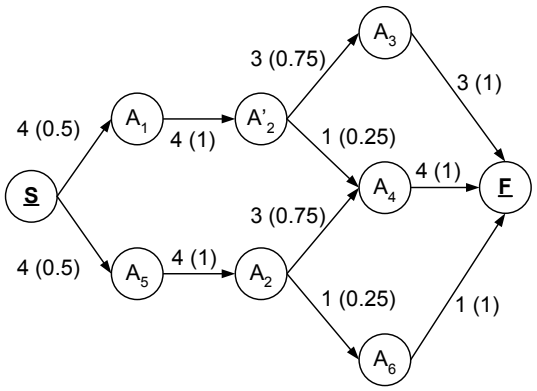

Fig. 2 The modified second-order model corresponding to the collection of sessions given in Figure 1

We may now estimate the probability of a trail $T = s_1, s_2, \ldots, s_m$, using (1), by replacing the probabilities with their estimates from the dynamic higher-order model using (2) and (3). In fact, the conditional probability of a trail is given by the transition probabilities of the last link in the trail. In the example in Figure 2, we have that $P_M(A_4|A_1, A_2) = 0.25$ and $P_M(A_4|A_5, A_2) = 0.75$.

VLMCs can be used to compute the probability of a trail as described above, or can be used, with the aid of (3), to predict the next web page, s_i, that a user may view given that he or she have already viewed pages, $s_1, s_2, \ldots, s_{i-1}$.

3 Testing the predictive power of a VLMC

Recently, several Markov model extensions have been proposed based on the intuition that, in many cases, a first-order model does not provide good enough accuracy in representing user navigation records. However, this intuition was rarely supported by statistical evidence. To the best of our knowledge, Deshpande and Karypis (2004) is the only related work that has reported on the use of statistical tests to assess the power of a Markov model in predicting web page accesses. In particular, the authors present the results of tests on the difference in the number of times that a first-order and a higher-order model make a correct prediction.

Herein, we propose a different approach to assess the predictive power of a Markov model. Instead of evaluating the overall proportion of correct predictions, as in the test described in Deshpande and Karypis (2004), we make use of the χ^2 statistical test, thus providing a finer level of detail in comparing the probability distributions induced by different models.

In order to test the predictive power of a VLMC we propose to use k-fold cross validation, Mitchell (1997), on the collection of input trails, I, where each trail represents a single trail (i.e. navigation session) derived from the web log input data. (Note that a trail may appear several times in I; see Table 1.)

We partition the set I into k random subsets of trails of roughly equal size, and build a VLMC, denoted by VLMC-i, from $k-1$ of the subsets by omitting the ith subset, where $1 \leq i \leq k$. In addition, we build a VLMC from the full data set, which we denote by VLMC-0. We will use VLMC-0 as the model we will test the prediction against, as it is the most complete one we have given the data set.

For each subset of trails we use VLMC-i to predict the probability of all the trails, $T_i = \{t_1^i, t_2^i, \ldots, t_w^i\}$, that are *not* in this subset, where we consider duplicate occurrences of a trail to be distinct. (Note that, in general, there may be duplicate trails in T_i, so strictly speaking T_i is a bag of trails.) Moreover, there may be trails in T_i that do not appear in VLMC-i, in which case we test the longest non-null suffix of the trail that appears in VLMC-i instead of the full trail, if such a suffix exists, otherwise we eliminate such a trail from T_i. (If no suffixes of a certain trail appear in VLMC-i then, as mentioned before, the best prediction we can make is a uniform random guess.)

Let P_0 and P_i be the probability distributions of VLMC-0 and VLMC-i, respectively, and let $P_0(t_j^i)$ be the estimate of the probability of the trail t_j^i using VLMC-0. In addition, let

$$P_i(t_j^i) = \frac{C(t_j^i)}{C^-(t_j^i)},$$

where $C(t_j^i)$ denotes the count of trail t_j^i in VLMC-i, and $C^-(t_j^i)$ denotes the count of the prefix of t_j^i in VLMC-i, i.e. the count of the trail resulting from removing the last node from t_j^i.

The χ^2-test statistic for goodness of fit, Siegel and Castellan Jr. (1988), of VLMC-i to VLMC-0, is now given by

$$\chi_i^2 = \sum_{j=1}^{\nu_i} \frac{(C(t_j^i) - C^-(t_j^i)P_0(t_j^i))^2}{C^-(t_j^i)P_0(t_j^i)}, \tag{4}$$

where ν_i is the number of degrees of freedom of the set T_i, which in this case is the number of unique trails in T_i minus one.

