Computing the Entropy of User Navigation in the Web

Mark Levene and George Loizou
School of Computer Science and Information Systems
Birkbeck University of London
Malet Street, London WC1E 7HX, U.K.
Email: {mark,george}@dcs.bbk.ac.uk

Abstract

Navigation through the web, colloquially known as "surfing", is one of the main activities of users during web interaction. When users follow a navigation trail they often tend to get disoriented in terms of the goals of their original query and thus the discovery of typical user trails could be useful in providing navigation assistance. Herein we give a theoretical underpinning of user navigation in terms of the entropy of an underlying Markov chain modelling the web topology. We present a novel method for online incremental computation of the entropy and a large deviation result regarding the length of a trail to realise the said entropy. We provide an error analysis for our estimation of the entropy in terms of the divergence between the empirical and actual probabilities. We then indicate applications of our algorithm in the area of web data mining. Finally, we present an extension of our technique to higher-order Markov chains by a suitable reduction of a higher-order Markov chain model to a first-order one.

Key words. Web user navigation, Web data mining, navigation problem, Markov chain, entropy

1 Introduction

The World-Wide-Web (known as the web) has become a ubiquitous tool, used in day-to-day work, to find information and conduct business, and it is growing at an exponential rate. The activity of searching for information consists of the cycle: (i) submitting a query to a search engine, (ii) selecting a page for browsing from the returned list of pages, and (iii) navigating, i.e. link following. We concentrate on the *navigation* process where the user forms *trails* of pages through the database graph, starting from a page chosen from the search engine's result list. The central concept of a trail, which describes some logical association amongst its pages, is based on Bush's visionary idea [Bus45] (see [Ore91]). During the navigation process users often tend to "get lost in hyperspace" [Nie90] meaning that when following links users tend to become disoriented in terms of the goals of their original query and the relevance to the query of the information they are currently browsing; we refer to this problem as the *navigation problem* [LL02b].

Our previous work on hypertext involved the formalisation of an underlying model, motivated by the navigation problem just mentioned [LL99a, LL99b]. In this model a hypertext database consists of an information repository, which stores the contents of the database in the form of pages, and a directed graph describing the structure of the database; the nodes

of the directed graph are the pages in the repository and the arcs of the directed graph are the links. We call a set of trails in the database graph a *web view*. To accommodate for the stochastic nature of navigation we represent each page by a state and we associate probabilities with the links, resulting in a Markov chain model. The probabilities can have two separate interpretations: firstly denoting the frequencies that users traversed links, taking into account the time users spend at the destination page, and secondly the weighted relevance the user attaches to a link given a particular query. Using these interpretations we have devised data mining algorithms for constructing a web view of the user's navigation behaviour patterns [BL00] and have also devised adaptive algorithms for constructing a web view based on automating the navigation process according to the user's query [ZL99].

Following from the above work we present a model for web user navigation based on the entropy of a finite ergodic Markov chain [KS60]. In this model, each user navigation session can be viewed as a trail through an ergodic Markov chain such that as soon as one session is completed a new one can be started at any state with nonzero probability according to an empirical initial distribution. Over a period of time we assume that the empirical distribution of the Markov chain probabilities stabilises in accordance with the actual transition probabilities. The entropy of the Markov chain is central to this approach, since once the empirical distribution stabilises, the entropy of a *typical* trail is "close" to the entropy of the Markov chain as a consequence of the *Asymptotic Equipartition Property* (AEP) [CT91]. Such a typical trail can be seen to represent the user's navigation behaviour over a period of time. Herein we concentrate on the analysis of an iterative method for computing this entropy by considering a long navigation session, which can be viewed as the concatenation of shorter sessions each starting from a predetermined "home page".

The main results of the paper are twofold. Firstly, we obtain a large deviation result for the length of a trail through a Markov chain needed to compute its entropy in terms of the mean waiting and recurrence times between states. Secondly, we formalise an iterative online algorithm for computing the entropy of a Markov chain, which is shown to converge to the true entropy from below. Moreover, we provide an error analysis for this algorithm in terms of the relative entropy between the empirical and actual transition probabilities, which is shown to be a chi-squared statistic.

Being able to compute the entropy of a Markov chain representing the part of the web navigated by a user or a group of users allows us to compute the probability of a typical trail (see Section 2), enabling us to mine such trails employing usage mining techniques similar to those presented in [BL00]. Moreover, knowledge of the entropy of a typical trail and the stationary distribution of the underlying Markov chain can be used to personalise ranking algorithms, such as Google's PageRank [PBMW98], for individual users or groups of users.

Although the Markov chain assumption is somewhat controversial, it is justifiable as a first attempt to obtain analytic results to aid our understanding of the navigation problem. The recent empirical results of Pirolli and Pitkow [PP99], based on web log data that summarise user navigation sessions, support our initial choice of a (first-order) Markov chain model as opposed to a higher-order Markov chain model, for two reasons. Firstly, navigation sessions are typically short, i.e. they do not tend to exhibit long range dependencies. In this context Huberman et al. [HPPL98] have suggested a "universal law of surfing", backed-up by evidence from web log data, which predicts that typical trails are short. Secondly, the experimental results of Pirolli and Pitkow [PP99] suggest that a first-order Markov chain model is substan-

tially more stable over a period of time than higher-order Markov chain models and is thus more reliable. However, higher-order Markov chains can be reduced to first-order Markov chains by aggregating states [Bil61] and therefore the techniques we present herein extend to higher-order Markov chain models. In fact, we extend our basic technique to higher-order Markov chains of bounded order utilising *dynamic Markov modelling* [CH87, BM89, TR93], which is an adaptive context modelling method in which a finite-state model is built dynamically.

