Web User Categorization and Behavior Study Based on Refreshing

Ratnesh Kumar Jain[1]
Dr. Sanjay Singh Thakur[1]
Dr. Ramveer Singh Kasana[1]
[1]Department of Computer Science and Applications,
Dr. H. S. Gour Central University, Sagar, MP (India)
jratnesh@rediffmail.com, sanjaymca2002@yahoo.com, irkasana7158@gmail.com
Suresh Jain[2]
[2]Department of Computer Engineering, Institute of Engineering & Technology,
Devi Ahilya University, Indore, MP (India)
suresh.jain@rediffmail.com

---**ABSTRACT**---

As the information available on World Wide Web is growing the usage of the web sites is also growing. Since each access to the web pages are recorded in the web logs it is becoming a huge data repository which when mined properly can provide useful information for decision making. The designer of the web site, analyst and management executives are interested in extracting this hidden information from web logs for decision making. In this research paper we proposed a method to categorize the users into faithful, Partially Impatient and Completely Impatient user, page wise so that study of user behavior can be easier. To categorize the user we proposed one new information in the web log that represent each instance of refreshing. We used the markov chain model in which we treated the clicking of Refresh button as another state i.e. Refresh State. We derive some theorem to study each type of user behavior and show that how do users behavior differ.

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

Paper submitted: 02 June 2009 Accepted: 13 Sept. 2009

1. Introduction

World Wide Web is a huge data repository and is growing with the explosive rate of about 1 million pages a day. As the information available on World Wide Web is growing access to the web sites by its users is also growing. 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. Useful information about Web users, such as their interests indicated by the pages they have visited, could be used to make predictions on the pages that might interest them. But, this type of information has not been fully utilized to provide a satisfactory answer to the third question. 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 to use Markov model to represent the users' past visiting behavior and apply the Markov chain to predict the web page that may be visited next by the user. In this approach web logs are used to represent the previous visiting behavior. Pages that can be accessed by a user in the next click are evaluated based on the current position of user in the web site and his/her visiting history in the Web site stored in the web log. Web site can be represented using a graph called next link graph which is based on user previous access behavior stored in web log. This graph is presented in the form of transition probability matrix called transition matrix in short which contains one-step transition probabilities (T.P.) 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.

In general reaching on a web page is represented as one state of the Markov chain model but many times we find that after typing the URL the web page is not open. In such case user either switches to other web page or press the refresh button. But web log does not have any entry for refresh button. Thus there is no state to represent this situation. 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 represent this situation by introducing a new state that will represent the refresh button click. Then we can study the impact of the refreshing (using refresh state) on the Transition Probability (T.P.) which is the theme of this chapter. In this research work based on the refresh button click we divided the user into three categories: faithful user, partially impatient user and completely impatient user and did the simulation study to compare the users' behavior.

2. Related Work

With the prevalence of the World Wide Web and people's reliance on it in society today, ensuring satisfactory performance and reliability of web servers and web sites is becoming increasingly important. Because of 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. Sarukkai [3] proposed Markov models for predicting the next page accessed by the user. Cadez et al. [4] utilize Markov models for classifying browsing sessions and a clustering technique to group users with similar navigation patterns (each cluster is represented by a Markov model). Eirinaki et al. [5] propose a method that incorporates link analysis, such as the pagerank measure, into a Markov model in order to provide Web path recommendations.

Jianhan Zhu, Jun Hong, John G. Hughes in [6] used Markov models to find conceptual clusters of Web pages based on link similarities between Web pages. In [7], they presented PageRate algorithm to give search results ratings based on past users' accumulated navigation behaviors on Web sites, and PageClustering algorithm to cluster Web pages with similar in-links to form conceptual categories to integrate with search results. In [8], they used a transition matrix compression algorithm to compress the Markov model of a Web site to an optimal size for efficient link prediction on the Web site. Kleinberg [9] proposed HITS algorithm to find authorities and hubs based on the Web link structure.

Several extensions to the simple Markov chain model have introduced. For example: Anderson et al. [10] proposed a Markov model's extension that incorporates relational predicates into states. Deshpande et al. [11] propose techniques for combining different order Markov models to obtain low state complexity and improved accuracy. Albrecht et al. [12] built a hybrid Markov model which combined four Markov models for pre-fetching documents. They assumed that the page sequence a user had visited was a Markov chain and applied the time factor in the Markov model. Deshpande and Karypis [11] proposed a technique that builds kth-order Markov models and combines them to include the highest order model covering each state; a technique to reduce the model complexity is also proposed. Pitkow et al[13] built a path-based system. They wanted to find out the longest repeating page subsequence (a path) that all users have visited. Su et al. [14] applied the n-gram language model into the pre-fetching systems. They considered a sequence of n web pages as an n-gram. By counting the times each n-gram appears, they give the prediction based on the maximal count.

We can find so many research papers of Jose´ Borges and Mark Levene based on Markov chain model. In [15] they proposed a first-order Markov model for a collection of user navigation sessions, and, more recently, they have extended the method to represent higher order conditional probabilities by making use of a cloning operation [16], [14]. In addition, they have proposed a method to evaluate the predictive power of a model that takes into account a variable-length history when estimating the probability of the next link choice of a user, given his or her navigation access sequence [17]. In [18] they proposed a new method to measure the summarization ability of a model, by which we mean the ability of a variable-length Markov model to summarize user access pattern up to a given length. An alternative approach to model user sessions is tree-based models. For example Dongshan and Junyi [19] propose a hybrid-order tree-like Markov model to predict web page access which provides good scalability and high coverage of the state space, also to predict the next page access. Chen and Zhang [20] utilized a Prediction by Partial Match forest that restricts the roots to popular nodes; assuming that most user sessions start in popular pages, the branches having a non popular page as their root are pruned.