The statistic, χ_i^2, approximately follows a χ^2 distribution with ν_i degrees of freedom. Therefore, we can compute the probability of having a value greater than or equal to χ_i^2 given the corresponding χ^2 distribution. If the probability is below the significance level of the test (usually taken at 5%) the null hypothesis is rejected, which means that the probability deviation between VLMC$-i$ and VLMC-0 is concluded as being significant. Otherwise, if the probability is above the 5% level, the null hypothesis is not rejected, meaning that there does not exist sufficient evidence to reject the null hypothesis that there is no difference between the two models. The probability of wrongly rejecting the null hypothesis if it is in fact true, is called the p-value, and the larger it is the smaller the evidence in favour of rejecting the hypothesis being tested.

For a large number of degrees of freedom, such as the case here, the Wilson-Hilferty cube root transformation, Wilson and Hilftery (1931), can be applied to a χ^2 random variable to obtain approximate normality. In particular, the test statistic

$$Z = \frac{(\chi_i^2/\nu_i)^{1/3} - (1 - 2/(9\nu_i))}{(2/(9\nu_i))^{1/2}}, \tag{5}$$

is approximately normally distributed, with expected value 0 and standard deviation 1, and can therefore be tested against a standard one-sided null hypothesis that VLMC-i and VLMC-0 come from the same probability distribution, at a level of significance of our choice.

Alternatively, the power-approximation proposed in Canal (2005) can be used, which has numerically been shown to be more accurate for χ^2 variables with a large number of degrees of freedom. It is given by

$$L = \left(\frac{\chi_i^2}{\nu_i}\right)^{1/6} - \frac{1}{2}\left(\frac{\chi_i^2}{\nu_i}\right)^{1/3} + \frac{1}{3}\left(\frac{\chi_i^2}{\nu_i}\right)^{1/2}, \tag{6}$$

which follows a normal distribution with the following expected value and variance,

$$E(L) = \frac{5}{6} - \frac{1}{9\nu_i} - \frac{7}{648\nu_i^2} + \frac{25}{2187\nu_i^3}, \tag{7}$$

$$Var(L) = \frac{1}{18\nu_i} + \frac{1}{162\nu_i^2} - \frac{37}{11664\nu_i^3}. \tag{8}$$

Combining the results from the k tests for each VLMC-i, we use k'/k as our metric for the predictive power of VLMC-0, where k' is the number of VLMCs for which the null hypothesis was not rejected according to (5) or (6). We call k'/k the *predictive power metric*. In the latter case we have to standardize the value given in (6) by subtracting its expected value and dividing by the standard deviation, which are computed according to equations (7) and (8).

We will now give a short example to illustrate the procedure. Assume a collection of trails that was split into k folds, where VLMC-0 is induced from the full collection of trails and VLMC$-i$ is induced from the collection of trails included in the first $k-1$ folds. The collection of trails in the k^{th}-fold will be our test set.

Assume that the first trail in the k^{th}-fold is $t_1^k = A_1, A_2, A_3$, and that from the $n-$gram counts we obtain that the count of its prefix A_1, A_2 in VLMC$-i$, is $C^-(t_1^k) = 20$. In addition, assume that from VLMC-0 we obtain, $P_0(t_1^k) = 0.45$ and from VLMC$-i$ we obtain, $P_k(t_1^k) = 0.50$. The conditional probabilities correspond to the probability of following link A_2, A_3 given that the user arrived at A_2 from A_1 according to VLMC-0 and VLMC$-i$, respectively.

Now, $P_k(t_1^k) \cdot C^-(t_1^k) = C(t_1^k) = 20 \cdot 0.50 = 10$, gives the number of times the path A_1, A_2, A_3 was followed according to the model represented by VLMC$-i$, and $P_0(t_1^k) \cdot C^-(t_1^k) = 20 \cdot 0.45 = 9$ gives the corresponding expected value according to the full model, which we consider to be our ground truth. As a result, the contribution to the test statistic χ_i^2 given by this trail is $(10-9)^2/9 = 1/9$.

Each trail in the test set, that is, the k^{th}-fold collection of trails, is processed in the same way. However there are two cases when a trail is ignored: (i) in the case when an identical trail was processed before; and (ii) in the case when the count $P_i(t_j^i) \cdot C^-(t_j^i)$ is below 5, since it is a requirement of the χ^2-test to have the expected counts above 5 in order to guarantee that the test statistic follows a χ^2 distribution.

In situations, as in case (ii) above, when the count of a trail of length j is below 5, we discard the first state on the trail and re-assess its $j-1$ length suffix. If, the suffix count is also below 5 we remove another state, until the count is above 5 or there are no more states to remove.

