CIRJE-F-194

A Two-Stage Prediction Model for Web Page Transition

Makoto Abe
The University of Tokyo

February 2003

A Two-Stage Prediction Model for Web Page Transition

Makoto Abe

Associate Professor of Marketing
Graduate School of Economics
University of Tokyo
Hongo, Bunkyo-ku
Tokyo. 113-0033 JAPAN

E-mail: abe@e.u-tokyo.ac.jp

TEL & FAX: 81-3-5841-5646

February 1, 2003

A Two-Stage Prediction Model for Web Page Transition

ABSTRACT

Utilizing data from a log file, a two-stage model for step-ahead web page prediction that permits adaptive page customization in real-time is proposed. The first stage predicts the next page of a viewer based on a variant of a Markov transition matrix computed from page sequences of other visitors who read the same pages as that viewer did thus far. The second stage re-analyzes the incorrect exit/continuation predictions of the first stage through data mining, incorporating the visitor's viewing behavior observed from the log file. The two-stage process takes advantage of a robust, theory-driven nature of statistical modeling for extracting the overall feature of the data, and a flexible, data-driven nature of data mining to capture any idiosyncrasies and complications unresolved in the first stage.

The empirical result with a test site implies that the first stage alone is sufficiently accurate (50.3%) in predicting page transitions. Prediction of site exit was even better with 100% of the exit and 90.8% of the continuation predictions being correct. The result was compared against other models for predictive accuracy.

Keywords: page transition, prediction model, log file, statistical modeling, Markov transition matrix, data mining

BIBLIOGRAPHICAL NOTES

Temporarily suppressed to hide the identity of the authors for double-blind review.

1. INTRODUCTION

For firms, large and small alike, internet has become a vital sales channel and a communication link with prospective customers. If a website is selling goods or services, the firm wants site visitors to order them. If the purpose of a website is communication, the firm wants visitors to request further information. A fraction of the visitors taking the intended action of the site is called a conversion rate. Effective personalized marketing can increase a conversion rate.

One of the advantages of the net over traditional channels is its ability for personalization in real-time at low cost. Many websites are open to public and do not require visitors to register or even identify themselves. When implementing personalization in such an open-site environment, the only source of information on an anonymous visitor is her page view history since her site entry up to that point, which is recorded in the log file. If one could predict her next page transition based on her page view observed thus far, it would become possible to take appropriate marketing actions in real-time. These actions can be in the form of customized page design or pop-up windows (adaptive page modification/generation). And a firm may offer personalized marketing incentives, such as special pricing and deals, bundling and tie-ins, prizes, and so forth, all of which aim to improve the intended conversion rate.

The objective of this research is to predict the page transition of visitors from the log file. In particular, given the viewing history of a visitor thus far, the proposed model predicts her next page (or site exit) stochastically using page view of other visitors accessing the site. The sole source of information used to predict page transition is a log file. The model does not require the page information on attributes such as page size, word-count, text vs. graphic proportion, and so forth, nor the visitor information on attributes such as demographic variables.

The paper is organized as follows. In Section 2, the relevant literature is reviewed first. To plan our strategy for modeling, an exploratory analysis is conducted on a test website in Section 3, which leads to the description of the proposed model in Section 4. Section 5 discusses the empirical result of our prediction model including the comparison with other benchmark models, followed by conclusions and future research in Section 6.

2. EXISTING LITERATURE

Because the field of predicting page transition is rather new, the relevant literature is quite limited. The author could not find any research specifically addressing step-ahead prediction of page transition using a log file.

Spiliopulou discusses a general approach to data mining of log files for site evaluation [1]. He raises several difficulties when analyzing page sequence of visitors recorded in a log file. Let us make brief notes on them here.

(1) Absence of user identification

Because the standard log file only contains the host name and the version of the user's web browser, exact identification of whether page requests came from the same user or not must be carried out by other means such as cookies, scripts, and authentication mechanism.

(2) Non-human visitors

Automated web robots for search engines often crawl through pages indiscriminately, leaving eccentric transitional patterns not expected from human visitors. The SurfAid project of IBM provides some solution to identify such visits by robots [2].

(3) Caching by browser

Clients' browsers usually cache previously visited pages to reduce network traffic, so that the page sequence in the log file may not coincide with the actual ordering of pages viewed. Cooley, Mobasher and Srivastava offer some solutions to this caching problem [3].

The first two issues did not apply to our study using an experimental site because authentication was required by all visitors, who had to log-in every time they entered the site. The third issue did not pose a problem either, because the proposed model required information only on which pages had been visited by a visitor thus far but not on their viewing order. A request for a new page is always recorded in a log file, and caching affects only the ordering information. In the actual implementation of the model, however, the above issues must be addressed with care since they can contaminate information of the log file in an unexpected manner.

