Mining for User Navigation Patterns Based on Page Contents

Yue Xu
School of Software Engineering and Data Communications
Queensland University of Technology
Brisbane, QLD 4001, Australia
E-mail: yue.xu@qut.edu.au

Abstract

This paper describes two novel methods that capture the most popular or preferred navigation sessions visited by web users. Most existing techniques used by web log mining are access-based approaches that statistically analyze the log data that reflects the way past users have interacted with the site and do not pay much attention on the content of the pages. The methods proposed in this paper take page contents into account for mining user navigation patterns. The page content is characterized as a topic vector. The first method determines the popular sessions by examining the deviation probability of a session as well as comparing the distance between topic vectors. The second method determines the popular sessions by using entropy of a session as an estimator. Experiments have been conducted and the results show that, by considering page contents, the sessions which contain irrelevant pages can be ruled out by the two methods.

1 Introduction

Navigating through a large web site for finding relevant information can be tedious and frustrating. Recently, there is a growing interest in developing adaptive web site agents that assist users in navigating web sites to find the information of their interests by recommending relevant web documents to the users [4, 6]. For achieving this goal, an essential problem should be solved that is to find the web documents which are relevant to the user interests. Recent years, an increasing number of researchers have focused their study on applying data mining techniques to web log data analysis for finding regularities in web users' navigation patterns. The user navigation patterns are then used to determine the documents that are relevant to the user's interests. When users interact with a web site, the user's navigation tracks are stored in web server logs and the log data is a good collection of data for being analysed to capture the user navigation behaviour patterns. Some approaches have been proposed that analyze previous users' web logs to discover user navigation patterns such as popular navigation sessions, page clusters, and popular paths between web pages [1, 5, 6, 8, 9, 11]. These patterns are then used to classify a new user into a category and the pages related to that category will be recommended to the user.

The data available for web-based systems has two forms: the content of the web site itself and the access patterns of users to the site. Most techniques used by web log mining are access-based approaches that statistically analyze the log data that reflects the way past users have interacted with the site and do not pay much attention on the content of the pages. One such approach is the statistical model proposed by Borges and Levene [1]. In this model, the user navigation information, obtained from web logs, is modelled as a hypertext probabilistic grammar(HPG). The set of highest probability sessions generated by the grammar corresponds to the user preferred navigation trails (also called association rules). However, the association rules are generated by only analysing web log data. Nothing of the page contents has been considered. In this paper, we propose a technique that takes page contents into account for finding user popular navigation sessions. The content of a page can be expressed by a conceptual description languange for describing the topics involved in this page. The conceptual language may be simple conjunctions of attributes or complex and cognitively inspired descriptions as in [7]. In this paper, associated with a web page there is a topic vector which characterizes the conceptual aspect of that page. Firstly, in section 2 we give a brief review to the HPG model proposed by Borges and Levene. Then, in section 3, we present two novel methods to find the most popular navigation sessions based on user access data and page content as well. Finally, section 4 summarizes the paper.

2 A Review of the HPG Model

A log file is an ordered set of web page requests made by users. The requests are stored in the order that the server receives them. If multiple users are browsing the site concurrently, their requests are interminingled in the log file. The page requests made by one user can be extracted from the log file. Since it is expected that a user may visite a web site more than once, a user navigation session is usually defined as a sequence of pages visited by the same user such that no two consecutive pages are separated by more than a certain amount of time, for example, 30 minutes as many authors have adopted [1]. Techniques to infer user navigation sessions from log data are described in [2]. A collection of user navigation sessions can be described by a hypertext probabilistic language [3] generated by a hypertext probabilistic grammar (HPG) [1] which is a proper subclass of probabilistic regular grammars [10]. A HPG is a probabilistic regular grammar which has a one-to-one mapping between the sets of nonterminal and terminal symbols. Each nonterminall symbol corresponds to a web page and a production rule corresponds to a link between pages. Moreover, there are two additional artificial states, S and F, which represent the start and finish states of the navigation sessions. The probability of a grammar string is given by the product of the probabilities of the productions used in its derivation. The productions with S on its left-land side are called *start productions* and the productions corresponding to links between pages are called *transitive productions*.