4 Assessing next link-choice prediction ranking

As an alternative to the method presented in Section 3 we present a method to assess the capability of a VLMC model to predict the next link choice of a user based on the user's previous navigation steps. To do this we randomly split the set of navigation trails into a training set and a test set, and we then induce a VLMC model from the training set. For each test trail of length, say m, we use its $m-1$ length prefix to predict the last page on the trail. The VLMC model gives the set of pages reachable from the tip of the prefix, after the prefix has been followed. The reachable pages are ranked by the probabilities of the VLMC model, where rank 1 refers to the page with the highest probability and rank n refers to the page with the lowest probability. The rank, r, of the last page on the test trail is used to measure the prediction accuracy, and thus we measure the strength of the prediction as $r-1$. We use the *Mean Absolute Error* (MAE) as the overall metric, which is defined as the sum of the prediction accuracies for each of the trails in the test set divided by the number of trails in the test set.

We now illustrate the method with an example. Consider the model in Figure 2 as the VLMC we want to assess, and a test set containing three trails: (i) A_1, A_2, A_4, (ii) A_5, A_2, A_4 and (iii) A_1, A_2, A_6. For the first trail we have the prefix A_1, A_2 which according to the model gives A_3 as the most probable next page and A_4 as the second more probable, since the last page in the trail is A_4 we have that $r-1 = 1$. For the second trail the prefix is A_5, A_2 and since A_4 is the most probable next page we have that $r-1 = 0$. For the last trail the last page is not in the set of reachable pages from the prefix, therefore the rank of A_6 is the number of reachable pages plus one, that is, $r-1 = 2$. In this example MAE will be $(1+0+2)/3 = 1$.

The method described in this section has the advantage that it does not make use of a ground truth that is model dependent, but has the disadvantage that it does not provide a statistical test which assesses the prediction accuracy of the model. The formal description of the method now follows as Algorithm 1.

Algorithm 1 (MAE(I, k))

1. *Partition the set of input trails, I, into k random subsets of roughly equal size.*
2. *Let w_i be the cardinality of a subset of trails $T_i = \{t_1^i, t_2^i, \ldots, t_w^i\}$, where $1 \leq i \leq k$.*
3. *Build a VLMC$-i$ from $k-1$ subsets by omitting the ith subset.*
4. *For each trail t in the ith subset let t' be the trail t minus its last state.*
5. *Find the longest suffix of t', say t'', that has a match in VLMC$-i$.*
6. *Rank all the pages reachable from the last state of t'' from 1 to n, where 1 refers to the page with highest probability and n to the page with the lowest probability.*
7. *Find the rank, r, of the last state in trail t and subtract 1 from it to obtain $r-1$, which is defined as the partial MAE_t.*
8. *Compute the overall metric for VLMC$-i$ as $MAE_i = \sum_{t=1}^{w_i} MAE_t / w_i$.*
9. *Compute the overall metric for the complete set of input trails as $MAE = \sum_{i=1}^{k} MAE_i / k$.*

5 Experimental Evaluation

5.1 Data sets description

We conducted experiments with three real data sets and one randomly generated data set. The first data set (CS) was made available by the authors of Spiliopoulou et al (2003). It is from a university site and corresponds to two weeks of site usage in 2002; sessions were inferred using cookies. The second data set (MM) corresponds to two weeks of usage records from the Music Machines site (`machines.hyperreal.org`) in 1999 and was obtained from the authors of Perkowitz and Etzioni (2000). The third data set (LTM) represents a month of usage in January 2003 from the London Transport Museum web site (`www.ltmuseum.co.uk`).

The CS data set had sessions already identified. For the other three data sets a session was defined as a sequence of requests from the same IP address with a time limit of 30 minutes between consecutive requests. Erroneous and image requests were eliminated, but .jpg requests were left for MM data set, since in that specific site they correspond to page views.

We also created a random data set (RDS) using the same method used in Borges and Levene (2005a). Briefly, given the number of pages in the site, we generate a topology by matching in-links and out-links according to power-law distributions with exponents that have been observed in "real" web sites. We then compute the pagerank as a measure of users interest in pages and generate sessions according to the random surfer model principle underlying pagerank.

Table 2 gives the summary characteristics of each data set. For each data set (DS) we indicate the number of pages occurring in the log file (pages) and the total number of requests recorded in the log file (requests). Also, we give the total number of sessions in each data set (NS) and the number of sessions of length one (l=1), two (l=2) and three (l=3); a session length is measured by the number of requests it is composed of.

			sessions			
DS	pages	requests	NS	l=1	l=2	l=3
CS	547	115448	24548	7148	3474	2202
LTM	1362	372434	47021	13489	3428	1893
MM	8237	303186	50192	12644	7891	4925
RDS	3855	84801	29663	4416	13049	6501

Table 2 Summary characteristics of the data sets

Figure 3 shows the variation in the number of resulting states necessary to accurately represent higher-order probabilities. The results show that for the RDS data set, the largest increase of the number of states occurs when moving from a first-order model to a second-order model. The other data sets reveal a larger increase in the number of states for higher-order models. These results suggest that the behaviour represented by RDS requires a shorter memory when deciding which in-link to follow.