The rest of the paper is organised as follows. In Section 2 we give the necessary preliminaries regarding ergodic Markov chains. In Section 3 we obtain a large deviation result for the length of a trail through a Markov chain needed to compute its entropy. In Section 4 we present an algorithm for computing incrementally the entropy of a Markov chain from a trail induced by a random walk and analyse its convergence properties. In Section 5 we analyse the performance of this algorithm by deriving an error term for it in terms of the relative entropy between the empirical and actual transition probabilities. In Section 6 we discuss applications of our algorithm in the area of web data mining [KB00]. In Section 7 we present an extension of our technique to higher-order Markov chains. Finally, in Section 8 we give our concluding remarks.

2 Ergodic Markov chains

Herein we present results on Markov chains which are needed to develop the main contribution. We first show that viewing the home page as the starting point of all user navigation sessions leads to an ergodic Markov chain, which is central to our approach. We then formalise the notion of the entropy of such a Markov chain and explain the AEP. The rest of the section presents some technical results needed for the estimates we derive in later sections.

An *ergodic Markov chain*, $\mathcal{M} = (G, P)$, is a finite Markov chain which is aperiodic and irreducible [Fel68] (such a Markov chain is called *regular* in [KS60]); we will often refer to an ergodic Markov chain simply as a Markov chain. In particular, $G = (N, E)$ is its underlying finite directed graph; the nodes in N are called *states*, the cardinality of N is n, and the cardinality of E is called the *size* of G. (At times we refer to the set N as the state space of $\mathcal{M}$.) In addition, the probabilities P_{ij} associated with arcs (or links) $(s_i, s_j) \in E$ are called *transition probabilities*, and the probabilities P_i associated with states $s_i \in N$ are called *initial probabilities*. We note that, since the Markov chain is ergodic, for some $m > 0$ we have that for all $i, j, P_{ij}^m > 0$.

As an example, consider the following transition probability matrix of an irreducible Markov chain $\mathcal{M}_1$ with four states s_0 to s_3, where s_0 represents the user's home page, whose initial probability is one.

$\mathcal{M}_1$	s_0	s_1	s_2	s_3
s_0	0	0.3	0.5	0.2
s_1	0.4	0	0	0.6
s_2	0	0.9	0	0.1
s_3	0.5	0	0.5	0

The induced Markov chain summarises the navigation statistics of a user through s_1, s_2 and s_3. After eliminating the home page s_0, which we consider to be an artificial starting

point, we get the ergodic Markov chain $\mathcal{M}_2$ with the following transition matrix and the initial probability vector, $\langle 0.3, 0.5, 0.2\rangle$.

$\mathcal{M}_2$	s_1	s_2	s_3
s_1	0.12	0.2	0.68
s_2	0.9	0	0.1
s_3	0.15	0.75	0.1

In general, assume that we are given an irreducible Markov chain (which may or may not be aperiodic) having a distinguished starting state, s_0, whose initial probability is one, and such that all the probabilities P_{0i}, with $i > 0$, are positive (assume $P_{00} = 0$). We can then eliminate s_0 via the state reduction technique of Sonin [Son99] to obtain an ergodic Markov chain modelling the user's behaviour.

A *trail* in $\mathcal{M}$ is a sequence

$$T = s_1, s_2, \ldots, s_t$$

of states in N such that $(s_i, s_{i+1}) \in E, i \in \{1, \ldots, t-1\}$, where t is the length of T.

The probability $p(T)$ of a trail T of length t in $\mathcal{M}$ is given by

$$p(T) = P_{k_1} P_{k_1 k_2} P_{k_2 k_3} \cdots P_{k_{t-1} k_t}, \tag{1}$$

where $k_1, k_2, \ldots, k_t$ is a permutation of $1, 2, \ldots, k$.

It is well known that an ergodic Markov chain has a *stationary distribution*, π, satisfying

$$\pi P = \pi,$$

such that

$$\lim_{m \to \infty} P^m = Q,$$

where each row of Q is identical to the stationary distribution π, which is positive.

It can be verified that for the ergodic Markov chain $\mathcal{M}_2$ above, we have

$$\pi = \langle 0.37246, 0.311512, 0.316027\rangle.$$

The *entropy* of a Markov chain [Khi57, CT91] is given by

$$H(\mathcal{M}) = -\sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i P_{ij} \log P_{ij}, \tag{2}$$

where as usual logarithms are taken to the base 2 and by convention $\log 0 = 0$ and $\log 0/0 = 0$.

It can be verified that for the ergodic Markov chain $\mathcal{M}_2$ above, we have

$$H(\mathcal{M}_2) = 0.929797.$$

The following result dates back to Shannon [Sha48] (see [Khi57] for a lucid proof).

Theorem 2.1

$$H(\mathcal{M}) = \lim_{t \to \infty} \frac{-\log p(T)}{t} \quad \text{(almost surely)}, \tag{3}$$

where T is a trail in $\mathcal{M}$ of length t. □

We will call a trail T *typical* if its entropy $-\log p(T)/t$ is "close" to $H(\mathcal{M})$. The AEP, already mentioned in Section 1, asserts that all typical trails have the same probability, given by

$$p(T) \approx 2^{-tH(\mathcal{M})}. \tag{4}$$

By the AEP the sum of the probabilities of *all* the typical trails can be made as "close" to one as we like by increasing t.

The next proposition, where R_{ij} is a random variable giving the recurrence time of the transition from state s_i to state s_j in one step, follows from Theorem 2 in [Kac47] (see [WZW98]).

Proposition 2.2 The mean recurrence time for R_{ij}, denoted by $E(R_{ij})$, is given by

$$E(R_{ij}) = \frac{1}{\pi_i P_{ij}}. \quad \square \tag{5}$$

As in Section 4.3 of Kemeny and Snell [KS60] (cf. [LL02a]), let

$$Z = (z_{ij}) = \Big(I - (P - E\ Diag(\pi))\Big)^{-1}$$

be the fundamental matrix for the ergodic Markov chain with transition matrix P, where E is the matrix whose elements are all one and $Diag(\pi)$ is a diagonal matrix whose diagonal elements are the components of the stationary distribution π. (Note that $Q = EDiag(\pi)$.)