In our research we use a simple Markov model with a new state named 'Refresh State'. And study the impact of refreshing by varying the probability of refreshing on the next page access probability.

3. Markov Chain Model

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. In other words, the description of the present state fully captures all the information that could influence the future evolution of the process. That means future states will be reached through a probabilistic process instead of a deterministic one.

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 (T.P.).

In general Markov chain model can be defined by a tuple $< X , T , \lambda >$, where

1. X is a set of states called state space
2. Transition Matrix T, and
3. λ is the initial probability distribution on the states in X.

3.1 Building Markov Models from Web Log Files

Markov models have also been used to analyze web navigation behavior of users. A user's web navigation on a particular website can be modeled using first- or second-order Markov models and can be used to make predictions regarding future navigation and to personalize the web page for an individual user. To understand how is Markov chain model is used to model user access patterns we use the following example-

Example : Suppose we have a web site of three pages {I, J, K}. Following are the sequence of user access per session mined from web logs-

(1) I, J, K, J (2) I, K, J, K, K (3) J, K, J (4) R, I, J.

Each navigation session suggests the order in which sequence of pages accessed by a user. We can think each page accessed by the user as a state. We introduce two more states say start state S: representing the access to the first page and a final state F: representing the last state. Hence our state space X={S, I, J, K, F}.

The navigation from one page to another can be thought as state transition. In each session, each sequence of two pages say I, J corresponds to a transition from one state (I) to another (J). Figure 1 suggests the transition diagram for first session according to our example.

Fig 1 shows the complete model for the set of navigation sessions given in the example. Since final state is reachable from every other state and there is no out transition from final state, the state F is an absorbing state, hence this model is absorbing Markov chain model. According to the discussion above we can formally define the Markov chain model for our example as follows:

X={S, I, J, K, F}, λ=<3/15, 6/15, 6/15, 0>

$$T=\begin{array}{c|cccc|} & I & J & K & F \\ I & 0 & 2/3 & 1/3 & 0 \\ J & 0 & 0 & 1/2 & 1/2 \\ K & 1/6 & 3/6 & 1/6 & 1/6 \\ F & 0 & 0 & 0 & 1 \end{array}$$

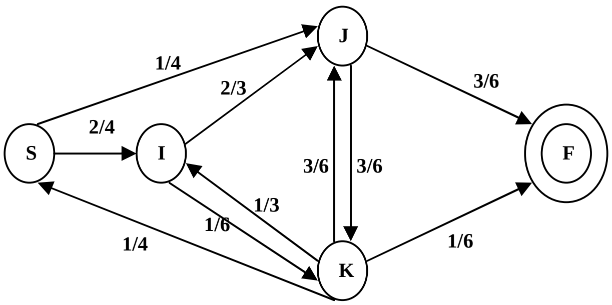


Figure 1: The complete Markov model for all sessions given in example

4. Markov Model with Refresh State

Let our web site is consist of three web pages named A, B, C. As in the example in previous section we can introduce two more state namely Start State and Finish State represented by S and F respectively. Since there is no information in the web log for refresh button click. We assume that web log has an entry per refresh button click or refresh button click can be detected from the web logs, and we can introduce a new state say Refresh State R, that will represent the refresh button click. Then the model is shown in fig 2.

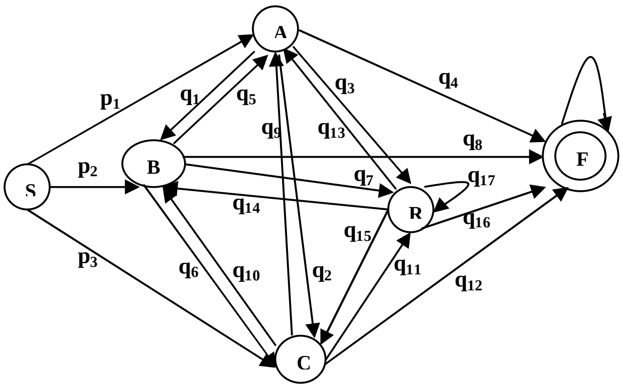


Figure 2: The Markov model with refresh state

As shown in the fig 2, we have the variables to represent transition probability (T.P.). Name of the variables indicating the T.P. are shown as the label of the edges. A user can start accessing the Web site by accessing any of the three web pages. Labels on edges from S to other states are indicating the initial probability. As shown in the figure there in no edge from S to R and S to F indicating that without accessing any web page user neither presses the refresh button nor does he close the web site. Hence the tuple < X , T , λ > that define Markov chain model is as follows:

X = { S, A, B, C, R, F }

$$T=\begin{array}{c|ccccc|} & A & B & C & R & F \\ A & 0 & q_1 & q_2 & q_3 & q_4 \\ B & q_5 & 0 & q_6 & q_7 & q_8 \\ C & q_9 & q_{10} & 0 & q_{11} & q_{12} \\ R & q_{13} & q_{14} & q_{15} & q_{16} & q_{17} \\ F & 0 & 0 & 0 & 0 & 1 \end{array}$$

$$\lambda = \{p_1, p_2, p_3, 0, 0\}$$

4.1 Some Results for nth Attempts

As shown above, the starting conditions are: $P\left[X^{(0)}=A\right]=p_1$, $P\left[X^{(0)}=B\right]=p_2$, $P\left[X^{(0)}=C\right]=p_3$, $P\left[X^{(0)}=R\right]=0$, $P\left[X^{(0)}=F\right]=0$.

As we have discussed, there are three types of users Faithful Users, Partially Impatient Users and Completely Impatient Users. We formally define each type of user:

4.1.1. **Faithful User:** Users that try to open a particular page but press refresh button if page is not open soon and do not switch to other web pages, can be categorized into faithful users.