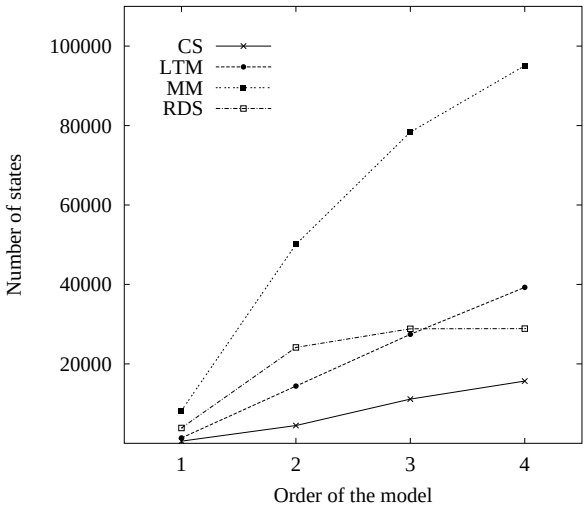

Fig. 3 The increase in the number of states with the order of the model

5.2 χ^2 statistical test evaluation

For the experimental evaluation each data set was randomly split into $k = 5$ folds, and VLMC-i was constructed from $k-1$ folds while VLMC-0 was constructed from the full data set. VLMC-0 is the closest model we have to the ground truth, since it was constructed from the full data set. By comparing the probabilities given by VLMC-i to those given by VLMC-0 we aim to assess the ability of the VLMC-i model to predict the probability of unseen trails, as described in Section 3.

Table 3 presents the χ^2-test statistic for each goodness of fit test and each result given corresponds to the average of the five runs, where each run corresponds to one of the five partitions of the data. As a further validation we carried out several additional tests using 10-fold cross validation with similar results. We note that for larger values of k we expect the results to be at least as good as for lower k since the test set is smaller and, therefore, the models created are closer to each other.

For each data set (DS) and each Markov chain order (Or.) we give the test-statistic value (χ^2) computed according to (4), the corresponding number of degrees of freedom (df) and the p-value (p-val). Note that, in this case the p-value corresponds to the probability of having a value above the test statistic value according to the χ^2 probability distribution with the given number of degrees of freedom. For example, in the evaluation of the first-order model on the CS data set, for a χ^2 distribution

DS	Or.	χ^2	df	p-val	WH	PA	p-val
CS	1	3576.9	676	0	40.8	42.2	0.000
	2	586.8	746	1	-4.5	-4.4	0.999
	3	207.9	744	1	-20.0	-20.3	1.000
	4	135.1	755	1	-25.4	-26.0	1.000
LTM	1	9801.1	1591	0	70.3	73.0	0.000
	2	1037.9	1788	1	-15.0	-14.9	1.000
	3	347.8	1822	1	-38.4	-39.2	1.000
	4	230.4	1821	1	-45.1	-46.6	1.000
MM	1	3458.8	1400	0	27.9	28.2	0.000
	2	425.1	1695	1	-32.3	-32.8	1.000
	3	297.8	1724	1	-39.0	-40.0	1.000
	4	292.6	1723	1	-39.3	-40.3	1.000
RDS	1	74.9	446	1	-20.0	-20.5	1.000
	2	80.6	484	1	-20.9	-21.5	1.000
	3	80.6	484	1	-20.9	-21.5	1.000
	4	80.6	484	1	-20.9	-21.5	1.000

Table 3 Summary results of the χ^2 statistical test for up to fourth order with each data set and the corresponding test results while using normal approximations for the p-value computation

with 676 degrees of freedom we have that the p-value is given by $P(\chi^2 > 3576.9) \approx 0.0$. As usual, if the p-value is below 5% the null hypothesis is rejected, meaning that there is strong evidence that the probabilities of VLMC-i are different from those of VLMC-0. In cases where p-value is above 5% we can say that there is no statistical evidence against the null hypotheses and, therefore, VLMC-i was able to capture the essential characteristics from the training data in order to accurately predict the probability of the trails in the test set.

We observe that the results were very consistent across the different k-fold partitions. In fact, the predictive power metric gives $k'/k = 5/5$ when the p-value $= 1$ and $k'/k = 0/5$ when the p-value $= 0$, meaning that results were identical across the partitions.