Spiliopulou and Pohle analyze viewers' page sequence recorded in a log file to diagnose a site structure [4]. In their study, a website is regarded as a network of connected pages. Pages are first divided in terms of their functions into major groups, such as "action pages" and "target pages", and within each group, pages are further classified into similar categories according to the role they take. Using these categories rather than individual pages as the unit of analysis, they seek typical navigation patterns left by visitors in order to detect any abnormality in the page and site design using a data mining technique.

When mining for navigation rules, possible sequences of previously visited pages can increase exponentially as the path becomes longer and the site becomes larger. Their idea of introducing the page categories in navigational analysis certainly alleviates this combinatorial problem. Still, other problems arise. The classification of pages into categories requires human intervention, which is subjective as well as time consuming [5].

A major finding from this literature survey is that, past page sequence is the only means to differentiate visitors, and that this source of heterogeneity must be incorporated somehow in our step-ahead prediction model. Yet, if all possible combinations and the ordering of the past pages visited were to be considered, it would create a huge combinatorial problem.

3. EXPLORATORY ANALYSIS WITH A TEST WEBSITE

To obtain further insights into our modeling strategy, we conduct an exploratory analysis on a test website.

3.1 The Website

The test site is an actual e-commerce website describing LOTUS document management software for business. It discusses problems in today's office with overflowing papers and suggests this particular software as its solution. The site also explains five steps that should be followed for a successful introduction of the software. The ultimate objective of this site is to have the visitors click a "request further information" button, which brings them to a goal page soliciting personal information, so that the firm can send out the brochure and a sales person can contact them later. The structure of this website consisting of 16 pages is shown in Figure 1.

< Insert Figure 1 about here >

An arrow indicates the existence of a hyperlink that permits a direct jump to the page. In addition to these arrows, jumps in the reversal direction is always possible with a "back" button of a browser. The top page is P1, and the goal page is P16. As can be seen from Figure 1, P3, P10, and P15 contain the "request further information" button that navigates the visitors to P16. P6 through P10 provide a basic description for each of the five steps, respectively, for successful introduction, whereas P11 through P15 elaborate each of these steps in further detail.

The log file contains the total of 3,551 page views recorded from 464 sessions between December 26, 2000 and March 22, 2001. Of the 464 sessions, 150 sessions have reached P16, the goal page.

3.2 Markov Transition Matrix and Correlation Matrix

Useful descriptive statistics to understand page sequence are the Markov transition matrix and the correlation matrix of the viewed pages. The former depicts the viewing tendency of two consecutive pages, whereas the latter describes two pages that are likely to be viewed together within a single session but they need not be visited in consecutive order.

To describe the entry into and the exit from the website, two hypothetical pages, "enter" and "exit", were added, respectively. Table 1 shows the Markov transition matrix. The columns and rows correspond to origins and destinations, respectively. Lighter shade (yellow) indicates the existence of a hyperlink, whereas darker shade (blue) indicates that the jump is possible with a “back” button.

< Insert Table 1 about here >

The transition matrix provides useful hints on addition and removal of hyperlinks for improving the navigation. If no direct link exists yet the transition probability is sufficiently high, a web designer might consider adding a link for the sake of user convenience [6]. When a transition probability is low even if there exits a link, a web designer might consider removing the link to simplify the page layout and prevent unintended jumps by the visitors [7].

< Insert Table 2 about here >

The correlation matrix is shown in Table 2. For clarity, only statistically significant figures at the one-tail 1% level are reported. Let us focus our attention to the correlation with the goal page (P16). P3, P4, and pages describing individual steps for successful introduction (P7~P15) had significant correlation with the goal page. The only exception was P6. The general tendency was that, within the same hierarchical level (P6~P10 or P11~P15), correlation with goal increased as deeper steps were reached (i.e., P6<P7<P8<P9<P10 and P11<P12<P13<P14<P15).

Just from these two matrices alone, much insight into page transition and website diagnostics can be obtained. These matrices provide the two extreme prospects on the relationship between any pair of pages. The transition matrix depicts a navigational pattern in the unit of single transitions by decomposing a session into the shortest sequences --- transitions. In contrast, the correlation matrix does so in the unit of the longest sequence of a session --- a complete session consisting of multiple transitions. For more insight, however, we felt the need to have measures with the unit of analysis somewhere in-between.

3.3 Exploratory Analysis on Site Exit

Once an anonymous visitor leaves a site, the firm loses means to contact her for potential business, just like in a real store. Therefore, the most important page transition to understand and predict would be site exit. We investigated whether site exit was in any way related to page and viewer characteristics.