4.1.2. **Partially Impatient Users:** A user that try to open a particular page if the page does not open he either press refresh button or he switch to any other page can be categorized as Partially Impatient Users. User has little patient to wait and again try to open that particular page.

4.1.3. Completely Impatient Users: Users that try to open a particular page if it does not open they switch to any other page but do not refresh can be categorized as completely impatient user. User can not wait or retry he give up the access to the page and go for any other page. The probability function for nth attempt is different for different type of users. Deriving the probability function requires very large space. Therefore, we are not giving the derivation in this paper. In this section we only summarize the nth attempt probability function for each type of user in form of theorem.

4.1.1. nth Attempt Results for Faithful User

A Faithful user at the n-1th 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 nth attempt can be calculated using the following theorems-

Theorem 1: The T.P. of reaching on state A at nth attempt when user is faithful

(i) When n is even

$$P\left[X^{(2n)}=A\right]=p_1q_3^nq_{13}^n\ ,\quad \text{For n=0,1,2,3….}$$

(ii) When n is odd

$$P\left[X^{(2n+1)}=A\right]=0\ ,\ \text{For n= 0,1,2,3……}$$

Theorem 2: The T.P. of reaching on state B at nth attempt when user is faithful

(i) When n is even

$$P\left[X^{(2n)}=B\right]=p_2q_7^nq_{14}^n\ ,\quad \text{For n=0,1,2,3…….}$$

(ii) When n is odd

$$P\left[X^{(2n+1)}=B\right]=0\ ,\ \text{For n= 0,1,2,3……}$$

Theorem 3: The T.P. of reaching on state C at nth attempt when user is faithful-

(i) When n is even

$$P\left[X^{(2n)}=C\right]=p_3q_{11}^nq_{15}^n\ ,\quad \text{For n=0,1,2,3…….}$$

(ii) When n is odd

$$P\left[X^{(2n+1)}=C\right]=0\ ,\ \text{For n= 0,1,2,3……}$$

4.1.2. nth Attempt Results for Partially Impatient User

If user continuous with attempts to access web pages then at the n-1th attempt he can be at state A, B, C or user can do refresh which we represent the state R that means user can be at state A or B or C or R then the probability of reaching on states A, B, C on the nth attempt can be calculated using the following theorems-

Theorem 4: The T.P. of reaching on state A at nth attempt when user is partially impatient-

(i) When n is even

$$P\left[X^{(2n)}=A\right]=p_1q_1^nq_5^n+p_1q_2^nq_9^n+p_2q_2^{n-1}q_6q_9^n+p_3q_1^{n-1}q_5^nq_{10}$$

For n=1,2,3…

(ii) When n is odd

$$P\left[X^{(2n+1)}=A\right]=p_2q_1^nq_5^{n+1}+p_3q_2^nq_9^{n+1}$$

For n= 0,1,2,3……

Theorem 5: The T.P. of reaching on state B at nth attempt when user is partially impatient-

(i) When n is even

$$P\left[X^{(2n)}=B\right]=p_1q_2^nq_9^{n-1}q_{10}+p_2q_1^nq_5^n+p_2q_6^nq_{10}^n+p_3q_1^nq_5^{n-1}q_9$$

For n=1,2, …

(ii) When n is odd

$$P\left[X^{(2n+1)}=B\right]=p_1q_1^{n+1}q_5^n+p_3q_6^nq_{10}^{n+1},$$

For n= 0,1,2,3……

Theorem 6: The T.P. of reaching on state C at nth attempt when user is partially impatient-

(i) When n is even

$$P\left[X^{(2n)}=C\right]=p_1q_1^nq_5^{n-1}q_6+p_2q_2^nq_5q_9^{n-1}+p_3q_2^nq_9^n+p_3q_6^nq_{10}^n$$

For n=1,2,3…

(ii) When n is odd

$$P\left[X^{(2n+1)}=C\right]=p_1q_2^{n+1}q_9^n+p_2q_6^{n+1}q_{10}^n$$

For n= 0,1,2,3……

4.1.3. nth Attempt Results for Completely Impatient User

A Completely impatient user if continuous with attempts to access web pages then at the n-1th attempt he can be at state A, B and C then the probability of reaching on states A, B, C on the nth attempt can be calculated using the following theorems-

Theorem 7: The T.P. of reaching on state A at nth attempt when user is completely impatient:

(i) When n is even

$$P\left[X^{(2n)}=A\right]=p_1q_1^nq_5^n+p_1q_2^nq_9^n+2^np_1q_3^nq_{13}^n+p_2q_6^nq_9q_{10}^{n-1}$$
$$+p_3q_1^{n-1}q_5^nq_{10}\ ,\text{For n =1,2,3…….}$$

(ii) When n is odd

$$P\left[X^{(2n+1)}=A\right]=2^n(p_2q_5+p_3q_9)q_3^nq_{13}^n,$$

For n =0,1,2,…

Theorem 8: The T.P. of reaching on state B at nth attempt when user is completely impatient:

(i) When n is even

$$P\left[X^{(2n)}=B\right]=p_2q_1^nq_5^n+p_2q_6^nq_{10}^n+2^np_2q_7^nq_{14}^n+$$
$$p_1q_2^nq_9^{n-1}q_{10}+p_3q_1^nq_5^{n-1}q_9\ ,\text{For n = 1,2,3...}$$

(ii) When n is odd

$$P\left[X^{(2n+1)}=B\right]=2^n(p_1q_1+p_3q_{10})q_7^nq_{14}^n,$$

For n= 0,1,2,3……

Theorem 9: The T.P. of reaching on state C at nth attempt when user is completely impatient:

(i) When n is even

$$P\left[X^{(2n)}=C\right]=p_1q_1q_6^nq_{10}^{n-1}+p_2q_1^{n-1}q_2q_5^n+p_3q_2^nq_9^n+$$
$$p_3q_6^nq_{10}^n+2^np_3q_{11}^nq_{15}^n\ ,\text{For n = 1,2,3,.... .}$$

