Linear Approximation of Web user Categorization

Ratnesh K. Jain¹, Diwakar Shukla², Saurabh Jain³
¹Kendriya Vidyalaya, Arah, Bhojpur, Bihar, INDIA
²Department of Computer Science & Applications, Dr. H. S. Gour, University, Sagar, MP, INDIA
³Department of Computer Applications, Shri Vaishnav Institute of Technology and Science, Indore, MP, INDIA

ABSTRACT

Web logs are being utilized as huge data repository for mining interesting and potentially useful patterns. When mined properly these patterns provide the support to designer of the web site; analyst and management executives in strategic decision making such as improvement of content, structure or in making adaptive web sites. Web logs although contains many information but doesn't clearly indicate the page refreshing. If we are able to record the refreshing then in the Markov chain model we can add one more state which can be utilized to categorize the users of website into three categories i.e. faithful, Partially Impatient and Completely Impatient users. Jain et al. [3] in his paper derived some theorem to study each type of users’ behavior and shown that how do users behavior differ. In this paper we approximate the expression using straight line and least square method and did comparative study. We find that approximate expressions can be used to predict the behavior of the users which is independent of parameters also.

Keywords---- Web mining, Pattern discovery, Adaptive web sites, Markov chain model, and Transition probability.

I. INTRODUCTION

In a Web site with a large number of Web pages, users often have navigational questions, such as, where am I? Where have I been? And where can I go [1]? We can easily find out the answer of the first two questions i.e. Where am I? and Where have I been? But, the question where can I go? Is trickier and require prediction based on the previous trends. A good web site is that which has capability to help the users to find answers to all three questions. Web sites that change their presentation and organization to help users find the answer to the third question on the basis of next page access prediction are known as adaptive Web sites. There are so many methods are available to predict the next page access. One of them is Markov Chain model. Naldi [6] in 2002 gave an interesting contribution for the use of markov chain model towards the determination of traffic sharing among multiple operators in competitive market. In 2003 Zhao Li and Jeff Tian[4] tested the suitability of Markov Chain as web usage model and validated the model. After that many of the researchers uses this model to answer the third question where can I go?

Markov chain, named after Andrey Markov, is used to model a stochastic (random) process with the Markov property. Having the Markov property means that, given the present state, future states are independent of the past states. At each instant the system may change its state from the current state to another state, or remain in the same state, according to a certain probability distribution. The changes of state are called transitions, and the probabilities associated with various state-changes are termed transition probabilities.

II. RELATED WORK

The close resemblance between web applications and the state transition mechanism, Markov models have been widely used for modeling users’ web navigation behavior [2]. Several researchers have proposed the use of Markov chain models to model user access pattern.

Jain et al. [3] suggested new information in the web log to represent users refreshing activity. He used this refreshing information as a state named as ‘Refresh State’ and study the impact of refreshing on the transition probability. In their research work based on the refresh button click they divided the users into three categories: faithful user, partially impatient user and completely impatient user and did the simulation study to compare the users’ behavior.

The latest work in the field of web usage mining which is based on Markov chain model was proposed by Marques and Belo [5]. In their article they presented the application of Markov chains in the establishment of
students’ profiles for a target eLearning oriented Web site and described the entire process of discovering student profiles on an eLearning Web based platform.

III. MARKOV MODEL WITH REFRESH STATE (DUE TO [3])

![Figure 1.1. The Markov model with refresh state](image)

When we utilize the Markov model for web user categorization each web page of a website represents a state and navigation from one page to another is called transition. The whole website can be represented by a transition probability matrix (called transition matrix in short) which contains one-step transition probabilities in the Markov model. The Markov model is then used for next page access prediction by calculating the conditional probabilities of visiting other pages in the future given the user’s current position and/or previously visited pages.

Let our web site is consist of three web pages named A, B, C. We can introduce two more state namely Start State and Finish State represented by S and F respectively. If we assume that web log has an entry per refresh button click or refresh button click can be detected from the web logs, we can introduce a new state which we named Refresh State and represented by R. The model is shown in figure 1.1.

Name of the variables indicating the transition probability are shown as the label of the edges as shown in fig 1.1. A user can start accessing the Web site by pressing the refresh button nor he closes the web site. Hence indicating that without accessing any web page user neither.