The next proposition follows from Proposition 2.2 and the results of Section 4.4 in [KS60]. Let W_{ij} be a random variable giving an upper bound on the waiting time for the transition from state s_i to state s_j to first occur in one step; we define the expectation of W_{ij} to be equal to the maximum mean time it takes to get to state s_j plus the mean recurrence time of the transition from state s_i to state s_j in one step.

Proposition 2.3 Let $E(W_{ij})$ denote the mean waiting time for W_{ij}. Then,

$$E(W_{ij}) = \frac{z_{jj} - \min_k z_{kj}}{\pi_j} + \frac{1}{\pi_i P_{ij}}, \tag{6}$$

where

$$z_{kj} = \delta_{kj} + \sum_{m=1}^{\infty} (P_{kj}^m - \pi_j),$$

with δ_{kj} being the Kronecker delta and $j, k \in \{1, \ldots, n\}$. $\square$

Let β be the minimum natural number such that for all $i, j, P_{ij}^{\beta} > 0$ (see Section 8.5 in [HJ85] for various upper bounds on β, which is known as the *index of primitivity*). It follows that

$$|\, z_{ij} \,| \leq \beta + \sum_{m=0}^{\infty} |\, P_{ij}^{\beta+m} - \pi_j \,|\,. \tag{7}$$

Now, let $1/\eta$ be given by

$$\frac{1}{\eta} = \min_{i,j} P_{ij}^{\beta} \leq \frac{1}{n}.$$

The next lemma gives us a rough idea of how large z_{ij} can be, independently of i and j.

Lemma 2.4 An upper bound on $|z_{ij}|$ is given by

$$|z_{ij}| \leq \beta + \eta.$$

Proof. The result follows immediately, since by a standard coupling argument (see Fact 10 in [Ros95]) we can deduce from (7) that

$$|z_{ij}| \leq \beta + \sum_{m=0}^{\infty} \left(1 - \frac{1}{\eta}\right)^m. \quad \square$$

We note that

$$\frac{z_{jj} - z_{ij}}{\pi_j},$$

is the mean waiting time for the first occurrence of s_j starting from s_i. Moreover, the interpretation of $z_{jj} - z_{ij}$ is the difference (in the limit) between the mean number of visits to state s_j starting from s_j and the mean number of visits to state s_j starting from s_i. As a final remark we note that by Lemma 2.4

$$|z_{jj} - z_{ij}| = |1 - \delta_{ij} + \sum_{m=1}^{\infty} (P_{jj}^m - P_{ij}^m)| \leq 2(\beta + \eta). \tag{8}$$

Define a *tour* of G to be a trail in $\mathcal{M}$ which contains all the states in N and such that the first and last states of the trail are the same. The following result is easily obtainable, where the mean waiting time for a tour is the mean waiting time to get to the first state of the tour, say s_i, plus the mean waiting time to get to each consecutive state in the tour and then back to s_i.

Proposition 2.5 The minimum mean waiting time for a tour of G, denoted by $W(G)$, is given by

$$\begin{aligned} W(G) &= \min_{i_1, i_2, \ldots, i_n} \left(\frac{z_{i_1 i_1} - \min_k z_{k i_1}}{\pi_{i_1}} + \sum_{j=2}^{n} \frac{z_{i_j i_j} - z_{i_{(j-1)} i_j}}{\pi_{i_j}} + \frac{z_{i_1 i_1} - z_{i_n i_1}}{\pi_{i_1}} \right) \\ &\leq (n+1) 2\eta(\beta + \eta), \end{aligned} \tag{9}$$

where $i_1, i_2, \ldots, i_n$ is a permutation of $1, 2, \ldots, n$. $\square$

3 A large deviation result for Markov chains

By Theorem 2.1 we can compute the entropy of a Markov chain by computing the probability of a "long" random walk (trail) through the navigation space. In order to get an estimate on the length of such a trail we derive an inequality which yields the probability of a large deviation of the empirical probability from the corresponding stationary probability. We use the method of bounded differences directly [McD89] (for Hoeffding's seminal paper see [Hoe63]) rather than relying on the eigenvalue gap as in [Gil98], where it is additionally assumed that the Markov chain is reversible.

Let T be a trail of length t and $\mathcal{F}_k$ denote the prefix of length k of T, with $k \in \{0, 1, \ldots, t\}$; by convention $\mathcal{F}_0$ is the empty sequence. Moreover, let $m_{i,j}(k)$ (or simply $m_{i,j}$ whenever $k = t$) be the number of transitions from s_i to s_j in $\mathcal{F}_k$.

We observe that, by the law of large numbers for ergodic Markov chains, $m_{i,j}/t$ converges to $\pi_i P_{ij}$ as t tends to infinity. Moreover, by Khinchin's proof of Theorem 2.1 [Khi57] once $m_{i,j}/t$ is "close" to $\pi_i P_{ij}$, for all $i, j \in \{1, \ldots, n\}$, the trail T becomes *typical*. By collecting the trail statistics, i.e. $m_{i,j}$ and m_i, where m_i is the number of visits to state s_i in T, we can compute $H(\mathcal{M})$ on using (2), noting again that by the law of large numbers $m_{i,j}/m_i$ converges to P_{ij} as t tends to infinity.

On using Hoeffding's inequality and the method of bounded differences, we now obtain an upper bound on the length t of a trail T; the length is needed to compute the entropy given by (3). Define a martingale $Y_0, Y_1, \ldots, Y_t$, with

$$Y_k = E(m_{i,j}(k) \mid \mathcal{F}_k),$$

where $k \in \{0, 1, \ldots, t\}$, viewing T as a sequence of random variables generated by a simulation of $\mathcal{M}$ (cf. Example 12.2.20 in [GS92]). Theorem 6.7 in [McD89] then implies that

$$p\left(\left|\frac{m_{i,j}}{t} - \pi_i P_{ij}\right| \geq \epsilon\right) \leq 2\exp\left(\frac{-\epsilon^2}{2tC^2}\right), \tag{10}$$

where $\epsilon > 0$ and for all $k \in \{1, \ldots, t\}$, $|Y_k - Y_{k-1}| \leq C$.