Page Characteristics

To characterize pages, we chose straightforward and objective measures of page attributes, shown in Table 3. Each of the 16 pages were evaluated according to these attributes, and various analyses were conducted to examine whether any of them is related to site exit to some degree.

< Insert Table 3 about here >

Unfortunately, we did not detect any meaningful relationship. It appeared that these page attributes were not sufficient to characterize site exit, which occurred as a result of the complex interaction between these objective page measures with page content iteself, website structure (e.g., relative location of pages and their accessibility within the site), as well as viewer behavior (e.g., time spent in the site). Due to this interaction, these page attributes alone played only a minor role in explaining the site exit. We, therefore, concluded that incorporating page characteristics would not be a fruitful direction, at least for building a transition model, from the following reasons.

(1) The attributes must be quantified by hand for each page, and if there are thousands of pages in a site, the process would be time-consuming and unpractical.
(2) There was a substantial amount of multicollinearity across these attributes, which must be dealt with by some means.
(3) If qualitative attributes such as page content were to be included, they would have to be quantified somehow, introducing certain subjectivity.

Viewer Characteristics

For general non-registered sites, visitors can enter freely without providing a positive identification. Hence, the only viewer heterogeneity that is manifested through a log file is a viewing history of each visitor, such as which pages are viewed, in what order, and for how long. Sessions were divided into two mutually exclusive groups, those reaching the goal and those exiting the site without reaching the goal. We then studied its relationship with the cumulative time spent in a session, the cumulative number of pages viewed in a session, and the average time spent per page, using various methods from simple descriptive statistics to statistical modeling like logistic regression. For the sake of space, only the major findings are reported below.

(1) The cumulative time spent in the session was not related to the probability of site exit.
(2) As more pages were viewed, the probability of reaching the goal increased [8].
(3) The probability of site exit was lowest when the average time spent per page took the medium values. Site exit was more prominent for either longer or shorter average time per page, exhibiting a complicated nonlinear relationship [9].

4. MODEL

Two requirements we seek in our page prediction model are (1) it can be calibrated automatically only from data contained in a log file, and (2) the viewer heterogeneity is accounted for. The first requirement implies that page attributes such as those in Table 3 are not included because coding of individual pages is necessary. Their poor explanatory power found in the exploratory analysis also supports this notion. Regarding the second requirement, the viewer heterogeneity must be captured through the difference in past page sequence observed thus far as well as viewing behavior such as cumulative time spent in the site and average time per page.

The proposed model consists of two stages. The first stage uses a statistical model based on a variant of a Markov process by accounting for viewer difference in the past page sequence for stochastic page prediction. The second stage employs data mining to re-analyze the incorrect exit/continuation predictions of the first stage by incorporating that visitor's viewing behavior extracted from a log file. In the second stage, transitions that are not predicted correctly in the first stage, i.e., the residuals from the statistical model, are analyzed by data mining that incorporates viewing behavior variables to further improve the prediction. The two-stage process takes advantage of a robust, theory-driven nature of statistical modeling for extracting the overall feature of the data, and a flexible, data-driven nature of data mining to capture any idiosyncrasies and complications unresolved in the first stage. Cooper and Giuffrida successfully apply such a combined approach of statistical modeling and data mining to the forecasting of market shares using POS data [10].

4.1 The First Stage: Statistical Modeling

An elaborate real-time step-ahead prediction should depend not only on the current location of a visitor but also on the past sequence arriving at the current location. The transition matrix that accounts for the current location alone does not suffice [11]. In addition to the current location, however, considering even just the previous page location increases the size of the matrix substantially [12]. If rules for predicting the next page transition were constructed from the entire history of the past pages visited, their computation would become unmanageable as there exit countless combinations of possible past sequences. To circumvent this problem, a viewer's next page is predicted by a transition matrix that is computed from data using only those viewers who visited the same pages as this viewer did thus far. Because this transition matrix is updated every time a viewer makes a transition to a new location, we call it the Baysian Conditional Markov Transition (BCMT) matrix.

Let us explain the BCMT matrix with an example. Suppose Linda entered a site from P3 and visited P5 and P6 thereafter. To predict Linda's next page, we extract page sequences of all viewers who visited P3, P5 and P6 from the log file and compute the transition matrix. A probabilistic prediction of her next page is obtained directly from this matrix, and the deterministic prediction would be the page with the highest transition probability. Suppose its deterministic prediction was P7, but Linda jumped to P10. This time, a transition matrix is computed from page sequences of all viewers who visited P3, P5, P6, and P10. The process is repeated as Linda moves through the site, each time with a new transition matrix [13].