![Figure 1.1. The Markov model with refresh state](image)

As per Jain, R. K. [3], the starting conditions are:

\[
\begin{align*}
\Pr[X^{(0)} = A] &= p_1, \\
\Pr[X^{(0)} = B] &= p_2, \\
\Pr[X^{(0)} = C] &= p_3, \\
\Pr[X^{(0)} = R] &= 0, \\
\Pr[X^{(0)} = F] &= 0.
\end{align*}
\]

And there are three types of users Faithful Users, Partially Impatient Users and Completely Impatient Users. The \(n\)th attempt probability for each type of user is specified below as derived in [3].

### 3.1 Some Results for \(n\)th Attempts (Due to [3])

A Faithful user at the \(n\)-th attempt can be either on a particular web page of which he is a Faithful User (for example he can be at state A if he is a Faithful user of page A) or it can be at state R. Then the probability of reaching on states A, B, C on the \(n\)th attempt can be calculated using the following theorems-

**Theorem 1:** The transition probability of reaching on state A at \(n\)th attempt when user is faithful

(i) When \(n\) is even

\[
\Pr[X^{(2n)} = A] = p_1 q_3^n q_1^n,
\]

For \(n = 0, 1, 2, 3, \ldots\)

(ii) When \(n\) is odd

\[
\Pr[X^{(2n+1)} = A] = 0,
\]

For \(n = 0, 1, 2, 3, \ldots\)

**Theorem 2:** The transition probability of reaching on state B at \(n\)th attempt when user is faithful

(i) When \(n\) is even:

\[
\Pr[X^{(2n)} = B] = p_2 q_7^n q_{14}^n,
\]

For \(n = 0, 1, 2, 3, \ldots\)

(ii) When \(n\) is odd:

\[
\Pr[X^{(2n+1)} = B] = 0,
\]

For \(n = 0, 1, 2, \ldots\)

**Theorem 3:** The transition probability of reaching on state C at \(n\)th attempt when user is faithful

(i) When \(n\) is even:

\[
\Pr[X^{(2n)} = C] = p_3 q_{11}^n q_{15}^n,
\]

For \(n = 0, 1, 2, 3, \ldots\)

(ii) When \(n\) is odd:

\[
\Pr[X^{(2n+1)} = C] = 0,
\]

For \(n = 0, 1, 2, 3, \ldots\)

### 3.1.2 \(n\)th Attempt Results for Partially Impatient User (Due to [3])

If user continuous with attempts to access web pages then at the \(n\)-th attempt he can be at state A, B, C or user can do refresh which we represent the state R that means user can refresh which we represent the state R that means user can

\[
\begin{align*}
\Pr[X^{(0)} = A] &= p_1, \\
\Pr[X^{(0)} = B] &= p_2, \\
\Pr[X^{(0)} = C] &= p_3, \\
\Pr[X^{(0)} = R] &= 0, \\
\Pr[X^{(0)} = F] &= 0.
\end{align*}
\]

\[
\begin{bmatrix}
0 & q_1 & q_2 & q_3 & q_4 \\
q_5 & 0 & q_6 & q_7 & q_8 \\
q_9 & 0 & q_{10} & q_{11} & q_{12} \\
q_{13} & q_{14} & 0 & q_{15} & q_{16} \\
0 & 0 & 0 & 0 & 1
\end{bmatrix}
\]

\[
T = \begin{bmatrix}
0 & q_1 & q_2 & q_3 & q_4 \\
q_5 & 0 & q_6 & q_7 & q_8 \\
q_9 & 0 & q_{10} & q_{11} & q_{12} \\
q_{13} & q_{14} & 0 & q_{15} & q_{16} \\
0 & 0 & 0 & 0 & 1
\end{bmatrix}
\]
When \( n \) is even

\[
\begin{align*}
P[X^{(2n)} = A] &= \frac{1}{2^n} q_1^n q_5^n + \frac{1}{2^n} q_1^n q_6^n + \frac{1}{2^{n+1}} q_5^n q_9^n + \frac{1}{2^{n+1}} q_6^n q_{10}^n \\
&\quad + \frac{1}{2^{n+1}} q_5^n q_{11}^n \\
&\quad \text{For } n=1,2,3,\ldots
\end{align*}
\]

When \( n \) is odd

\[
\begin{align*}
P[X^{(2n+1)} = A] &= \frac{1}{2^n} q_1^n q_5^n + \frac{1}{2^n} q_1^n q_6^n + \frac{1}{2^{n+1}} q_5^n q_9^n + \frac{1}{2^{n+1}} q_6^n q_{10}^n \\
&\quad + \frac{1}{2^{n+1}} q_5^n q_{11}^n \\
&\quad \text{For } n=0,1,2,3,\ldots
\end{align*}
\]

**Theorem 5:** The transition probability of reaching on state B at \( n^{th} \) attempt when user is partially impatient-