We now derive a value for C, which is one plus the number of transitions that can be packed into a sequence of states whose length is the mean waiting time for W_{ij}. Define

$$m_{i,j}(k, t) = m_{i,j}(t) - m_{i,j}(k),$$

where $0 \leq k \leq t$, i.e. the number of transitions from s_i to s_j in the suffix of T starting from the kth state. Thus, where $1 \leq k \leq t$, we have

$$\begin{aligned} |Y_k - Y_{k-1}| &= |[m_{i,j}(k) + E(m_{i,j}(k,t) \mid \mathcal{F}_k)] - [m_{i,j}(k-1) + E(m_{i,j}(k-1,t) \mid \mathcal{F}_{k-1})]| \\ &\leq 1 + \frac{E(W_{ij})}{E(R_{ij})}, \end{aligned}$$

since we can bound $E(m_{i,j}(k,t) \mid \mathcal{F}_k)$ above and below by

$$\frac{(t-k) - E(W_{ij})}{E(R_{ij})} \leq E(m_{i,j}(k,t) \mid \mathcal{F}_k) \leq \frac{(t-k)}{E(R_{ij})}.$$

Thus on using (8) and Propositions 2.2 and 2.3 we can rewrite (10) as

$$p\left(\left|\frac{m_{i,j}}{t} - \pi_i P_{ij}\right| \geq \epsilon\right) \leq 2\exp\left(\frac{-\epsilon^2}{2t(2 + 2P_{ij}\eta(\beta+\eta))^2}\right), \tag{11}$$

since

$$\frac{E(W_{ij})}{E(R_{ij})} = 1 + \frac{\pi_i P_{ij}(z_{jj} - \min_k z_{kj})}{\pi_j} \leq 1 + 2P_{ij}\eta(\beta+\eta).$$

Finally, in order to obtain an estimate of the trail length for computing $H(\mathcal{M})$, we multiply our estimate of t by the size of G. Given ϵ and p, our estimate of t, for a given pair (i, j), can be derived from (11).

4 Computing the entropy of a Markov chain

A straightforward method for computing the entropy of a Markov chain, when P is given, is to compute the stationary distribution π and then to use (2) to compute $H(\mathcal{M})$.

In the case when we do not have P readily available we cannot use this straightforward method. Moreover, the size of G and the complexity of determining π may prevent us from such a computation of the entropy. An incremental method for computing the entropy arises from (3) simply by navigating through the state space, where the transitions occur according to the underlying probabilities, and then collecting the trail statistics. We will show that with such a method the empirical entropy converges to the true entropy from below, i.e. the empirical entropy increases monotonically until it converges to the true entropy.

We assume that a function $random(\{P_i\})$ is available which returns an initial state distributed according to the initial probabilities $\{P_i\}$, and a function $random(s_i, \{P_{ij}\})$ is available which returns a state adjacent to s_i distributed according to the transition probabilities $\{P_{ij}\}$, where $i, j \in \{1, \ldots, n\}$. We now give the pseudo-code of an algorithm, designated ENTROPY$(\{m_i\}, \{m_{i,j}\})$, which returns an estimate of the entropy of the Markov chain $\mathcal{M}$; for brevity in the algorithm we overload the index k of a state s_k so that it refers both to the kth state in the trail induced by the random walk and also to the kth state in N. In the algorithm $\{m_i\}$ and $\{m_{i,j}\}$ are variables indicating, respectively, the number of visits to s_i and the number of transitions from s_i to s_j so far, where $i, j \in \{1, \ldots, n\}$. Moreover, CONVERGED is a Boolean variable which is assumed to be initially false.

Algorithm 1 (ENTROPY$(\{m_i\}, \{m_{i,j}\})$)

1. **begin**
2. $H, \{m_i\}, \{m_{i,j}\}, k := 0$;
3. $s_1 := random(\{P_i\})$;
4. **while not** CONVERGED **do**
5. $k := k + 1$;
6. $s_{k+1} := random(s_k, \{P_{kj}\})$;
7. $H := H + (m_{k,(k+1)} + 1) \log((m_{k,(k+1)} + 1)/(m_k + 1)) - m_{k,(k+1)} \log(m_{k,(k+1)}/m_k)$;
8. **for all** $j \neq (k + 1)$ **do**
9. $H := H + m_{k,j} \log(m_k/(m_k + 1))$;
10. **end for**
11. $m_k := m_k + 1$;
12. $m_{k,(k+1)} := m_{k,(k+1)} + 1$;
13. **end while**
14. **return** $-H/k$;
15. **end.**

Regarding the convergence criterion, it is possible to use (11) in order to obtain an upper bound on the *number of iterations*, say t, of the while loop beginning at line 4 and ending at line 13, so that the *empirical entropy* $-H/t$ returned by the algorithm is as "close" to $H(\mathcal{M})$ as we wish in the sense of almost sure convergence; in this case, we say that t is *large enough* and write $-H/t \approx H(\mathcal{M})$.

The next theorem, which states the correctness of Algorithm 1, follows from the fact that

$$-H/t = -\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{m_{i,j}}{t}\log\frac{m_{i,j}}{m_i}. \tag{12}$$

Theorem 4.1 Assume that t is large enough. Then the value $-H/t$ returned by Algorithm 1 correctly approximates $H(\mathcal{M})$, i.e. $-H/t \approx H(\mathcal{M})$. $\square$

We observe that the space complexity of Algorithm 1 is linear space in the size of G regardless of the length t of a typical trail. On the other hand, the time complexity of Algorithm 1 is $O(t\ \log|G|)$, where $|G|$ denotes the size of G, assuming that a binary search can be performed on G to find a given state.

In the following let H_k denote the value of H in Algorithm 1 after $k \geq 0$ executions of the while loop beginning at line 4 and ending at line 13. The next theorem implies that $-H/t$ converges to $H(\mathcal{M})$ from below, where t is the final value of k.

Theorem 4.2 $H_{k+1} \leq H_k$.