From the set of user sessions we obtain the number of times a page was requested, the number of times it was the first state in a session, and the number of times it was the last state in a session. The number of times a sequence of two pages appears in the sessions gives the number of times the corresponding link was traversed. The probability of a production from a state that corresponds to a web page is proportional to the number of times the corresponding link was chosen relative to the number of times the user visited that page. A parameter α is used to attach the desired weight of a state corresponding to the first page browsed in a user navigation session. If α =0, only states which were the first in an actual session have probability greater than zero of being in a start production. In this case, only those strings which start with a state that was the first in a session are induced by the grammar. If α=1, the probability of a start production is proportional to the overall number of times the corresponding state was visited. In this case, the destination node of a production with higher probability corresponds to a state that was visited more often.

The HPG can be formally defined as follows.

Definition 2.1 (HPG) *: An HPG is a five-tuple* $< H, M, S, F, \Gamma >$ *which is denoted as* $\Phi(H, M, S, F, \Gamma)$*, where:*

- H *is a finite set of nodes,* $H = \{h_1, \ldots, h_m\}$ *which represents the set of states involved in the HPG. Each state represents a page request.*
- $M = \{p(h_i, h_j)\}$*,* $h_i, h_j \in H$*, is a* $m \times m$ *matrix, that is,* $\forall h_i, h_j \in H$*,* $1 \geq p(h_i, h_j) \geq 0$*,* $\Sigma_{k=1}^{m} p(h_i, h_k) = 1$*.*
  The matrix M *is called the* **transition matrix** *of* Φ *and the probabilities* $p(h_i, h_j)$ *are called the* **transition probabilities** *of* Φ *which are calculated by the mapping function* Γ*.* $\forall h_i \in H$*,* $p(S, h_i)$ *is the probability that a user will start his/her navigation by visiting the page associated with* h_i*.* $\forall h_i, h_j \in H$*,* $p(h_i, h_j)$ *is the probability that a user who is browsing the page associated with* h_i *will next browse the page associated with adjacent state* h_j*.*
- S *is the start state of the HPG.*
- F *is the finish state of the HPG.*
- Γ *is a function from* $H \times H$ *to* $[0, 1]$*, which is used to calculate* Γ*, i.e.,* $\forall h_i, h_j \in H$*,* $p(h_i, h_j) = \Gamma(h_i, h_j)$*.*

In the model proposed by Borges [1], the probabilities of the transition matrix M are determined from statistical information collected from the logs. For simplicity, we use h_i to denote the page which is associated with h_i. For a single user, let R be the total number of page requests, R_i be the number of requests to page h_i, N_s be the total number of sessions, and N_{si} be the number of sessions with h_i being the first page. The probabilities $p(S, h_i)$ are determined by the following equation, where S is the start state, h_i is a state in $H - \{S\}$, $1 \geq \alpha \geq 0$:

$$p(S, h_i) = \alpha \frac{R_i}{R} + (1 - \alpha) \frac{N_{si}}{N_s} \tag{2.1}$$

Similarly, let R_j^i be the number of page requests to page h_j immediately after visiting page h_i, R^i be the total number of page requests immediately after visiting page h_i, the probabilities $p(h_i, h_j)$ are determined

Table 2.1: An example set of user's sessions

Session Number	User Sessions
1	$A_1 \rightarrow A_2 \rightarrow A_3 \rightarrow A_4$
2	$A_1 \rightarrow A_5 \rightarrow A_3 \rightarrow A_4 \rightarrow A_1$
3	$A_5 \rightarrow A_2 \rightarrow A_4 \rightarrow A_6$
4	$A_5 \rightarrow A_2 \rightarrow A_3$
5	$A_5 \rightarrow A_2 \rightarrow A_3 \rightarrow A_6$
6	$A_4 \rightarrow A_1 \rightarrow A_5 \rightarrow A_3$

Table 2.2: The association rules obtained from the HPG in Figure 2.1 ($\lambda = 0.3$)

Session Number	User Sessions	Derivation Probabilities
1	$A_4 \rightarrow A_1$	0.5
2	$A_5 \rightarrow A_2 \rightarrow A_3$	0.45
3	$A_5 \rightarrow A_3$	0.4
4	$A_3 \rightarrow A_4$	0.4
5	$A_2 \rightarrow A_3 \rightarrow A_4$	0.3
6	$A_1 \rightarrow A_5 \rightarrow A_2$	0.3

by the following equation, where h_i, h_j are two states in $H - \{S\}$:

$$p(h_i, h_j) = \frac{R_j^i}{R^i} \tag{2.2}$$

The strings generated by the grammar correspond to the user navigation sessions. For each string $< h_{i1}, \ldots, h_{ir} >$ generated by the grammar, its derivation probability is defined as follows:

$$p(< h_{i1}, \ldots, h_{ir} >) = \prod_{k=1}^{k=r-1} p(h_{ik}, h_{i(k+1)}) \tag{2.3}$$

The aim of the HPG is to identify the subset of these sessions, which correspond to users' preferred sessions also called the association rules. A session is included in the grammar's language if its derivation probability is above a *cut-point* λ. The *cut-point* is responsible for pruning out strings whose derivation contains transitive productions with small probabilities.