The concept of the BCMT matrix is similar to that of collaborative filtering, which is popular in recommendation systems. In collaborative filtering, data on customers who purchased the same products as the person in question (let's call her Linda again) are extracted, their purchase patterns are analyzed, and products Linda has not bought yet are recommended. In general, attributes on customers and products are not considered in collaborative filtering. In our approach, next page transition is predicted using data on viewers who had visited the same pages as Linda did thus far, without considering viewer and page attributes.

4.2 The Second Stage: Data Mining

In the first stage, the probability of site exit was predicted from the transition matrix by adding a hypothetical page "exit". The second stage focuses on site exit and improves its prediction by incorporating viewing behavior. The exploratory analysis on site exit hinted complicated nonlinear and interaction effects of viewing behavior that was unlikely to be uncovered by traditional linear-based statistical models. We, therefore, employ a data mining technique in this stage.

Site exit at each page view is predicted by incorporating the exit probability computed in the first stage as well as viewing behavior variables. The response variable is binary, indicating the site exit (1 for exit, 0 for continuation). The explanatory variables are exit probability, cumulative time spent, cumulative number of pages, and average time spent per page. For binary prediction, neural net and decision tree are two popular data mining techniques [14]. While both techniques can uncover the underlying nonlinearity and interaction effects of explanatory variables hidden in data, the former is used when only forecasting is needed but not its interpretation. The latter has an advantage in that the classification rules can be made explicit, which is an important feature when we want to know the reason.

Decision tree does the same thing as cluster analysis. A sample is divided into finer, more homogeneous groups that exhibit the similar response pattern for explanatory variables. The division is carried out one explanatory variable at a time in a sequential fashion, such that a variable with the most discriminating power for the binary response is applied first.

Depending on the division strategy and the measure of group homogeneity, several algorithms exist. For ease of interpretation, we adopt the most popular CART (classification and regression tree) algorithm that divides data into two clusters at each step using the Gini's diversity index measure [14].

5. RESULTS

5.1 The First Stage: Bayesian Conditional Markov Transition (BCMT) Matrix

5.1.1 Page Transition

In step-ahead prediction, there is no need to divide data into a calibration and validation samples. □e can make full use of the data for the maximal predictive accuracy. Table 4 compares the actual page visited (or site exit) and predicted page (or exit) for each of the 3,551 page views. The page with the highest probability was used as the predicted page in this result. Overall, 50.3% (1,787 pages) were predicted correctly. The mean probability of the actual pages jumped was 0.3056 [15].

< Insert Table 4 about here >

For comparison, the prediction was repeated with the homogeneous Markov transition matrix that did not account for the difference in past page sequence across visitors. In particular, step-ahead page transitions of each visitor were predicted with a transition matrix that was computed from the page sequences of all other viewers without updating [16]. This naïve model resulted in the overall prediction accuracy of 44.9% (1,593 pages) and the mean probability of the actual pages of 0.2871. The result implies the importance and the power of accounting for viewer heterogeneity by computing a transition matrix from the page sequences of other viewers visiting the same pages and updating as the visitor makes a transition.

5.1.2 Site Exit Prediction

Let us now focus our attention to the exit prediction result shown in Table 5. The rows and columns represent the actual and predicted transitions, respectively. 100% (150/150) of the exit predictions and 90.8% (3087/3401) of the continuation predictions were correct, and the overall accuracy was 91.2% (3237/3551). The other way to state the result is that 100% of the actual continuations and 32.3% of the actual exits were identified correctly, with the overall accuracy of 91.2%. In a real setting, perhaps the former interpretation is more relevant because we only have prediction rather than the actual state of exit or continuation.

< Insert Table 5 about here >

5.1.3 Comparison of Accuracy for Exit Prediction with Other Methods

We now compare this result regarding site exit with other prediction models. Three naïve models were evaluated.

(a) Logistic Regression
The dependent variable was the same exit/continuation indicator, and the explanatory variables were selected by the stepwise method from the same set of the page attributes shown in Table 3 and the viewing behavior variables described in the previous exploratory analysis (cumulative time spent in the site, average time per page, cumulative number of pages viewed thus far). Due to lack of statistical power (degrees of freedom), no holdout sample could be set aside for validation. Because we used all the data for calibration, the result reflected the best case scenario, thereby the prediction accuracy represented the upper bound.

(b) Decision Tree Model
The model classified each page view into either continuation or exit, incorporating the same page attributes and viewing behavior variables as above. The model also searched for the optimal data allocation between calibration and validation samples for predictive performance, which resulted in the 90% and 10% proportion.

(c) Logistic Regression followed by its Residuals Re-analyzed by Decision Tree
This approach combined the above statistical method and the data mining method together. Site exit was first predicted by logistic regression as in (a). Then the incorrect predictions were re-analyzed by the decision tree model of (b) for any nonlinear and interaction effects. As with the logistic regression of (a), all data had to be used for calibration. Because the result reflected the best case scenario, the actual performance would likely to be much worse.