(ii) When n is odd

$$P\left[X^{(2n+1)} = C\right] = 2^n (p_1 q_2 + p_2 q_6) q_{11}^n q_{15}^n ,$$

For n= 0,1,2,3……

5. Simulation Study

Initially we observe the behavior of each type of user individually by varying the value of refreshing probability (q_3 and q_{13} for page A, q_7 and q_{14} for page B and q_{11} and q_{15} for page C) of each type of user. We draw the graphs between the number of attempt and T.P. The graphs are as follows:

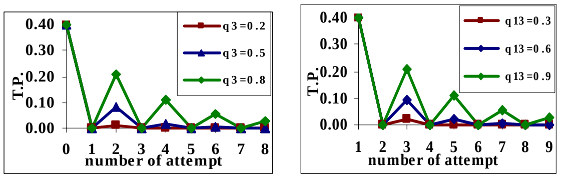


Fig 3(a) Fig 3(b)

Figure 3(a) and 3(b): Faithful user of page A.

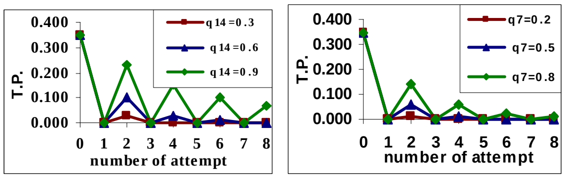


Fig 4(a) Fig 4(b)

Figure 4(a) and 4(b): Faithful user of page B.

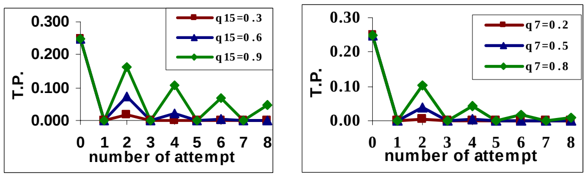


Fig 5(a) Fig 5(b)

Figure 5(a) and 5(b): Faithful user of page C.

As shown in the fig 3(a), 3(b), 4(a), 4(b), 5(a) and 5(b) above, we find that for any value of $q_{3,}$ $q_{13,}$ $q_{7,}$ $q_{14,}$ q_{11} and $q_{15,}$ T.P. decreases as the number of attempts increases. With higher value of q_3**,** initially T.P. is high but it decreases rapidly as attempt increases. While for lower value of q_3 initially T.P. is low and it decreases slowly. The same results obtained when we vary the T.P. q_{13} i.e. the T.P. of going from refresh state R to state A. The second trend we find is that for faithful user varying the probability of going from any page to refresh state and from Refresh state to any page does not affect the T.P. That is T.P. from any page to refresh state and from Refresh state to any of the page is same. The T.P. for even attempt is decreasing as the attempt is increasing. While the T.P. for odd attempt is zero. That is why we got fluctuating graph. The graphs for web page B and C have same pattern as for page A only difference is their initial probability. This indicates that behavior of the faithful user is same regardless of web pages.

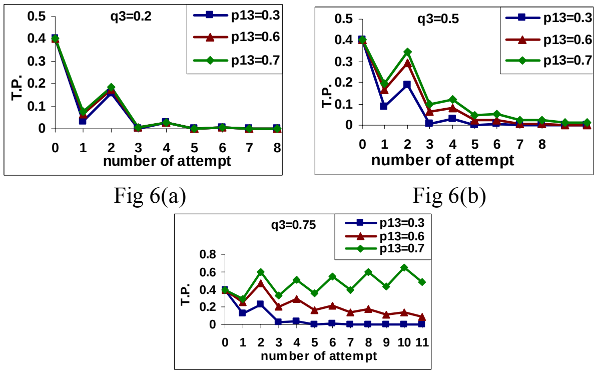


Fig 6(a) Fig 6(b)

Fig 6(c)

Figure 6: Partially Impatient user of page A.

The graphs for Partially Impatient Users are different than faithful users and Completely Impatient users as shown in fig 6(a), 6(b), 6(c). With the low value (0.2) of q_3 when we vary q_{13} from 0.3 to 0.7 completely impatient user's behavior is same. The graph shows that till 5th attempt for even number of attempt, T.P. is high and it drops at odd number of attempt and increases at even number of attempt but for further attempts it becomes zero irrespective of even or odd. For medium value (0.5) of q_3 and different values of q_{13} T.P. with respect to number of attempt is fluctuating but reducing towards zero as the attempt is increasing. Also for higher value of q_{13} graph is fluctuating at the higher side of T.P. and for higher value of q_{13} it fluctuating comparatively lower side of T.P. But after 8th attempt it becomes zero.

Graphs for higher value of q_3 (q_3=0.75) shows very different results. It shows that for higher value of q_{13} (i.e. q_{13}=0.7), possibility of having Partially Impatient user at page A is increasing as the number of attempt increasing, while for moderate value of q_{13} (i.e. q_{13}=0.5) T.P. is decreasing slowly and for lower value of q_{13} (i.e. 0.3) the T.P. is decreasing and it becomes zero from 5th attempt onwards.