The results presented in Table 3 show that for the three real data sets a first-order model is not accurate enough. In fact, only for second, or higher, order models we are able to accurately predict the probability of a trail. On the other hand, for the RDS data set probabilities are accurate even for the first-order model. The latter result suggests that the navigation represented by RDS corresponds to an inherently low-order behaviour. In fact, our experience with random data generators suggests that it is difficult to devise a random data generator that is not first-order. If we randomly generate trails on a web topology the trails tend to be first-order unless we use a more sophisticated model of users' behaviour. (We note that, we could not find in the literature a good random generator for user web navigation sessions.)

In addition, Table 3 gives the Wilson-Hilferty normal approximation (WH) according to (5), the standardized power approximation (PA) according to (6), and the p-value corresponding to PA; the p-value corresponding to WH was almost the same. Using such approximations is important, since for a very large number of degrees of freedom it may not be possible to obtain the p-value for the χ^2 probability distribution. Since PA follows a standard normal distribution (Z), for example, the p-value for the CS data set on the first-order model corresponds to the $P(Z > 42.2) \approx 0.0$. The results presented for the normal approximation are consistent with those reported for the χ^2-test, providing further verification of the validity of the approximation in this particular context. A negative value for WH or PA corresponds to a test statistics value on the left tail of the standard normal distribution, which in the case of a right-tailed test and for the values given corresponds to a p-value approximately equal to 1.0. Therefore, in this case, there is no statistical evidence against the null hypothesis.

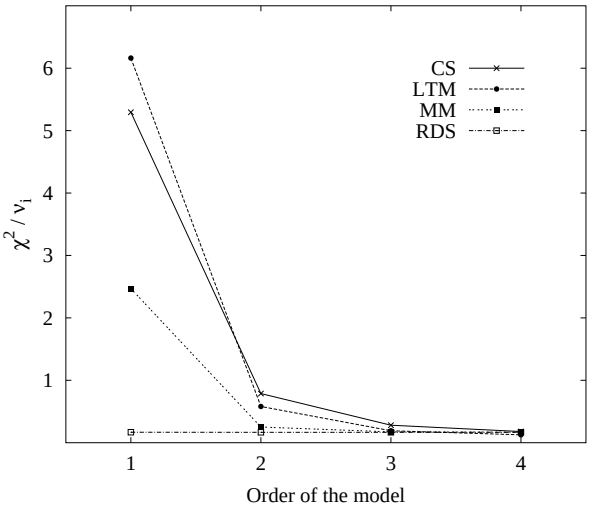

Fig. 4 Variation of the ratio between the χ^2 test statistic and the number of degrees of freedom with the order of the model

Equation 4 provides a measure of relative proximity between two probability distributions, and the number of degrees of freedom corresponds to the number of trails whose probability was tested. Therefore, it is possible to look at the ratio between the χ^2 statistic and the number of degrees of freedom, ν_i, as the average deviation between the trail probability given by VLMC-0 and the corresponding probability given by VLMC-i. Figure 4 shows that, as the order of a model increases, the probabilities given by VLMC-i get closer to the corresponding probabilities given by VLMC-0. The ratio analysis provides an intuitive measure of a model's relative accuracy in representing the input navigation data, which is consistent with the statistical tests we have conducted.

In our opinion the results presented in this section are important on two counts. First, to the best of our knowledge this is the first time that a statistical test is used to evaluate the depth of memory required in user navigation. Second, we are able to confirm that for some web sites a second-order model provides good accuracy

in representing navigation sessions, while for other sites a higher-order model may be necessary.

5.3 Rank-based prediction evaluation

This section presents the results for the Mean Absolute Error (MAE) metric for model evaluation. In this set of experiments we also use the data sets described in Section 5.1. Each data set was split into $k = 5$ subsets and a model was built from $k - 1$ subsets, leaving the kth subset as a test set.

In Table 4, for each data set (DS) we present an average result of the five k-fold evaluations. For each run we give the number of trails in the test set (trails), the respective average trail length (ATL), and the average number of out-links from the penultimate state in a trail, that is, the branching factor (bf). The branching factor indicates the number of link-choice options available when choosing the link to the last state on the trail, and is therefore an indication of the difficulty in predicting the last state of a trail based on the sequence of states preceding it.

For each data set we built models up to the fourth-order. For each model we indicate its order (Or.) and the corresponding mean absolute error value (MAE) as computed by Algorithm 1. Moreover, we make use of a metric which corresponds to the complement of MAE (MAE_c). MAE_c is similar to MAE but measures the rank relative to the bottom of the list of link choices. A higher MAE_c corresponds to a higher branching factor implying that it is harder to predict the actual link choice. Since for given data set the sum of MAE and MAE_c is constant and equal to the branching factor value , we use the ratio as our normalised measure of the quality of the prediction (MAE_c/bf). In addition, we give the average length of the matching suffixes (SL) in the model for the test trails. The difference between trail length and the suffix length measures the amount of information discarded from a given test trail when finding the longest matching suffix on the model. Finally, (top1), (top2) and (top3) indicate the percentage of predictions which where in the top 1, top 2 and top 3 ranks, respectively.