< Insert Table 6 about here >

Table 6 summarizes the accuracy rate of exit and continuation predictions by each method. The proposed model out-performed the other three methods, especially for exit prediction, which is crucial in real application. To illustrate the point, when site exit is predicted, intensive marketing (for example, offering a large discount) to prevent this exit would not be wasted at all since all viewers would exit the site without such incentives (because of the 100% prediction accuracy). If continuation is predicted with the proposed model, 9.2% would still exit the site. Though this is better than (a) or (b) and comparable to (c), there is some room for improvement here.

5.2 The Second Stage: Decision Tree

In the second stage, we tried to improve the performance of the BCMT matrix by incorporating additional information from the visitors' viewing behavior. In particular, we wanted to improve the present accuracy of 90.8% for the continuation prediction made on the 3,401 page views. The dependent variable was whether site exit occurred or not, and the explanatory variables were the three viewing behavior variables, cumulative time spent in site, average time per page, and cumulative number of pages viewed thus far, in addition to the exit probability predicted in the first stage. A decision tree software called AnswerTree by SPSS was used. As a first cut, we applied all the data for calibration. The predictive performance is shown in Table 7.

< Insert Table 7 about here >

Although the accuracy rate had improved, its increase from 90.8% to 91.2% (3103/3401) was only minor. Figure 2 shows the estimated decision tree. The derived rules suggested a site exit if the cumulative time (DECTIME) was greater than 2.5 seconds and the probability of exit predicted in the first stage (ERATE) was between 0.040 and 0.048. Because
(1) the prediction performance improved only slightly even though all the data were used for calibration,
(2) these rules had to be validated by a holdout sample, which would likely to depress the predictive performance even further,
(3) the extracted rules were rather unintuitive,
we could safely conclude that the advantage of introducing the second stage was small

< Insert Figure 2 about here >

6. CONCLUSIONS

In this research, a prediction model for step-ahead page transition using data from a log file was proposed. The model had two stages. The first stage predicted the next page of a viewer based on a variant of a Markov transition matrix computed from page sequences of other visitors who had read the same pages as that viewer did thus far. The second stage re-analyzed the incorrect exit/continuation predictions of the first stage through data mining, incorporating the visitor's viewing behavior observed from the log file. The two-stage process takes advantage of a robust, theory-driven nature of statistical modeling for extracting the overall feature of the data, and a flexible, data-driven nature of data mining to capture any idiosyncrasies and complications unresolved in the first stage.

The empirical result with a test site implied that the first stage alone was sufficiently accurate in predicting transitions, and the second stage did not improve the exit/continuation prediction.

With the first stage alone, correct pages were predicted for 50.3% of the transitions, whereas the naïve model that used a homogeneous transition matrix resulted in only 44.9% accuracy. The performance was even better with the exit/continuation prediction. 100% of the exit predictions and 90.8% of the continuation predictions were correct, resulting in the overall accuracy of 91.2%. The benchmark models using logistic regression and decision tree performed much worse. These results implied the importance and the power of accounting for viewer heterogeneity in predicting page transition.

Although viewer behavior information was not found to be related to page transition and site exit in this test site consisting of 16 pages, it might well be the case in other websites. Hence, it is always a good idea to have the option of pursuing the second stage improvement.

Future research should apply this model to larger websites to explore its practicality. When a site is much larger than this empirical site, computing a transition matrix for hundreds and thousands of pages becomes unrealistic. One way to overcome this difficulty is to focus only key pages of interest. Another is to group similar pages into categories to form "conceptual pages", which become the unit of analysis. What would be the realistic and practical strategy in dealing with a large site? Addressing the question would surely bring marketers a step closer to the practical implementation of adaptive page operation that could raise the conversion efficiency.

ACKNOWLEDGEMENTS

Temporarily suppressed to hide the identity of the authors for double-blind review.

REFRENCES AND NOTES

[1] Spiliopoulou, Myra (2000), "Web Usage Mining for Web Site Evaluation: Making a site better fit its users," Communications of the ACM, 43 (8), 127-134.

[2] www.surfaid.dfw.ibm.com

[3] Cooley, Robert, Bamshad Mobasher and Jaideep Srivastava (1999), "Data Preparation for Mining World Wide Web Browsing Patterns," International Journal of Knowledge and Information Systems (Springer), 1 (1), 5-32.

[4] Spiliopoulou, Myra and Carsten Pohle (2001), "Data Mining for Measuring and Improving

the Success of Web Sites," Data Mining and Knowledge Discovery (Kluwer), 5 (1, 2), 85-114.