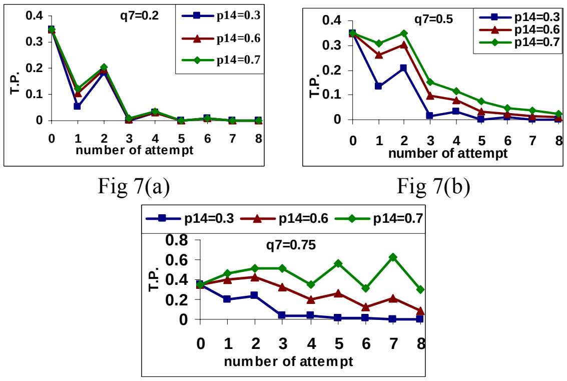


Fig 7(a) Fig 7(b)

Fig 7(c)

Figure 7: Partially Impatient user of page B

In fig 7, we find same behavior for the lower and medium value of q_7 but for higher value of q_7 (i.e. q_7=0.75) each line shows different behavior. Line for q_{14}=0.3 shows initially fluctuating pattern then it become zero. Line for q_{14}=0.6, increased up to second attempt then decrease up to forth then fluctuate upward and downward on every next attempt but slowly declining towards zero. Line for q_{14}=0.7, increases up to fourth attempt very slowly but after that fluctuate upward and downward but upward peak goes higher slowly. That means at the higher number of attempt probability of having at page B of the Partially Impatient user is higher if both q_7 and q_{14} (i.e. probability of going from page B to refresh state and from refresh state to page B) are high, zero if any of the q_7 and q_{14} are low and very less (about zero) if both q_7 and q_{14} are medium or if one is medium and one is high.

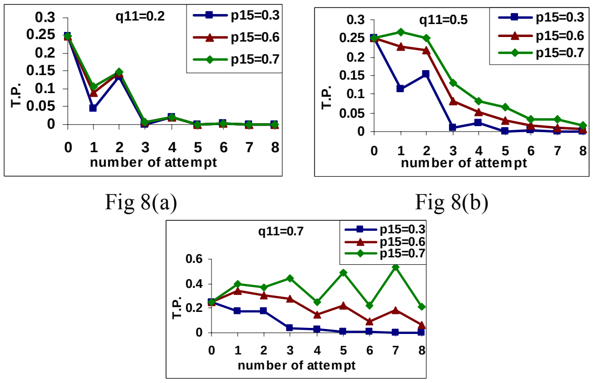


Fig 8(a)

Fig 8(b)

Fig 8(c)

Figure 8: Partially Impatient user of page C

Fig 8(b) shows that for q_{11}=0.5 and q_{15}=0.7, T.P. increased on the first attempt, on second attempt it decreases very slowly but on the third attempt T.P. is dropped drastically. And for the further attempt it slowly decline towards zero. For q_{11}=0.5 and q_{15}=0.6, T.P. decreased slowly till second attempt but on the third attempt it decreases drastically and on the further attempt it goes towards zero.

Behavior of the Partially Impatient user when q_{11}=0.7 and q_{15}=0.7 is again fluctuating, for odd terms it is increasing and for even terms it is decreasing. That is, if we draw a line for odd terms only we find an inclined line which means probability of having at page C is increase as the attempt increases. But when q_{11}=0.7 and q_{15}=0.6 graph is declined till forth attempt then we get fluctuating graph. For q_{11}=0.7 and q_{15}=0.3 as attempt increases probability of having on page C is decreases rapidly and very soon becomes zero.

One important conclusion we can derive from the graphs shown in figures from 3 to 7 that to study the users' behavior we should restrict the attempt number till five.

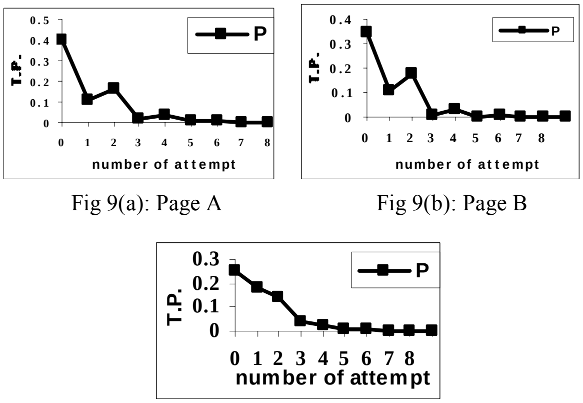


Fig 9(a): Page A

Fig 9(b): Page B

Fig 9(c): Page C

Figure 9: Completely Impatient User.

To study the behavior of the Completely Impatient user we again draw the graphs, page wise, between number of attempt and T.P. as shown in the fig 9(a), 9(b) and 9(c). With the initial probability for page A p_1=0.4 we find again fluctuating graph. Where for both even and odd terms, T.P. is decreasing rapidly as attempt increases. After sixth attempt T.P. becomes zero. Same pattern is shown for page B with initial probability p_2=0.35. But with the initial probability p_3=0.25, graph for page C shows some different pattern. T.P. for page C is decreasing as the number of attempt is increasing but constantly without any influence of even number of attempt or odd number of attempt. Since expression for calculating T.P. for page A, B and C have symmetry. We can conclude that if the initial probability is low then behavior of partially impatient user is different that is reflected as constant declining graph while for higher initial probability Completely impatient user behavior is different that is indicated by fluctuating graph.

Fig 5.10(a) Fig 5.10(b)

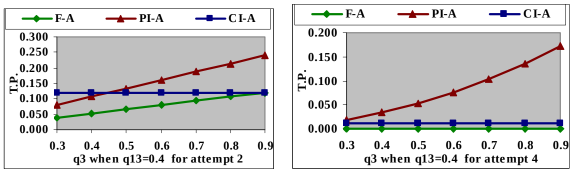


Fig 10(a)

Fig 10(b)

Figure 10: Comparison between different users of page A for even attempt when q_{13} is low.

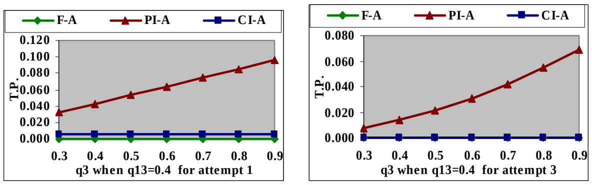


Fig 11(a)

Fig 11(b)

Figure 11: Comparison between different users of page A for odd attempt when q_{13} is low.