Proof. Suppose that at line 6 of Algorithm 1 a transition occurred from state s_i to state s_j and that the number of states outgoing from s_i is m, with $m \geq 1$; $m = 0$ is not possible since $\mathcal{M}$ is ergodic. If $m = 1$, then the result is immediate since $H_{k+1} = H_k$ due to the fact that $m_{i,j} = m_i$. So assume that $m > 1$ and that the transition that occurred at line 6 of Algorithm 1 was from s_i to s_m.

Let X_j denote $m_{i,j}$, for $j \in \{1,\ldots,m\}$, and X denote m_i. It is sufficient to show that

$$\sum_{j=1}^{m-1} X_j\log\frac{X_j}{1+X} + (1+X_m)\log\frac{1+X_m}{1+X} \quad < \quad \sum_{j=1}^{m-1} X_j\log\frac{X_j}{X} + X_m\log\frac{X_m}{X}. \tag{13}$$

(Note that we demonstrate strict inequality and that we omit to divide the left and right hand sides of (13), respectively, by $k+1$ and k.) On substituting

$$X - \sum_{j=1}^{m-1} X_j \text{ for } X_m$$

into (13), it remains to show, after some algebraic manipulation, that

$$(1+X-\sum_{j=1}^{m-1}X_j)\log(1+X-\sum_{j=1}^{m-1}X_j) - (1+X)\log(1+X) \quad < \\ (X-\sum_{j=1}^{m-1}X_j)\log(X-\sum_{j=1}^{m-1}X_j) - X\log X. \tag{14}$$

On a further substitution of

$$X - \sum_{j=1}^{m-1} X_j \text{ for } \alpha X$$

into (14) for some α, with $0 < \alpha < 1$, we need to establish that

$$(1+\alpha X)\log(1+\alpha X) - (1+X)\log(1+X) \quad < \quad \alpha X \log \alpha X - X \log X. \tag{15}$$

The result now follows since (15) can be transformed into

$$\alpha X \log\left(1+\frac{1}{\alpha X}\right) + \log(1+\alpha X) \quad < \quad X \log\left(1+\frac{1}{X}\right) + \log(1+X) \tag{16}$$

and $x\log(1+1/x)$ is strictly monotonically increasing for all $x > 0$, since its derivative is positive for all $x > 0$. $\square$

We observe that the difference

$$(1+X)\log(1+X) - X\log X$$

pertaining to the decrease of $m_{i,j}/m_i$ due to the increase in m_i, where $j \neq m$, is greater than the difference

$$(1+\alpha X)\log(1+\alpha X) - \alpha X \log \alpha X$$

pertaining to the increase of $m_{i,m}/m_i$ due to the increase in both m_i and $m_{i,m}$. All other empirical probabilities, $m_{h,j}/m_h$, with $h \neq i$, remain unchanged due to the increase in m_i. We further note that all empirical probabilities, m_h/k, with $h \neq i$, decrease due to the increase in k, apart from m_i/k which increases; this fact does not influence Algorithm 1, since H_k is divided by k only before exiting from the algorithm.

Building on Theorem 4.2 we can detect the *per-state convergence* of the empirical entropy to the true entropy by examining the rate of change indicated by (15). It is important to note that when X, i.e. m_i, is large enough then α is approximately equal to P_{im}, written $\alpha \approx P_{im}$. Let us now assume that we fix $\alpha > 0$ at a level which is a lower bound estimate of P_{im}. On using the fact that

$$\lim_{X\to\infty} \log\left(1+\frac{1}{\alpha X}\right)^{\alpha X} = \log e$$

we can approximate (16) by

$$0 < \log\left(\frac{1+X}{1+\alpha X}\right) \approx \log\left(\frac{1}{\alpha}\right)$$

when X is sufficiently large, noting that $(1+X)/(1+\alpha X)$ converges from below to $1/\alpha$.

So, on considering an error $\epsilon > 0$ and letting

$$\log\left(\frac{1+X}{1+\alpha X}\right) = \log\left(\frac{1}{\alpha}\right) - \epsilon$$

we can estimate X to obtain

$$X = \frac{1-\alpha 2^{\epsilon}}{\alpha(2^{\epsilon}-1)}, \tag{17}$$

with the constraint that $\log(1/\alpha) > \epsilon$. As an example, if we let $\epsilon = 1$ then our estimate of X is given by

$$X = \frac{1}{\alpha} - 2.$$

We emphasise that the estimate of X is per-state, so in order to obtain an estimate of t, i.e. an estimate on the number of times the while loop in Algorithm 1 should be executed, we need to multiply our estimate of X by the expected number of steps of a tour of all the states in the Markov chain. A lower bound for this expectation is he number of states of the Markov chain, i.e. n, and an upper bound is given by (9).

5 Error analysis

We next proceed to analyse the performance of computing the entropy given the empirical transition probabilities; to this end we derive the distance of the empirical entropy from the true entropy of the subtrail generated so far by the random walk. This distance can be viewed as the error.

Let $p(T)$ be as in (1). Then the entropy of T, denoted by $H(T)$, is given by

$$H(T) = \frac{-\log p(T)}{t} = \frac{-\log P_{k_1}}{t} - \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{m_{i,j}}{t} \log P_{ij}.$$

Let the *error*, denoted by $\mathcal{E}$, be the difference between the entropy of T and the empirical entropy given by (12). It can be verified that

$$\mathcal{E} = \frac{-\log P_{k_1}}{t} + \frac{1}{t} \sum_{i=1}^{n}\sum_{j=1}^{n} m_{i,j} \log \frac{m_{i,j}}{m_i P_{ij}}, \tag{18}$$

where the second term in the error is the *relative entropy* (or *divergence*) between the empirical transition probabilities $\{m_{i,j}/m_i\}$ and the actual transition probabilities $\{P_{ij}\}$; thus $\mathcal{E} \geq 0$, since the relative entropy is non-negative [CT91].