[5] Buchner, A. G. and M. D. Mulvenna (1998), "Discovering Internet Marketing intelligence through Online Analytical Web Usage Mining," ACM SIGMOD Record, 27 (4), 54-61.

[6] For example, transition P5⇒P3 and P6⇒P5 are candidates for adding a link because they both have a probability higher than 0.10. An explicit link may not be necessary, however, since these transitions can be invoked by a "back" button.

[7] For example, links for P5⇒P7, P8, P9 and P6⇒P8, P9, P10 and P7⇒P6, P9, P10 and P8⇒P6, P7, P10 and P9⇒P6, P7, P8 and P10⇒P6, P7, P8, P9 should be removed since their probabilities are lower than 0.10.

[8] For example, of the three pages (P3, P10, P15) that had a direct link to P16, the probability of reaching the goal was only 0.28 if viewers visited only P3 but neither P10 nor P15, and 0.39 if both P3 and P10 but not P15 were visited, and 0.63 if all three pages were visited. This observation was also consistent with the fact that correlation with goal increased as deeper steps were reached within the same hierarchical level.

[9] Viewers with short average time per page probably clicked away pages because they are not particularly interested in the site. Abnormally large average time per page was found to be caused by a long state of inactivity during the session, perhaps visiting other sites through another browser window or doing other things than web surfing while on line. Both of these cases represented a lack of interest and hence resulted in the higher chance of site exit.

[10] Cooper, Lee G. and Giovanni Giuffrida (2000), "Turning Datamining into a Management Science Tool: New Algorithms and Empirical Results," Management Science, 46 (2), 249-264.

[11] Consider two viewers currently reading the same page, but one who visited each of the pages in intended order and another who skipped most pages but visited only vital pages. Obviously, the prediction for the next page should be different between the two viewers. The next page depends not only on the current page alone. The entire past page sequence reaching the current page contains the important information on where she might jump next.

[12] Suppose a web site contains N pages. Then the size of the transition matrix increases

from N×N to N×N^2, an N times more.

[13] One difficulty with this approach is that, as Linda's sequence becomes longer, there would be fewer and fewer viewers who visited the same pages as she did. In the worst case, there remains no viewer to construct the transition matrix from, and no prediction can be made. The solution we used in the current study was to use the same transition matrix as the previous prediction by not updating the BCMT matrix when the number of other viewers were less than a predetermined value of five. Other possible solutions are (1) when extracting viewers on the basis of the past pages visited, focus only on key pages rather than all pages, (2) classifying pages into similar categories to reduce the number (for example, descriptions of different brands at the same level of detail can be grouped together). Their advantage is that researchers can limit their analysis to key pages that play important roles in the site. The obvious disadvantage is that human intervention is necessary for page classification and certain subjectivity is introduced in the process.

[14] Berry, Michael J. A. and Gordon Linoff (1997), Data Mining Techniques: For Marketing, Sales, and Customer Support, New York: Wiley.

[15] The mean probability of actual pages is defined as the average of predicted transition probabilities for actual pages jumped to. The higher this measure is, the more accurate the prediction becomes.

[16] We had to go through this tedious step for a model validation purpose because we should not predict the transitions of a visitor from data that included her actual page sequence.

Table 1: Transition Matrix

The columns and rows correspond to origins and destinations, respectively. Lighter shade (yellow) indicates the existence of a hyperlink whereas darker shade (blue) indicates that the jump is possible using a “back” button of a browser.

to/ from	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	enter
1	0	0.262248	0.060921	0.005464	0	0	0.005348	0	0	0	0	0	0.028571	0	0	0	0.915948
2	0.516014	0	0.05052	0	0.002551	0	0	0	0	0	0	0	0	0	0	0	0.047414
3	0.379004	0.622478	0	0.846995	0.186224	0	0	0	0	0	0	0	0	0	0	0	0.032328
4	0.005338	0	0.264487	0	0.005102	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0.444279	0.038251	0	0.167315	0.042781	0.063584	0.057692	0.068966	0	0	0	0	0	0	0.002155
6	0	0	0.001486	0.010929	0.489796	0	0.085561	0.011561	0	0.034483	0.35	0.022727	0	0	0	0	0.002155
7	0	0	0	0	0.045918	0.33463	0	0.080925	0.012821	0	0.57	0.227273	0	0	0	0	0
8	0	0	0	0	0.033163	0.011673	0.524064	0	0.115385	0	0	0.727273	0.257143	0	0	0	0
9	0	0	0	0	0.030612	0	0.010695	0.549133	0	0.08867	0	0	0.685714	0.178571	0	0	0
10	0	0	0	0	0.135204	0.007782	0.02139	0.011561	0.615385	0	0	0	0	0.821429	0.359375	0	0
11	0.001779	0	0	0	0	0.365759	0.02139	0	0	0.004926	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0.219251	0.017341	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0.196532	0.00641	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0.173077	0.004926	0	0	0	0	0	0	0
15	0	0	0	0.005464	0	0	0	0	0.00641	0.305419	0	0	0	0	0	0	0
16	0	0.002882	0.104012	0.005464	0	0	0	0	0	0.236453	0	0	0	0	0.46875	0	0
enter	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
exit	0.097865	0.112392	0.074294	0.087432	0.071429	0.11284	0.069519	0.069364	0.012821	0.256158	0.08	0.022727	0.028571	0	0.171875	1	0