Fig 10 shows that, there is no impact on the T.P. on Completely Impatient user for any value of q_3 for attempt number two and four. That means Completely Impatient user of page A is not affected by refreshing. But for attempt four T.P. of Completely Impatient users is very low. The T.P. of Faithful User and Partially Impatient user for attempt two is increasing with the refreshing probability but the difference between both is that T.P. for Partially Impatient users' growing more rapidly than faithful user. While on 4th attempt, for low value of refreshing, T.P. of all types of user are approximately same. But T.P. of faithful user is approximately zero and T.P of Partially Impatient users' is growing like exponential curve. Also, T.P. of Partially Impatient user for lower refreshing is lower than the Completely Impatient user but as the refreshing rate is increased the probability of having at page A of Partially Impatient user also increased.

For odd number of attempt and lower value of q_{13} T.P. of Partially Impatient user is growing with the refreshing probability(q_3) but T.P. of all others are very low approximately zero and not changing.

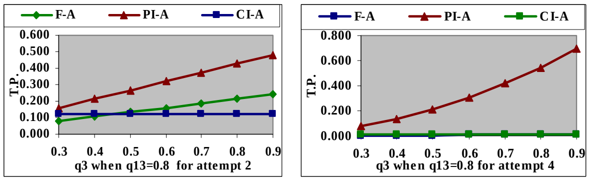


Fig 12(a) Fig 12(b)

Figure 12: Comparison between different users of page A for even attempt when q_{13} is high.

From fig 10(a) and 12(a) we can find that T.P. of Completely Impatient user will not be affected by q_{13}. Secondly the T.P. of Faithful user is lower than the T.P. of Completely Impatient user for any value of q_3 when q_{13} is low but higher value of q_{13}, T.P. of faithful user is initially lower than the Completely Impatient user but as the value of q3 is increased over 0.45 T.P. of faithful user crosses the T.P. of Completely Impatient user.

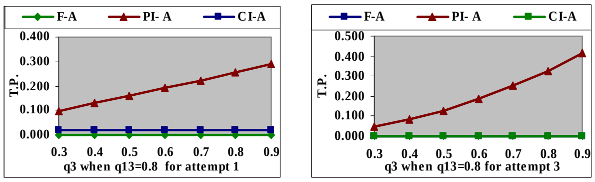


Fig 13(a) Fig 13(b)

Figure 13: Comparison between different users of page A for odd attempt when q_{13} is high.

As shown in the charts in fig 10(a), 10(b), 11(a), 11(b), 12(a), 12(b), 13(a) and 13(b) the T.P. is low for lower value of q_{13} and T.P. decreases with the attempt, but T.P. of Partially Impatient user is increased as the Refreshing probability increases.

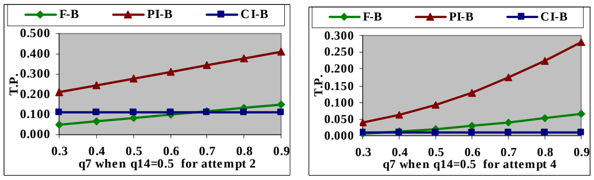


Fig 14(a) Fig 14(b)

Figure 5.14: Comparison between different users of page B for even attempt when q_{14} is low.

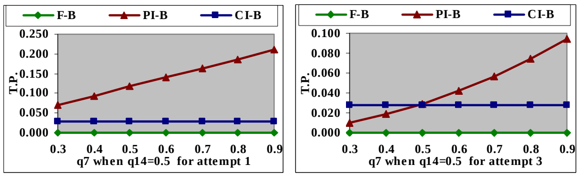


Fig 15(a) Fig 15(b)

Figure 15: Comparison between different users of page B for odd attempt when q_{14} is low.

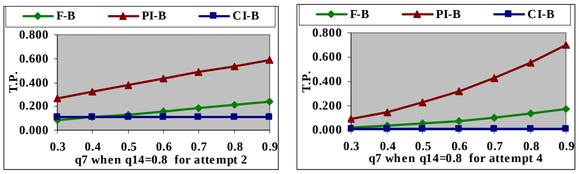


Fig 16(a) Fig 16(b)

Figure 16: Comparison between different users of page B for even attempt when q_{14} is high.

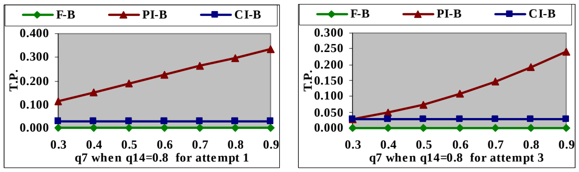


Fig 17(a) Fig 17(b)

Figure 17: Comparison between different users of page B for odd attempt when q_{14} is high.

The pattern for page B is similar to pattern for page A the only difference is that since here we taken the lower value of q_{14}=0.5 T.P. is higher. Since there is similarity in the probability calculating function of page A, B and C if we assign the same probability we can not get the difference in the T.P. hence to study the changes in the T.P. we have taken different probabilities. And the resultant changes can be seen from the charts. To study the behavior of different users of page B we used q_6=0.4 and q_{10}=0.6.

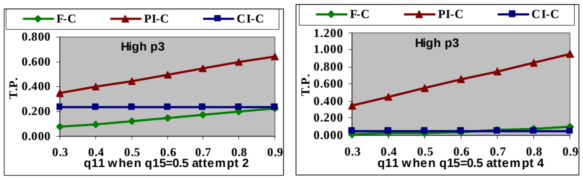


Fig 18(a) Fig 18(b)

Figure 18: Comparison between different users of page C for even attempt when initial probability p3 is high.