The results given in Table 4 indicate that the biggest drop in MAE occurs when moving from a first-order model to the corresponding second-order model; see also Figure 5 for a graphical representation of how the MAE varies with the order of the model. It follows that the biggest gain in accuracy occurs when extending a first-order model to accurately model second-order probabilities. For the three real data sets there is always some gain in moving to higher-order models, although in some cases the gain is quite small. On the other hand, for the RDS data set a second-order model is sufficient and there is no gain in moving for a more complex model.

Regarding the suffix length analysis we note that among the real data sets the MM data reveals the highest proportion of a trail being discarded when predicting. In fact, for an average trail length of 5.6 states only 3.5 are used for prediction. We note that, in general, the higher the order of the model the shorter is the suffix used to predict the last state. This is expected, since the set of trails modelled by a higher order model is a proper subset of the trails modelled by a corresponding lower order model. One possible interpretation for a large difference between the trail length and the suffix length is that the user's navigation behaviour is not stable in the sense that different trails of similar length do not tend to be followed in exactly the same way. In such cases lower order models are more adequate.

DS	trails	ATL	bf	Or.	MAE	$\frac{MAE\_c}{bf}$	SL	top1	top2	top3
CS	3278.8	4.4	58.8	1	4.8	0.918	3.9	37%	55%	64%
				2	3.5	0.940	3.3	40%	61%	68%
				3	3.1	0.947	3.0	40%	63%	71%
				4	3.0	0.949	3.0	41%	64%	72%
LTM	5968.2	5.7	71.5	1	7.6	0.894	4.9	57%	61%	64%
				2	4.7	0.934	4.2	59%	66%	71%
				3	4.2	0.943	3.9	59%	68%	73%
				4	4.0	0.945	3.7	59%	69%	74%
MM	7036.4	5.6	81.2	1	15.2	0.813	3.5	27%	39%	46%
				2	9.7	0.882	2.5	34%	48%	56%
				3	8.9	0.890	2.2	34%	51%	60%
				4	8.8	0.892	2.2	34%	52%	60%
RDS	5001.0	2.2	15.0	1	10.4	0.307	1.3	3%	15%	24%
				2	8.2	0.453	1.3	4%	23%	34%
				3	8.2	0.453	1.3	4%	23%	35%
				4	8.2	0.453	1.3	4%	23%	35%

Table 4 Summary results on the accuracy of predicting the last state of a trail for different models' orders

The analysis of the results reveals that for the LTM data set a first-order model is able to produce accurate predictions in 57% of the cases. In addition, for a fourth-order model the page chosen is among the top three suggestions in 74% of the cases. For the MM data set, although the top suggestion is accurate in less than 35% of the cases, if the top three suggestions are taken into account then the link chosen is among the suggestions in close to 60% of the cases for third-order (or above) models. The prediction for the RDS data set is not accurate even for a fourth-order model. This result supports our intuition that the model behind the user navigation session generator is inherently random, and, therefore, it is hard to base predictions on it. When assessing the quality of such results it is important to keep in mind the value of MAE_c/bf, which in this case is 0.453, indicating low prediction accuracy.

Finally, Tables 5, 7, 8 and 6 show examples of the application of the method to trails from the LTM data set. For this example we make use of a third-order model and the first partition. For each trail the sequence of URLs followed is given (seq.) and the target page (tg), corresponds to the page we want to predict based on the anterior URLs. The reachable pages correspond to the pages reachable from the page before the target and the probability according to the third order model is indi-

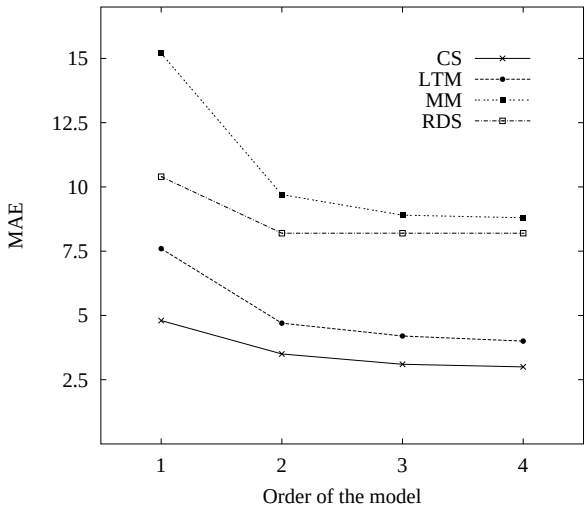

Fig. 5 Variation of the MAE metric value with the order of the model

seq.	URL
0	/
1	/exhibitions/index.html
2	/exhibitions
3	/exhibitions/past.html
4	/exhibitions
5	/exhibitions/online_exhibitions/tfl/index.html
tg	/exhibitions
prob.	reachable pages
0.75	# /exhibitions
0.25	/exhibitions/online_exhibitions/tfl/gla2.html
0	/visiting/index.html