By Corollary 5.2 in [McD89] (see [Hoe63])

$$p\left(\left|\frac{m_{i,j}}{m_i} - P_{ij}\right| \geq \delta\right) \leq 2\exp\left(-2\delta^2 m_i\right),$$

with $0 < \delta < 1$. In particular, when

$$m_i = \left\lceil \frac{\ln 2k}{2\delta^2} \right\rceil,$$

where ln stands for the natural logarithm, the probability of a deviation greater than or equal to δ is less than or equal to $1/k$. So, when k is sufficiently large and δ is sufficiently close to zero, (18) can be rewritten as

$$\mathcal{E} \approx O\left(\frac{1}{t}\right) + \frac{1}{t} \sum_{i=1}^{n}\sum_{j=1}^{n} m_{i,j} \log \left(\frac{m_{i,j}/m_i}{(m_{i,j}/m_i) \pm O\left(\frac{1}{\sqrt{m_i}}\right)} \right), \tag{19}$$

assuming that for all $i, j \in \{1, \ldots, n\}$, $m_{i,j}/m_i > \delta$, since

$$\delta \approx \sqrt{\frac{\ln 2k}{2m_i}} = O\left(\frac{1}{\sqrt{m_i}}\right).$$

We observe that t is, in general, much larger than $\sqrt{m_i}$, since $t = \sum_{i=1}^{n} m_i$, and thus, in general, the first error term vanishes much faster than the second error term.

As a closing remark we note that in the absence of any additional prior knowledge about the actual transition probabilities, the empirical entropy provides the maximum likelihood estimate of the true entropy. Moreover, the relative entropy is asymptotically chi-squared

with $n^2 - n - q$ degrees of freedom, where q is the number of P_{ij} with $P_{ij} = 0$ [Bil61] (cf. [Mil55]); that is,

$$2\sum_{i=1}^{n}\sum_{j=1}^{n} m_{i,j} \log \frac{m_{i,j}}{m_i P_{ij}} \approx \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{(m_{i,j} - m_i P_{ij})^2}{m_i P_{ij}}.$$

Thus the expected value of the error, i.e. the bias of our estimate of $H(T)$, is given by

$$E(\mathcal{E}) = \frac{-2 \log P_{k_1} + n^2 - n - q}{2t}.$$

6 Applications in web data mining

Herein we mention applications of our results in two subareas of *web data mining* [KB00]. The first subarea is that of *web usage mining* [CPY98, SKS98, BL00], which makes use of web log data to discover behavioural patterns in one or more users navigation history. In this context we also mention [HPPL98] wherein, backed-up by evidence from web log data, it was shown that navigation behaviour gives rise to regular statistical patterns and in particular to a power-law distribution. Further evidence to this effect was demonstrated in [LBL01], wherein a power-law distribuion was derived from a Markov chain model similar to the one assumed in this paper.

The online incremental algorithm that we have presented can be used to compute the Markov chain transition probabilities and the probability of a typical navigation trail, which is related to the entropy via (4). This incremental approach to computing the entropy is consonant with the way users surf the web and the manner in which web log data is collected. One possibility is to incorporate our results into existing web usage algorithms such as the one described in [BL00], so that typical trails can be mined.

The second subarea is that of *web structure mining* [CGP98, PBMW98, CDK$^+$99, HHMN99], which involves the discovery of linkage patterns in the structure of the web graph that can be used to improve the ranking of web pages. This approach originates from the area of citation analysis [PN76, Gel78], whose aim is to measure the influence of research in a given subfield using citation data; see also [Lar96], where co-citation analysis is used to cluster related web pages.

The stationary distribution π of the Markov chain computed from the user navigation log data indicates the relative weighting of the pages according to the user's preferance. (We note that, once Algorithm 1 has converged, π can be returned as a byproduct.) This is a form of page ranking, which could be weighted into search engine results to improve the relevance of answers to the user or a group of users, and can be viewed as an extension of Google's PageRank [PBMW98]. Currently search engine results do not take into account the preferences of an individual user or a group of users, so it is worth incorporating data mining results into the page ranking component of a search engine.

7 Extension to higher-order Markov chains

Although we have argued in the introduction that a (first-order) Markov chain model is a good starting point for developing a theory of user navigation patterns, it is only an approximation and therefore a non-parametric model for entropy estimation along the lines proposed in

[KASW98, WZW98] should be considered. One of the problems, however, in using the Wyner-Ziv approach in the context of the web is that navigation sessions are typically short and therefore it is probably better to fit a ν-order Markov chain model, for some $\nu \geq 1$, to the data. This can be done by using the techniques described in [Cha73, MGZ89], or by using an adaptive context modelling technique [Ris86, BWC89]. In our case fitting a ν-order Markov chain model is not practical due to the size of the model, i.e. $O(n^{\nu})$, since, in general, we expect only few transitions to have non-zero probability. So, we propose instead a technique based on *dynamic Markov modelling* [CH87, BM89, TR93], which is an adaptive context modelling technique in which a finite-state model is built dynamically.

Hereafter we describe an extension to Algorithm 1 realising the said technique; we refer to this extension as Algorithm 1e, and indicate how the results of the previous sections are affected.

We first set a limit, say $V \geq 2$, on the order ν of the Markov chain model, which is based on the expected length of a navigation session starting and ending at the user's home page. We then invoke Algorithm 1 $V-1$ times, for $\nu = 1, 2, \ldots, V-1$, where, additionally, at each step we maintain second-order probabilities relative to the current ν-order Markov chain model, and at the end of each step we add states to the current model if the second-order probabilities are sufficiently different from the current first-order probabilities, thereby extending the order of the current Markov chain model.

We now outline Algorithm 1e. As before we represent each page by a state but we extend the mapping from states to pages to be a many-to-one mapping rather than a one-to-one mapping, i.e. a page may now be associated with several states; let us denote this mapping by ρ.