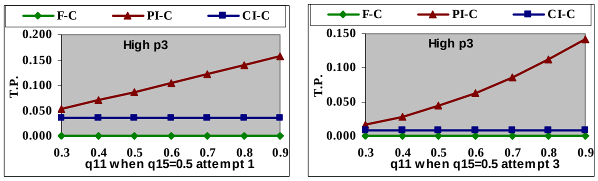


Fig 19(a) Fig 19(b)

Figure 19: Comparison between different users of page C for odd attempt when initial probability p3 is high.

When the initial probability of page C is high that is p_3=0.5 and p_2=p_1=0.25 then for odd attempt T. P. is very low and can not reach to 0.2 while on other hand for even attempt T.P. is reached to approx. 1. For any attempt T.P. of Partially Impatient users are higher than T.P. of other types of user. T.P. of Completely Impatient user is constant without affected by q_{11} and q_{15}. But T.P. is fluctuating (high for even number terms and low for odd number of terms). T. P. of faithful user is zero for odd number of terms and for even number of terms it is initially low and increased slowly with the value of q_{11}.

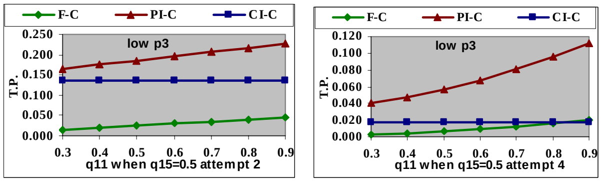


Fig 20(a) Fig 20(b)

Figure 20, Comparison between different users of page C for even attempt when initial probability p3 is low.

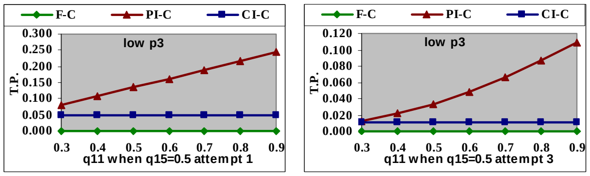


Fig 21(a) Fig 21(b)

Figure 21, Comparison between different users of page C for odd attempt when initial probability p3 is low.

For page C if its initial probability is low that is p_3=0.1, then the pattern found is same as found for p_3=0.5. For odd attempt there is not very much difference in the T.P. for low values of refreshing (q11) but as the refreshing rate increased the difference in the T.P. of different user expand. For odd number of attempt the T.P. of faithful user and T.P. of Completely Impatient user is constant with respect to refreshing. While for even number of attempt the T.P. of Faithful User is increasing with the refreshing but very slowly in comparison to the Partially Impatient user. Also, T.P. of all the users is decreasing with respect to attempt irrespective of even or odd attempt.

6. Conclusion and Future Directions

In this paper we focus on the problem of creating **adaptive web sites**: sites that automatically improve their organization and presentation by mining visitor access data collected in Web server logs. In our approach we used Markov chain model in which we treated the clicking of Refresh button as another state i.e. Refresh State. Based on this assumption we categorized all the users into three category faithful users, partially impatient users and completely impatient users. We proposed some new theorems for calculating the T.P. of different category of user (page wise) for odd number of attempt and for even number of attempt. Then using these theorems we did simulation study to prove the categorization. We studied the individual behavior with respect to attempt and refreshing. We also did comparative study of behavior of different types of user of each page. We find that-

1. Behavior of the faithful user is same regardless of web pages. That is probability of having at a page fluctuate, there is some probability (chances) to be in even number of terms but for odd number of terms probability is zero.
2. If the initial probability is low then behavior of Completely Impatient user is different that is reflected as constant declining graph while for higher initial probability Completely Impatient user behavior is different that is indicated by fluctuating graph.
3. The probability of having at a page for Partially Impatient user is fluctuating but in all declines towards zero for lower and medium probability. But for higher refreshing probability being on the page increase with the number of attempt.
4. For odd number of attempt and lower value of q_{13} T.P. of Partially Impatient user is growing with the refreshing probability(q_3) but T.P. of all others are very low approximately zero and not changing.
5. T.P. of Completely Impatient user of page A is not affected by refreshing. But T.P. of Completely impatient user decreases with respect to number of attempt.
6. The T.P. with respect to refreshing is increased exponentially for Partially Impatient user, constant for Completely Impatient user and slowly increases for faithful user. Hence the T.P. is approximately same for low refreshing but for high refreshing we can see grate difference in the T.P. of different type of user.

The different behavior proves the classification of users. It provides a good method of user classification that can be used for better behavior study. The designer of the web site can focus for any individual category of user.

The draw back of this approach is that if the refreshing rate will be very low then this categorization may be insignificant. Also this approach is basically statistical and approach can be useful only for high visitation sites. One another draw back is that if the web site has so many pages the dimension of the transition matrix will be very high that may cause presentation and memory problems. But this approach is novel and effective for many circumstances.

This approach is based on the assumption that each instance of refreshing can be identified. Thus some good algorithm should be developed so that either each attempt to refreshing a page can be recorded in the web log or by some means it can be identified from the web log.

ACKNOWLEDGEMENTS

Authors are grateful to the technical reviewers for the comments, which improved the clarity and presentation of the paper. Authors are thankful to their colleagues of Department Of Computer Science and Applications those help in many ways to complete the simulation study.

REFERENCES

[1] Su, Z., Yang Q., Lu Y., Zhang, H, *WhatNext: A Prediction System for Web Requests using N-gram Sequence Models*, Proc. of the International Conference on Web Information Systems Egineering (WISE2000), 2000.

[2] S. Karlin and H. M. Taylor. A First Course in Stochastic Processes, 2nd Ed. Academic Press, New York, 1975.

[3] R. Sarukkai, *Link prediction and path analysis using markov chains*, Proceedings of the 9th Int. WWW conference, 2000.