Table 2: Correlation Matrix

Only statistically significant figures at the one-tail 1% level are reported. Inspection of correlation with the goal page, P16, indicates pages that are visited for goal reaching sessions. Shade indicates pages describing the five steps of a successful introduction briefly (P6~P10) and in detail (P11~P15).

	1	2	3	4	5	6	7	8	9	0	11	12	13	14	15	16
1	1.000															
2	.	1.000														
3	.	- 0.220	1.000													
4	.	0.175	0.241	1.000												
5	.	.	0.429	0.185	1.000											
6	.	0.128	0.267	0.277	0.624	1.000										
7	.	.	0.215	0.349	0.499	0.707	1.000									
8	.	0.124	0.214	0.355	0.471	0.644	0.846	1.000								
9	.	0.144	0.201	0.376	0.453	0.598	0.781	0.879	1.000							
10	.	.	0.240	0.257	0.564	0.489	0.629	0.692	0.751	1.000						
11	.	0.129	0.162	0.349	0.352	0.564	0.550	0.480	0.457	0.363	1.000					
12	.	0.169	.	0.234	0.220	0.315	0.424	0.440	0.359	0.268	0.463	1.000				
13	.	0.142	.	0.240	0.200	0.298	0.348	0.416	0.402	0.313	0.440	0.686	1.000			
14	.	0.151	.	0.231	0.178	0.282	0.323	0.371	0.393	0.311	0.417	0.670	0.854	1.000		
15	.	.	0.124	0.255	0.266	0.295	0.413	0.437	0.454	0.488	0.351	0.350	0.478	0.463	1.000	
16	.	.	0.214	0.163	.	.	0.199	0.223	0.276	0.233	.	0.171	0.221	0.202	0.247	1.000

Table 3: Objective Measures of Page Characteristics

These page characteristics are investigated for any influence on site exit. It appears that these page attributes are not sufficient to characterize site exit, which occurs as a result of the complex interaction between these objective page measures with page content itself, website structure, as well as viewer behavior.

attribute category	attributes
file size	size of the page (bytes)
linked files	number of html files linked to the page size of the linked text files (bytes) number of the linked graphic files size of the linked graphic files (bytes) number of other files linked to the page size of the linked other files (bytes)
text in a page	text area of the page (mm^2) number of key words
graphics in a page	graphical area of the page (mm^2) proportion of the text area to the graphic area

Table 4: Actual and Predicted Page Transition

The rows and columns represent actual and predicted page jumps, respectively. The diagonal figures imply the number of transitions predicted correctly.

actual	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	EXIT
1	0	0	92	12	29	0	0	1	1	0	0	0	0	0	0	0	0
2	0	290	0	20	14	1	0	0	0	0	0	0	0	0	0	0	0
3	0	213	371	0	0	73	0	0	0	0	0	0	0	0	0	0	0
4	0	3	0	10	168	2	0	0	0	0	0	0	0	0	0	0	0
5	0	0	7	128	171	0	1	6	5	5	42	2	6	4	14	0	0
6	0	0	2	1	0	192	35	11	2	0	0	6	0	0	7	0	0
7	0	0	0	0	0	18	57	10	7	1	86	0	7	1	0	0	0
8	0	0	0	0	0	13	0	96	9	12	3	34	0	6	0	0	0
9	0	0	0	0	0	12	0	1	90	5	0	1	29	0	18	0	0
10	0	0	0	0	0	53	0	4	1	91	2	0	1	28	0	23	0
11	0	1	0	0	0	0	0	0	0	0	94	4	0	0	1	0	0
12	0	0	0	0	0	0	0	9	0	0	0	32	3	0	0	0	0
13	0	0	0	0	0	0	0	0	8	0	0	0	26	1	0	0	0
14	0	0	0	0	0	0	0	0	0	2	0	0	0	25	1	0	0
15	0	0	1	0	0	0	0	0	0	0	0	0	0	1	62	0	0
16	0	0	2	27	43	0	0	0	0	0	0	0	0	0	48	30	0
EXIT	0	53	55	8	41	28	8	9	9	0	29	5	4	2	52	11	150
total	0	560	530	206	466	392	101	147	132	116	256	84	76	68	203	64	150

Table 5: Actual and Predicted Site Exit

The rows and columns represent the actual and predicted transitions, respectively. 90.8% (3087/3401) of the continuation predictions and 100% (150/150) of the exit predictions are correct with the overall accuracy of 91.2% (3237/3551).