We utilise the operation of *cloning* [CH87], whereby a state, say s_k, is duplicated on the basis of a link (s_i, s_k), subject to the condition that there is another link (s_h, s_k), with s_h being a distinct state from s_i. Specifically, we create a new state s_u, with $\rho(s_u) = \rho(s_k)$, and modify the link structure of the underlying graph of the Markov chain model as follows:

1) remove the link (s_i, s_k) and add the link (s_i, s_u), and

2) for every link (s_k, s_j) we add the link (s_u, s_j).

It is evident that cloning preserves the deterministic nature of the finite-state model induced by the underlying graph of the Markov chain model.

We now introduce new notation to capture second-order probabilities. We denote the probability of a transition from s_k to s_j (i.e. associated with the link (s_k, s_j)), *given* that the previous transition that occurred was from s_i to s_k (i.e. associated with the link (s_i, s_k)), by $P_{i,kj}$. As before the probability of a transition from s_k to s_j is denoted by P_{kj}.

Assume that we are currently at the νth iteration of Algorithm 1e. Then in addition to the P_{kj} statistics we also temporarily store the $P_{i,kj}$ statistics, whenever $P_{i,kj} > 0$. That is,

$$P_{i,kj} = \frac{m_{i,k,j}}{m_{k,j}},$$

where $m_{i,k,j}$ is the transition count indicating the number of transitions from s_k to s_j given that the previous transition that occurred was from s_i to s_k. The criterion for the convergence of the current iteration of Algorithm 1e is the accuracy of the estimations, $\{m_{i,k,j}/m_{k,j}\}$, of

the current second-order probabilities, i.e. convergence occurs when they are close enough to the true probabilities $P_{i,kj}$. (This can be measured by using Hoeffding's inequality as in Section 5.)

At the end of the νth iteration of Algorithm 1e we clone all states that satisfy the following condition with respect to a transition going into them. A state s_k is *cloned on the basis of a transition* (s_i, s_k) if

$$P_{i,kj} - \frac{1}{L_k} > \gamma_k,$$

for some state s_j and for some $0 < \gamma_k < 1$, where L_k is the number of states going into state s_k, i.e. the number of links of the form (s_h, s_k) for some state s_h. That is, s_k is cloned if there is sufficient evidence that the transition from s_k to s_j is *not* independent of the transition from s_i to s_k. (We observe that if there are m states going into state s_k, then s_k is cloned at most $m - 1$ times.)

After cloning a state the transition counts are modified as follows: where s_k is the state cloned on the basis of s_i, the newly added state is s_u, and old_m_k and $old\_m_{k,j}$ denote the values of m_k and $m_{k,j}$ prior to the cloning operation taking place, namely

1) $m_k = \sum_h m_{h,k}$, where the sum is over links (s_h, s_k) (note that after cloning $h \neq i$),

2) $m_u = m_{i,k}$,

3) $m_{k,j} = (m_k / old\_m_k)\ old\_m_{k,j}$, for every link (s_k, s_j), and

4) $m_{u,j} = (m_u / old\_m_k)\ old\_m_{k,j}$, for every link (s_u, s_j).

We close this section with the following remarks:

(i) The entropy decreases after cloning takes place at the end of each iteration of Algorithm 1e, since we are moving to a higher-order Markov chain model [CT91]. Thus during an iteration the empirical entropy increases according to Theorem 4.2 but it is strictly less than the value returned by the previous iteration of Algorithm 1e due to the cloning. When Algorithm 1e terminates the empirical entropy has converged to its V-order true entropy within the accuracy we have specified via the γ_k parameters.

(ii) The parameters γ_k can be used to control the level of cloning, i.e. the lower γ_k is for a given state s_k the more cloning takes place. Overall if the parameters γ_k are high then we expect the final model to be closer to the first-order model, and correspondingly if the parameters γ_k are low then we expect the final model to be closer to the V-order Markov chain model.

(iii) The results from previous sections are still valid with reference to Algorithm 1e, since each iteration of the algorithm is not affected by the $\{P_{i,kj}\}$ statistics, which are additional parameters used only for cloning purposes and then discarded prior to the next iteration.

8 Concluding Remarks

We have presented a Markov chain model for analysing user navigation patterns through the web, which is based on the computation of the information contained in a typical navigation trail. We have constructed an online algorithm that computes the empirical entropy of such a Markov chain by simulating a user's random walk through the web; this algorithm converges from below to the true entropy (Theorem 4.2). We have also provided mechanisms for estimating the required length of the said navigation trail so that the empirical entropy returned by our algorithm is sufficiently close to the true entropy. In addition, we have indicated how our algorithm can be incorporated into web data mining algorithms to improve the quality of their output. Finally, we have presented an extension of our basic technique that deals with higher-order Markov chains of bounded order. Our theoretical investigation paves the way for experimentation with web log data, based on a sound statistical methodology.

Acknowledgements. The authors would like to thank the referees for their constructive comments, which improved the presentation of the results.

References

[Bil61] P. Billingsley. Statistical methods in Markov chains. *The Annals of Mathematical Statistics*, 32:12–40, 1961.

[BL00] J. Borges and M. Levene. Data mining of user navigation patterns. In B. Masand and M. Spiliopoulou, editors, *Web Usage Analysis and User Profiling*, Lecture Notes in Artificial Intelligence (LNAI 1836), pages 92–111. Springer-Verlag, Berlin, 2000.

[BM89] T. Bell and A. Moffat. A note on the DMC data compression scheme. *The Computer Journal*, 32:16–20, 1989.

[Bus45] V. Bush. As we may think. *Atlantic Monthly*, 176:101–108, 1945.

[BWC89] T. Bell, I.H. Witten, and J.G. Cleary. Modeling for text compression. *ACM Computing Surveys*, 21:557–591, 1989.

[CDK+99] S. Chakrabarti, B. Dom, S.R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J.M. Kleinberg. Mining the web's link structure. *IEEE Computer*, 32:60–67, 1999.

[CGP98] J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through URL ordering. In *Proceedings of International World Wide Web Conference*, pages 161–172, Brisbane, 1998.

[CH87] G.V. Cormack and R.N.S. Horspool. Data compression using dynamic Markov modelling. *The Computer Journal*, 30:541–550, 1987.