Table 5 The prediction rank-based method applied on a trail where the target page is the top rank prediction

seq.	URL
0	/
1	/exhibitions/current.html
2	/exhibitions
3	/exhibitions/online_exhibitions/wheels/index.html
4	/exhibitions
5	/exhibitions/online_exhibitions/wheels/template_02.html
tg	/exhibitions
prob.	reachable pages
0.8	# /exhibitions
0.15	/exhibitions/online_exhibitions/wheels/template_03.html
0.05	F (end session)
0	/exhibitions/online_exhibitions/wheels/template_04.html
0	/collections/visual_services.html
0	/exhibitions/future.html
0	/exhibitions/permanent_12_b.html

Table 6 A second example of the application of the prediction rank-based method applied on a trail where the target page is the top rank prediction

cated. We use the # symbol to emphasise the target page among the reachable pages. A probability with value 0 indicates that there is a link connecting the pages but the link was not followed given the exact sequence of links represented by the trail. For trails in Tables 5 and 6 the target page is the top rank, for trail in Table 7 it is

seq.	URL
0	/learning/school_resources/gnvq/teachers.html
1	/learning/school_resources/gnvq/index.html
2	/learning/school_resources/gnvq/scripts.js
3	/learning/school_resources/gnvq/case_study.html
tg	/learning/school_resources/gnvq/teachers.html
prob.	reachable pages
0.33	/learning/school_resources/gnvq/index.html
0.33	/learning/school_resources/gnvq/1
0.33	# /learning/school_resources/gnvq/teachers.html
0	/visiting/map_depot.html
0	/visiting/index.html
0	/learning/school_resources/gnvq/r1
0	F (end session)

Table 7 The prediction rank-based method applied on a trail where the target page is in the top three rank prediction

seq.	URL
0	/usl
1	/usl/usl2.html
2	/usl/teach/tz1.html
tg	/usl/teach/tz2.html
prob.	reachable pages
0.5	/learning/school_post.html
0.5	F (end session)
0	/usl/activities/ac3/ac32c.html
0	# /usl/teach/tz2.html

Table 8 An example of the application of the prediction rank-based on a trail where the target page, although reachable, the corresponding link was never followed after viewing the previous pages composing the trail

in the top three ranks, and for trail in Table 8, although the page is reachable, the link was never followed and thus the page is assigned rank $r = 3 - 1 = 2$ since there are two pages having a higher probability.

6 Concluding Remarks

In this work we have presented a study aiming at assessing the predictive power of a variable length Markov chain model induced from a collection of users web navigation sessions. Two methods have been proposed, the first uses a χ^2 statistical test to measure the distance between the probability distribution of a model built from the full collection of trails and a model induced from a subset of such collection. The χ^2 test enables us to assess the model's ability to generalise for unseen trails. The second method evaluates the model ability to predict the last page view of a navigation session based on the preceding page views. A metric was defined to measure the rank of the predicted page view from a choice of possible ones. The higher the rank the better the prediction is. While the first method has the advantage of providing a formal statistical method for testing the predictive power, the second method has the advantage of being independent of a model specific ground truth.

We reported on a set of experiments conducted with three independent real data sets and one randomly generated data set. We note that similar results were ob-

tained from the two methods conducted in the sense that they both indicate that a second-order model is, in most cases, enough to capture the the portion of a user's navigation behaviour which is necessary to predict the user's next link choice. It was also verified that both the Wilson-Hilferty and the power approximation are valid in the particular context of the statistical test presented.

A meaningful result was obtained for the LTM data set, which revealed that a first-order model is able to accurately predict the last page on a trail in 56.5% of the cases. The lowest first-order prediction rate from the real data sets was obtained for the MM data set with a 27% precision rate. Moreover, the first-order model is able to place the prediction target into the top three predictions in 46% of the cases for the MM data set, in 64% of the cases for the CS data set and in 64% of the cases in the LTM data set. With a fourth-order model, those values increase to 60%, 72% and 74% respectively. The results for predictions based on the RDS data set are, in general, not as good, suggesting that further work is needed on developing a more realistic user web navigation sessions generator.