Table 2.1 gives an example of a collection of user sessions. The collection of navigation sessions in the example contains a total of 24 page requests. The association rules generated by the HPG in the example are given in Table 2.2, where $\lambda = 0.3$.

3 Content-based HPG

As we have mentioned that HPG relies only on user access log data to discover the association rules. However, the data in web server logs may not really reflect user's navigation intention. One reason is that some data may be missing due to caching by the browser. This arises most commonly when the visitor users the browsers back button. For example, if the user returns back to a previous page p_i from the current page p_j and then goes to page p_k, this will appear in the log as $p_i \rightarrow p_j \rightarrow p_k$ but not $p_i \rightarrow p_j \rightarrow p_i \rightarrow p_k$. Another reason is that the user may visit some pages which are not relevant to his/her navigation interests and also recorded in the logs. These irrelevant pages in the user's navigation sessions can make great impact on the quality of the association rules. We argue that the page contents can be used to eliminate or alleviate the impact made by the missing data or the irrelevant pages.

3.1 A Modified Transition Matrix

From Equation 2.1 and Equation 2.2, we can see that the probabilities $p(h_i, h_j)$ in M are obtained from statistical information captured in logs only, none of the page content information has been used. In this subsection, we present a method to incorporate page content information into the probability calculation.

In this paper it is assumed that there are n topics involved in a web site, and that associated with each page in the site there is a n-dimensional vector which characterizes the relevancy of each topic to the page. The ith element in the vector represents the relevancy assigned to the ith topic. That is, $\forall h_i \in H$, there is a n-dimensional vector denoted as T_i, $T_i =< t_{i1}, \ldots, t_{in} >$, where t_{ij} represents the relevancy of the jth topic to the page h_i.

We observed that with a navigation goal in his/her mind, a user will visit the pages which are content relevant. This observation suggests that for two pages h_i and h_j whose topic vectors are $T_i =< t_{i1}, \ldots, t_{in} >$ and $T_j =< t_{j1}, \ldots, t_{jn} >$ respectively, if the distance between each pair of the corresponding vector elements t_{ik} and $t_{jk} (k = 1, \ldots, n)$, denoted as $D(t_{ik}, t_{jk})$, is big, the two pages can be thought of as irrelevant. The probability $p(h_i, h_j)$ should be increased if $D(t_{ik}, t_{jk})$ is small because h_j would be a good candidate to visit next after visiting h_i if the user wants to find some information which relevant to the information on page h_i. Under this consideration, we modify Equation 2.2 by taking the distance of page topic vectors into account to calculate the transition probabilities. Equation 3.1 below calculates the transition probability from two aspects. $p(h_i, h_j)$ is determined by Equation 2.2 which is the contribution from user access statistics. $\frac{\beta}{e^{2D(T_i, T_j)}}$ calculates the contribution from the page topics. In Equation 3.1, p_c denotes the modified probability and p is the probability calculated by Equation 2.2, $D(T_i, T_j)$ is the arithmetic average of the difference between T_i

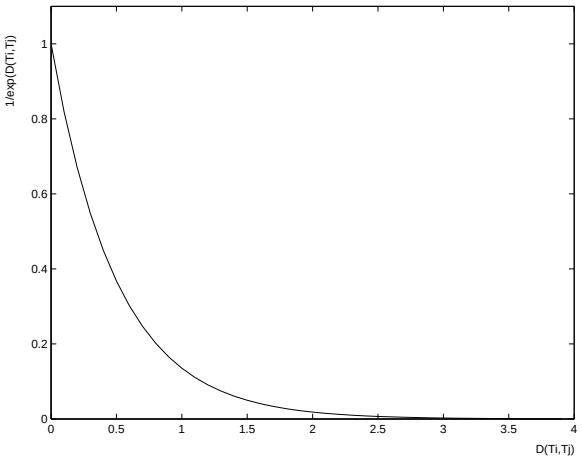

Figure 3.1: Function $\frac{1}{e^{2D(T_i,T_j)}}$

and T_j:

$$p_c(h_i, h_j) = p(h_i, h_j) + \frac{\beta}{e^{2D(T_i,T_j)}} \qquad (3.1)$$

$$where \quad D(T_i, T_j) = \frac{\Sigma_{k=1}^{k=m}|t_{ik} - t_{jk}|}{m}, 0 < \beta < 1$$

Figure 3.1 depicts the function $\frac{1}{e^{2D(T_i,T_j)}}$, from which we can see that the contribution from page topic vector to $p_c(h_i, h_j)$ is getting small as the distance between the two topic vectors gets larger.