[Cha73] C. Chatfield. Statistical inference regarding Markov chain models. *Applied Statistics*, 22:7–20, 1973.

[CPY98] M.-S. Chen, J.S. Park, and P.S. Yu. Efficient data mining for traversal patterns. *IEEE Transactions on Knowledge and Data Engineering*, 10:209–221, 1998.

[CT91] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley & Sons, Chichester, 1991.

[Fel68] W. Feller. *An Introduction to Probability Theory and its Applications*. John Wiley & Sons, New York, NY, 3rd edition, 1968.

[Gel78] N.L. Geller. On the citation influence methodology of Pinski and Narin. *Information Processing & Management*, 14:93–95, 1978.

[Gil98] D. Gillman. A Chernoff bound for random walks on expander graphs. *SIAM Journal on Computing*, 27:1203–1220, 1998.

[GS92] G.R. Grimmett and D.R. Stirzaker. *Probability and Random Processes*. Oxford University Press, Oxford, 2nd edition, 1992.

[HHMN99] M.R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. Measuring index quality using random walks on the web. In *Proceedings of International World Wide Web Conference*, pages 1291–1303, Montreal, 1999.

[HJ85] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, U.K., 1985.

[Hoe63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *American Statistical Association Journal*, 58:13–30, 1963.

[HPPL98] B.A. Huberman, P.L.T. Pirolli, J.E. Pitkow, and R.M. Lukose. Strong regularities in world wide web surfing. *Science*, 280:95–97, 1998.

[Kac47] M. Kac. On the notion of recurrence in discrete stochastic processes. *Bulletin of the American Mathematical Society*, 53:1002–1010, 1947.

[KASW98] I. Kontoyiannis, P.H. Algoet, Yu.M. Suhov, and A.J. Wyner. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *IEEE Transactions on Information Theory*, 44:1319–1327, 1998.

[KB00] R. Kosala and H. Blockeel. Web mining research: A survey. *SIGKDD Explorations*, 2:1–15, 2000.

[Khi57] A.I. Khinchin. *Mathematical Foundations of Information Theory*. Dover, New York, NY, 1957. Translated by R.A. Silverman and M.D. Friedman.

[KS60] J.G. Kemeny and J.L. Snell. *Finite Markov Chains*. D. Van Nostrand, Princeton, NJ, 1960.

[Lar96] R.R. Larson. Bibliometrics of the world wide web: An exploratory analysis of the intellectual structure of cyberspace. In *Proceedings of Annual American Society for Information Science Meeting*, pages 71–78, Baltimore, Md., 1996.

[LBL01] M. Levene, J. Borges, and G. Loizou. Zipf's law for web surfers. *Knowledge and Information Systems*, 3:120–129, 2001.

[LL99a] M. Levene and G. Loizou. Navigation in hypertext is easy only sometimes. *SIAM Journal on Computing*, 29:728–760, 1999.

[LL99b] M. Levene and G. Loizou. A probabilistic approach to navigation in hypertext. *Information Sciences*, 114:165–186, 1999.

[LL02a] M. Levene and G. Loizou. Kemeny's constant and the random surfer. *American Mathematical Monthly*, 109:741–745, 2002.

[LL02b] M. Levene and G. Loizou. Web interaction and the navigation problem in hypertext. In A. Kent, J.G. Williams, and C.M. Hall, editors, *Encyclopedia of Microcomputers*, pages 381–398. Marcel Dekker, New York, NY, 2002.

[McD89] C. McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics*, volume 141 of *London Mathematical Society Lecture Note Series*, pages 148–188. Cambridge University Press, Cambridge, U.K., 1989.

[MGZ89] N. Merhav, M. Gutman, and J. Ziv. On the estimation of the order of a Markov chain and universal data compression. *IEEE Transactions on Information Theory*, 35:1014–1019, 1989.

[Mil55] G.A. Miller. Note on the bias of information estimates. In H. Quastler, editor, *Information Theory in Psychology: Problems and Methods*, pages 95–100. Free Press, Glencoe, Il., 1955.

[Nie90] J. Nielsen. *Hypertext and Hypermedia*. Academic Press, Boston, Ma., 1990.

[Ore91] T. Oren. Memex: Getting back on the trail. In J.M. Nyce and P. Kahn, editors, *From Memex to Hypertext: Vannevar Bush and the Mind's Machine*, pages 319–338. Academic Press, San Diego, Ca., 1991.

[PBMW98] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Working paper, Department of Computer Science, Stanford University, 1998.

[PN76] G. Pinski and F. Narin. Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management*, 12:297–312, 1976.

[PP99] P. Pirolli and J.E. Pitkow. Distributions of surfers' paths through the world wide web: Empirical characterizations. *World Wide Web*, 2:29–45, 1999.

[Ris86] J. Rissanen. Complexity of strings in the class of Markov sources. *IEEE Transactions on Information Theory*, 32:526–532, 1986.

[Ros95] J.S. Rosenthal. Convergence rates of Markov chains. *SIAM Review*, 37:387–405, 1995.

[Sha48] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[SKS98] S. Schechter, M. Krishnan, and M.D. Smith. Using path profiles to predict HTTP requests. *Computer Networks and ISDN Systems*, 30:457–467, 1998.

[Son99] I. Sonin. The state reduction and related algorithms and their applications in the study of Markov chains, graph theory, and the optimal stopping problem. *Advances in Mathematics*, 145:159–188, 1999.

[TR93] J. Teuhola and T. Raita. Application of a finite-state model to text compression. *The Computer Journal*, 36:607–614, 1993.

[WZW98] A.D. Wyner, J. Ziv, and A.J. Wyner. On the role of pattern matching in information theory. *IEEE Transactions on Information Theory*, 44:2045–2056, 1998.

[ZL99] N. Zin and M. Levene. Constructing web views from automated navigation sessions. In *Proceedings of ACM Digital Library Workshop on Organizing Web Space (WOWS)*, pages 54–58, Berkeley, Ca., 1999.