(i) When \( n \) is even

\[
P[X^{(2n)} = B] = \frac{1}{2^n} q_1^n q_5^n + \frac{1}{2^n} q_1^n q_6^n + \frac{1}{2^n} q_2^n q_5^n + \frac{1}{2^n} q_2^n q_6^n \\
&\quad + \frac{1}{2^{n+1}} q_2^n q_{10}^n \\
&\quad \text{For } n=1,2,3,\ldots
\]

(ii) When \( n \) is odd

\[
P[X^{(2n+1)} = B] = \frac{1}{2^n} q_1^n q_5^n + \frac{1}{2^n} q_1^n q_6^n + \frac{1}{2^n} q_2^n q_5^n + \frac{1}{2^n} q_2^n q_6^n + \frac{1}{2^{n+1}} q_2^n q_{10}^n \\
&\quad + \frac{1}{2^{n+1}} q_2^n q_{11}^n \\
&\quad \text{For } n=0,1,2,3,\ldots
\]

**Theorem 6:** The transition probability of reaching on state C at \( n^{th} \) attempt when user is partially impatient-

(i) When \( n \) is even

\[
P[X^{(2n)} = C] = \frac{1}{2^n} q_1^n q_5^n + \frac{1}{2^n} q_1^n q_6^n + \frac{1}{2^n} q_2^n q_5^n + \frac{1}{2^n} q_2^n q_6^n \\
&\quad + \frac{1}{2^{n+1}} q_2^n q_{10}^n \\
&\quad \text{For } n=1,2,3,\ldots
\]

(ii) When \( n \) is odd

\[
P[X^{(2n+1)} = C] = \frac{1}{2^n} q_1^n q_5^n + \frac{1}{2^n} q_1^n q_6^n + \frac{1}{2^n} q_2^n q_5^n + \frac{1}{2^n} q_2^n q_6^n + \frac{1}{2^{n+1}} q_2^n q_{10}^n \\
&\quad + \frac{1}{2^{n+1}} q_2^n q_{11}^n \\
&\quad \text{For } n=0,1,2,3,\ldots
\]

IV. LINEAR APPROXIMATION

Consider two variables Y and X as

\[Y = P[X = .]\]

\[X = n\]

Now we can draw line \( Y = a + bX \) using principle of least square.

The two normal equations are

\[
\begin{align*}
\sum Y &= ma + \sum X \\
\sum XY &= a \sum X + b \sum X^2
\end{align*}
\]

Where \( m \) is number of pair points available on the graph,

\[
\hat{b} = \frac{m \sum XY - (\sum X)(\sum Y)}{m \sum X^2 - (\sum X)^2}
\]

\[
\hat{a} = \frac{1}{n} [\sum Y - m \sum X]
\]

The line is plotted as approximation:

\[Y = \hat{a} + \hat{b}X\]

Averaging over many varying parameters \( p, q \) etc. provide average relationship-

\[Y = \bar{a} + \bar{b}X\]

V. SIMULATION STUDY
The linear approximation is having downward trend in both the cases. It is matching with the original trend as observed in [3]. Further, the line pattern is independent of q-parameters.

The approximate relationship between state probability and transition number is showing downward trend as observed in above graph. The line is independent of q-values.

The probability of reaching on page C is decreasing with the number of attempt which is similar to the trends shown in the fig 6(a) and 6(b) of previous graph in [3]. The line in 7(a) and 7(b) is independent of the q-values.
The graph in Fig 9(a), 9(b) and 9(c) is declining with the number of transactions as in the fig 8(a), 8(b) and 8(c) which suggest that 9(a), 9(b) and 9(c) is linear approximation of the fig 8(a), 8(b) and 8(c) and it is independent of the values of p and q.

The approximate relationship between state probability and transition number for the completely impatient user of page A is shown in the fig 11(a), 11(b) and 11(c) which shows similar trends as shown in the figure 10(a), 10(b) and 10(c).

The approximated behavior of Completely Impatient user of page B is showing downward trend which means the transition probability is decreasing with respect to the number of attempts. This trend is similar to the trend shown in previous graph of [3]. Which means linear approximation is effective and independent of the p and q-values.
The behavior of completely impatient user of page C in the linear approximation is again trending downward which is similar to the previous trend in [3]. Here again the values of p and q are not affecting the behavior.

VI. CONCLUSION

The paper shows that linear approximation could be made effectively to the probability values and number of transitions. The basic feature of this approximation is that they are independent of model parameters. One can use the line approximation probabilities directly for computation of state probabilities in the Markov chain model.

REFERENCES