[4] I. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White. *Visualization of navigation patterns on a web site using model based clustering*. In Proc. of the 6th Int. KDD conference, pages 280-284, 2000.

[5] M. Eirinaki, M. Vazirgiannis, and D. Kapogiannis, *Web Path Recommendations Based on Page Ranking and Markov Models*, Proc. Seventh Ann. ACM Int'l Workshop Web Information and Data Management (WIDM '05), pp. 2-9, 2005.

[6] Zhu, J., *Using Markov Chains for Structural Link Prediction in Adaptive Web Sites*, Proc. of User Modeling, pp. 298-300, 2001.

[7] Zhu, J., Hong, J., and Hughes, J., *PageRate: Counting Web Users' Votes*, Proc. of ACM Hypertext'01, pp. 131-132, 2001.

[8] Zhu, J., Hong, J., and Hughes, J., *Using Markov Chains for Link Prediction in Adaptive Web Sites*, Proc. of Soft-Ware 2002: Computing in an Imperfect World, Springer-Verlag LNCS 2311, pp. 60-73, 2002.

[9] Kleinberg, J. M., Authoritative sources in a hyperlinked environment, Journal of ACM, 604-632, 1999.

[10] C. R. Anderson, P. Domingos, and D. S. Weld, *Relational markov models and their application to adaptive web navigation*, Proc. of the 8th Int. KDD conference, pages 143-152, 2002.

[11] M. Deshpande and G. Karypis, Selective Markov Models for Predicting Web Page Accesses, ACM Trans. Internet Technology, vol. 4, pp. 163-184, May 2004.

[12] Albrecht, D.W., Zukerman, I., and Nicholson, A.E. *Pre-sending documents on the WWW: A comparative study*, Proc. of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI99), 1999.

[13] Pitkow J. and Pirolli P., *Mining Longest Repeating Subsequences to Predict WWW Surfing*, Proc. of the 1999 USENIX Annual Technical Conference, 1999.

[14] J. Borges and M. Levene, *Generating Dynamic Higher-Order Markov Models in Web Usage Mining,* Proc. Ninth European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD), pp. 34-45, Oct. 2005.

[15] J. Borges and M. Levene, Data Mining of User Navigation Patterns, Web Usage Analysis and User Profiling, pp. 92-111, Springer, 2000.

[16] J. Borges and M. Levene, *A Clustering-Based Approach for Modelling User Navigation with Increased Accuracy*, Proc. Second Int'l Workshop Knowledge Discovery from Data Streams, pp. 77-86, Oct. 2005.

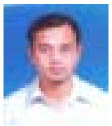
[17] J. Borges and M. Levene, Testing the Predictive Power of Variable History Web Usage, J. Soft Computing, special issue on Web intelligence, 2006.

[18] J. Borges and M. Levene, Evaluating Variable-Length Markov Chain Models for Analysis of User Web Navigation Sessions, 441 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19, APRIL 2007

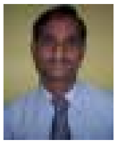
[19] Xing Dongshan and Shen Junyi, A new markov model for web access prediction, Computing In Science and Engineering, 4(6):34-39, 2002.

[20] X. Chen and X. Zhang, A Popularity-Based Prediction Model for Web Prefetching, 452 IEEE Transactions on Knowledge and Data Engineering, Vol 19, pp. 63-70, April 2007.

Authors Biography



Ratnesh Kumar Jain is currently a research scholar at Department of Computer Science and Applications, Dr. H. S. Gour Central University (formerly, Sagar University) Sagar, M P, India. He completed his bachelor's degree in Science (B. Sc.) with Electronics as special subject in 1998 and master's degree in computer applications (M.C.A.) in 2001 from the same University. His field of study is Operating System, Data Structures, Web mining, and Information retrieval. He has published more than 7 research papers and has authored a book.



Sanjay Thakur has completed M.C.A. and Ph.D. (CS) degree from H.S. Gour Central University, Sagar in 2002 and 2009 respectively. He is presently working as a Lecturer in the Department of Computer Science & Applications in the same University since 2007. He did his doctoral work in the field of Computer Networking and Internet traffic sharing. He has authored and co-authored 10 research papers in National/International journals and conference proceedings. His current research interest is Stochastic Modeling of Switching System of Computer Network and Internet Traffic Sharing Analysis.

R. S. Kasana completed his bacholar's degree in 1969 from Meerut University, Meerut, UP, India. He completed his master's degree in Science (M.Sc.-Physics) and master's degree in technology (M. Tech.-Applied Optics) from I.I.T. New Delhi, India. He completed his doctoral and post doctoral studies from Ujjain University in 1976 in Physics and from P. T. B. Braunschweig and Berlin, Germany & R.D. Univ. Jabalpur correspondingly. He is a senior Professor and HoD of Computer Science and Applications Department of Dr. H. S. Gour University, Sagar, M P, India. During his tenure he has worked as vice chancellor, Dean of Science Faculty, Chairman Board of studies. He has more than 34 years of experience in the field of academics and research. Twelve Ph. D. has awarded under his supervision and more than 110 research articles/papers has published.

Suresh Jain completed his bachelor's degree in civil engineering from Maulana Azad National Institute of Technology (MANIT) (formerly, Maulana Azad College of Technology) Bhopal, M.P., India in 1986. He completed his master's degree in computer engineering from S.G. Institute of Technology and Science, Indore in 1988, and doctoral studies (Ph.D. in computer science) from Devi Ahilya University, Indore. He is professor of Computer Engineering in Institute of Engineering & Technology (IET), Devi Ahilya University, Indore. He has experience of over 21 years in the field of academics and research. His field of study is grammatical inference, machine learning, web mining, and information retrieval. He has published more than 25 research papers and has authored a book.