Support there are five topics involved in the example above and each page in the example has a topic vetor as shown in Table 3.1, where value 1 indicates that the corresponding topic is involved in that page and values 0 indicates not. It can be figured out that the topic distance between page A_1 and A_4, A_5 and A_2, A_2 and A_3, A_5 and A_3 and A_1 and A_5 are the same, the distance between A_3 and A_4 is a bit larger. By using Equation 3.1, we get the association rules as given in Table 3.2. The probability of session 6 becomes larger than the probabilities of session 4 and session 5 after considering the impact of the page topics. The irrelevancy of A_3 and A_4 makes the probabilities of sessions 4 and 5 not increase much. This example shows that the consideration of the page content does have impact on the centainty of the association rules. If the cut-point λ increases, session 4 and session 5 will be ruled out from the set association rules before session 6 is.

Table 3.1: Topic vectors

Page	Topic Vector
A_1	$T_1 =< 1, 1, 0, 0, 1 >$
A_2	$T_2 =< 1, 1, 1, 0, 0 >$
A_3	$T_3 =< 1, 0, 1, 1, 0 >$
A_4	$T_4 =< 0, 1, 1, 0, 1 >$
A_5	$T_5 =< 1, 1, 0, 1, 0 >$

Table 3.2: The association rules obtained after considering topics ($\beta = 0.3, \lambda = 0.3$)

Session Number	User Sessions	Derivation Probabilities
1	$A_5 \rightarrow A_2 \rightarrow A_3$	0.67
2	$A_4 \rightarrow A_1$	0.64
3	$A_5 \rightarrow A_3$	0.54
6	$A_1 \rightarrow A_5 \rightarrow A_2$	0.47
4	$A_3 \rightarrow A_4$	0.46
5	$A_2 \rightarrow A_3 \rightarrow A_4$	0.36

3.2 Session Entropy and Pruning

In information theory, Shannon's measure of entropy is used as a measure of the information contained in a piece of data. For a random variable X with a set of possible values $< x_1, \ldots, x_n >$, having probabilities $p(x_i)$, $i = 1, \ldots, n$, if we had no information at all about the value X would be, the possibility for each value should be the same, i.e. $1/n$. In this case, X is in its most uncertain situation. According to information theory, the entropy of X reaches its maximum in this situation. On the other hand, if the entropy of X is close to zero, the value of X has few uncertainties. In this case, there should be a small set of values with high probabilities and others with very low probabilities. Based on this theory, we propose to use the entropies of topics in a session to prune the association rules.

$\forall h_i \in H$, its topic vector is $T_i =< t_{i1}, \ldots, t_{in} >$. Each topic can be treated as a random variable with two possible values: involved or not involved. The entropy of topic t_j to page h_i can be estimated by $H(t_{ij}) = -(p(t_{ij})logp(t_{ij}) + (1 - p(t_{ij}))log(1 - p(t_{ij})))$, where $p(t_{ij})$ is the probability of t_j being involved in h_i. Assume that $t_{i1}, \ldots, t_{in}$ are independent variables, according to entropy theory, we have $H(t_{i1}, \ldots, t_{in}) = \Sigma_{k=1}^{k=n} H(t_{ik})$, which estimates the certainty of topics $t_{i1}, \ldots, t_{in}$ being involved in h_i. Let $s_i =< h_{i1}, \ldots, h_{ir} >$ be a session which represents page sequence $h_{i1} \rightarrow h_{i2} \ldots \rightarrow h_{ir}$, $T_{ij} =< t_{ij_1}, \ldots, t_{ij_n} >$ be the topic vector of page h_{ij} with $1 \leq j \leq r$, $T_{s_i} =< t_{s_i1}, \ldots, t_{s_in} >$ be the topic vector of s_i, and $p(t_{ij_k})$ is the probability of the kth topic being involved in h_{ij}. The probability of the kth topic being involved in s_i denoted as $p(t_{s_ik})$ $(1 \leq k \leq n)$ can be estimated by the following equation:

$$p(t_{s_ik}) = \frac{\Sigma_{j=1}^{j=r} p(t_{ij_k})}{r} \qquad (3.2)$$

Table 3.3: Topic vectors of the sessions in the above example

Sessions	Topic Vector
s_1	$T_{s1} =< 0.5, 1, 0.5, 0, 0.5 >$
s_2	$T_{s2} =< 1, 0.67, 0.67, 0.67, 0 >$
s_3	$T_{s3} =< 1, 0.5, 0.5, 1, 0 >$
s_4	$T_{s4} =< 0.5, 0.5, 1, 0.5, 0.5 >$
s_5	$T_{s5} =< 0.67, 0.67, 1, 0.3, 0.3 >$
s_6	$T_{s6} =< 1, 1, 0.3, 0.3, 0.3 >$

Table 3.4: The association rules obtained from the HPG with entropies ($\lambda = 0.3$)

Sessions	User Sessions	Sesion Entropies
s_1	$A_4 \rightarrow A_1$	0.6
s_2	$A_5 \rightarrow A_2 \rightarrow A_3$	0.8
s_3	$A_5 \rightarrow A_3$	0.6
s_4	$A_3 \rightarrow A_4$	1.2
s_5	$A_2 \rightarrow A_3 \rightarrow A_4$	1.1
s_6	$A_1 \rightarrow A_5 \rightarrow A_2$	0.78

The entropy of a session can be estimated by the following equation:

$$H(s_i) = H(t_{s_i 1}, \ldots, t_{s_i n}) = \Sigma_{k=1}^{k=n} H(t_{s_i k}) \qquad (3.3)$$

where

$$H(t_{s_i k}) = -(p(t_{s_i k}) log p(t_{s_i k}) + (1 - p(t_{s_i k})) log(1 - p(t_{s_i k})))$$

The entropy of a session estimates the certainty of the topics involved in the session. If the entropy is small, then there must be some topics with high probabilities and the others with very low probabilities. In this case, this session is a good candidate to be selected as an association rule. On the other hand, if the entropy is large, then the probabilities of the topics must be very close and low as well. In this case, this session shouldn't be selected as an association rule since the pages in this session may not focus on some certain topics. For the example used in previous sections, Table 3.3 gives the topic vector of each session and Table 3.4 gives the entropy of each session. Table 3.4 suggests similar result as Table 3.2 does. That is, the session 4 and session 5 should be ruled out from the association rule set since their entropy is high.

4 Conclusion

In this paper we have proposed two methods that find the most popular navigation sessions based on both user access data and page contents. The first method modified the HPG model proposed by Borges [1] by taking the page content into account. This modification makes the HPG model more robust because the use of the page contents can eliminate or alleviate the impact made by the missing data or the irrelevant pages in access logs. The second method determines popular sessions according to session entropy which is based on well-established Information Theory.

References

[1] J. Borges and M. Levene. Data mining of user navigation patterms. In *Proceedings of the Web Usage Analysis and User Profiling*, volume 1, pages 31–36, 1999.

[2] R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, 1(1):5–32, February 1999.

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

[4] M. Pazzani and D. Billsus. Adaptive web site agents. *Autonomous Agents and Multi-Agent Systems*, 5:205–218, 2002.

[5] M. Perkowitz and O. Etzioni. Adaptive web sites: An ai challenge. In *Proceedings f IJCAI-97*, volume 1, 1997.

[6] M. Perkowitz and O. Etzioni. Towards adaptive web sites: Conceptual framework and case study. *Artificial Intelligence*, 118:245 – 275, 2000.

[7] S.Hanson and M.Bauer. Conceptual clustering, categorization, and polynorphy. *Machines Learning*, 3:343–372, 1989.

[8] T. Toolan and N. Kusmerick. Mining web logs for personalized site maps. In *Proceedings of the International Conference on Web Information Systems Engineering*, volume 1, 2002.

[9] S. Tso, H. Lau, and R. Ip. Development of a fuzzy push delivery scheme for internet sites. *Expert Systems*, 16:103–114, 1999.

[10] C. S. Wetherell. Probabilistic languages: A review and some open questions. *Computing Surveys*, 12(4):361–379, 1980.

[11] T. W. Yan, M. Jacobsen, H. Garcia-Molina, and U. Dayal. From user access patterns to dynamic hypertext linking. In *Proceedings of the fifth interntional world wide web*, volume 1, 1996.