In addition, we would like to mention that another way to test the predictive power of the model is to split I into k sets in such a way that all the trails in the ith subset, $i < k$, temporally precede the trails in the $i+1$th subset. We then form VLMC-i, $1 \leq i \leq k$, from all the trail subsets from 1 to i, and perform the χ^2-test statistic for goodness of fit of VLMC-i to VLMC-$(i+1)$, with $i < k$, with VLMC-$(i+1)$ replacing VLMC-0 in (4), and then utilising the approximation given in (5) or (6); note that for this test VLMC-k = VLMC-0. As further work we intend to compare this method to the one we have used in the paper.

Finally, we plan to measure the ability of the VLMC model to accurately represent sessions which are longer than the model's order and to study the impact this will have on the prediction accuracy.

References

Anderson CR, Domingos P, Weld DS (2002) Relational markov models and their application to adaptive web navigation. In: KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Press, New York, NY, USA, pp 143–152

Bejerano G (2004) Algorithms for variable length Markov chain modelling. Bioinformatics 20:788–789

Borges J, Levene M (2000) Data mining of user navigation patterns. In: Masand B, Spiliopoulou M (eds) Web Usage Analysis and User Profiling, Lecture Notes in Artificial Intelligence (LNAI 1836), Springer-Verlag, Berlin, pp 92–111

Borges J, Levene M (2005a) A clustering-based approach for modelling user navigation with increased accuracy. In: Proceedings of the 2nd International Workshop on Knowledge Discovery from Data Streams, Porto, Portugal, pp 77–86

Borges J, Levene M (2005b) Generating dynamic higher-order markov models in web usage mining. In: Jorge A, Torgo L, Brazdil P, Camacho R, Gama J (eds) Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), Springer-Verlag, Porto, Portugal, Lecture Notes in Artificial Intelligence (LNAI 3271), pp 34–45

Canal L (2005) A normal approximation for the chi-square distribution. Computational Statistics and Data Analysis 48:803–808

Chen Z, Fowler R, Fu AC, Wang C (2003) Fast construction of generalized suffix trees over a very large alphabet. In: Proceedings of International Conference on Computing and Combinatorics (COCOON), Big Sky, MT, pp 284–293

Deshpande M, Karypis G (2004) Selective Markov models for predicting web page accesses. ACM Transactions on Internet Technology 4:163–184

Dongshan X, Junyi S (2002) A new Markov model for web access prediction. IEEE Computing in Science & Engineering 4:34–39

Jespersen S, Pedersen T, Thorhauge J (2003) Evaluating the markov assumption for web usage mining. In: Proceedings of the fifth ACM international workshop on Web information and data management, pp 82–89

Kemeny J, Snell J (1960) Finite Markov Chains. D. Van Nostrand, Princeton, NJ

Levene M, Loizou G (2003) Computing the entropy of user navigation in the web. International Journal of Information Technology and Decision Making 2:459–476

Mitchell T (1997) Machine Learning. McGraw-Hill, New York, NY

Mobasher B (2004) Web usage mining and personalization. In: Singh MP (ed) Practical Handbook of Internet Computing, Chapman Hall & CRC Press, Baton Rouge

Perkowitz M, Etzioni O (2000) Towards adaptive web sites: Conceptual framework and case study. Artificial Intelligence 118(2000):245–275

Sarukkai RR (2000) Link prediction and path analysis using markov chains. Computer Networks 33(1-6):377–386

Schechter S, Krishnan M, Smith M (1998) Using path profiles to predict HTTP requests. Computer Networks and ISDN Systems 30:457–467

Siegel S, Castellan Jr N (1988) Nonparametric Statistics for the Behavioral Sciences, 2nd edn. McGraw-Hill, New York, NY

Spiliopoulou M, Mobasher B, Berendt B, Nakagawa M (2003) A framework for the evaluation of session reconstruction heuristics in web usage analysis. IN-FORMS Journal on Computing (15):171–190

Wilson E, Hilftery M (1931) The distribution of chi-square. Proceedings of the National Academy of Sciences of the United States of America 17:684–688

José Borges received his PhD in Computer Science in 2000 from University College of London and a Master degree in Electrical Engineering and Computer Science in 1994 from School of Engineering, University of Porto. He is currently an auxiliar professor at the School of Engineering in University of Porto. His main research interests are web data mining and web technologies and he has published several papers on the field of web mining.

Mark Levene received his PhD in Computer Science in 1990 from Birkbeck College University of London, having previously been awarded a BSc in Computer Science from Auckland University New Zealand in 1982. He is currently Professor of Computer Science at Birkbeck College, where he is a member of the Information Management and Web Technologies research group. His main research interests are web search and navigation, web data mining and stochastic models for the evolution of the web. He has published extensively in the areas of database theory and web technologies, and has recently published a book called An Introduction to Search and Engines and Web Navigation.