		Predicted		**Total**
		continuation	**exit**	
Actual	**continuation**	3,087	0	**3087**
	exit	314	150	**464**
Total		**3,401**	**150**	**3,551**

Table 6: Comparison of Predictive Accuracy for Site Exit with Other Models

The proposed model of the first stage outperforms the other three methods, especially for exit prediction, which is crucial in real application.

	Data	**Prediction Accuracy (%)**		
		Total	**Exit**	**Continuation**
proposed model	**prediction**	91	100	91
logistic regression	**calibration**	88	58	88
decision tree	**prediction**	88	53	90
logistic regression and decision tree	**calibration**	92	95	92

Table 7: Actual and Predicted Site Exit of the Second Stage

The second stage re-analyzes the 3,401 continuation predictions made in the first stage to improve their prediction by incorporating viewer behavior using a decision tree technique. Although the accuracy rate is improved, its increase from 90.8% to 91.2% (3103/3401) is only minor.

		Predicted		Total
		continuation	exit	
Actual	continuation	3067	20	3087
	exit	278	36	314
Total		3345	56	3401

Figure 1: Structure of the Experimental Web Site

An arrow indicates the existence of a hyperlink that permits a direct jump to the page. The top page is P1, and the goal page is P16. P3, P10, and P15 contain a “request further information” button that navigates the visitors to P16. P6 through P10 provide a basic description for each of the five steps, respectively, for successful introduction, whereas P11 through P15 elaborate each of these steps in further detail.

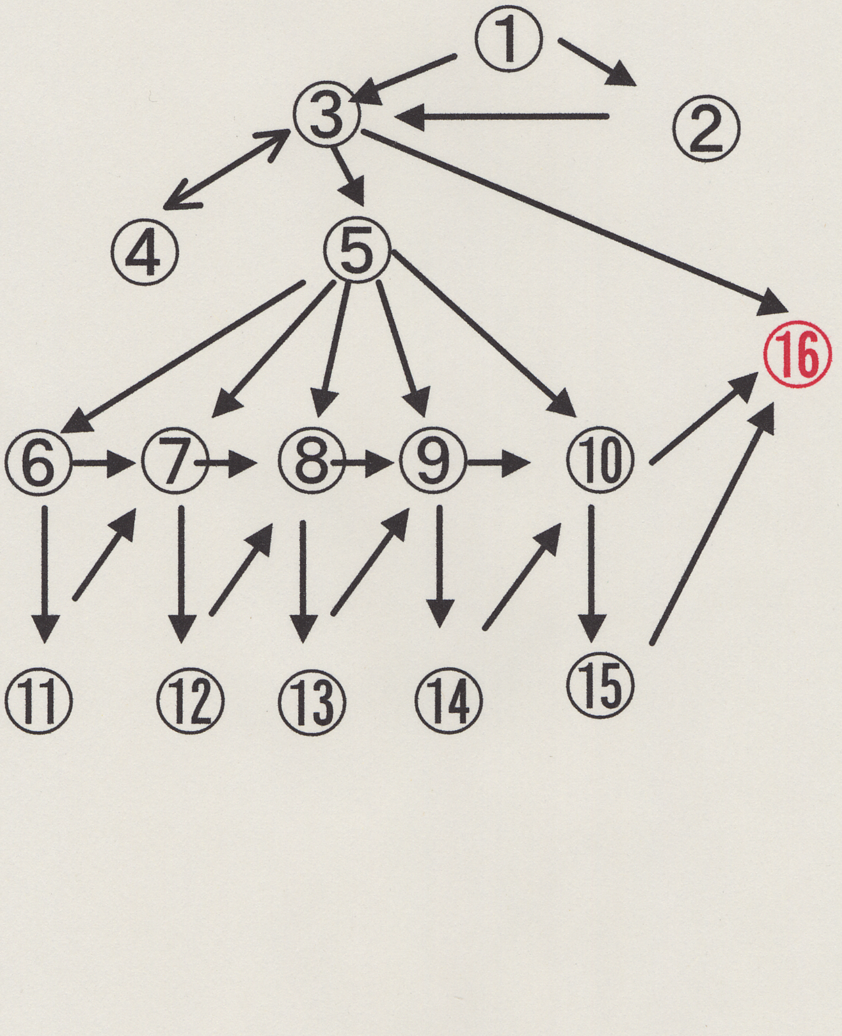
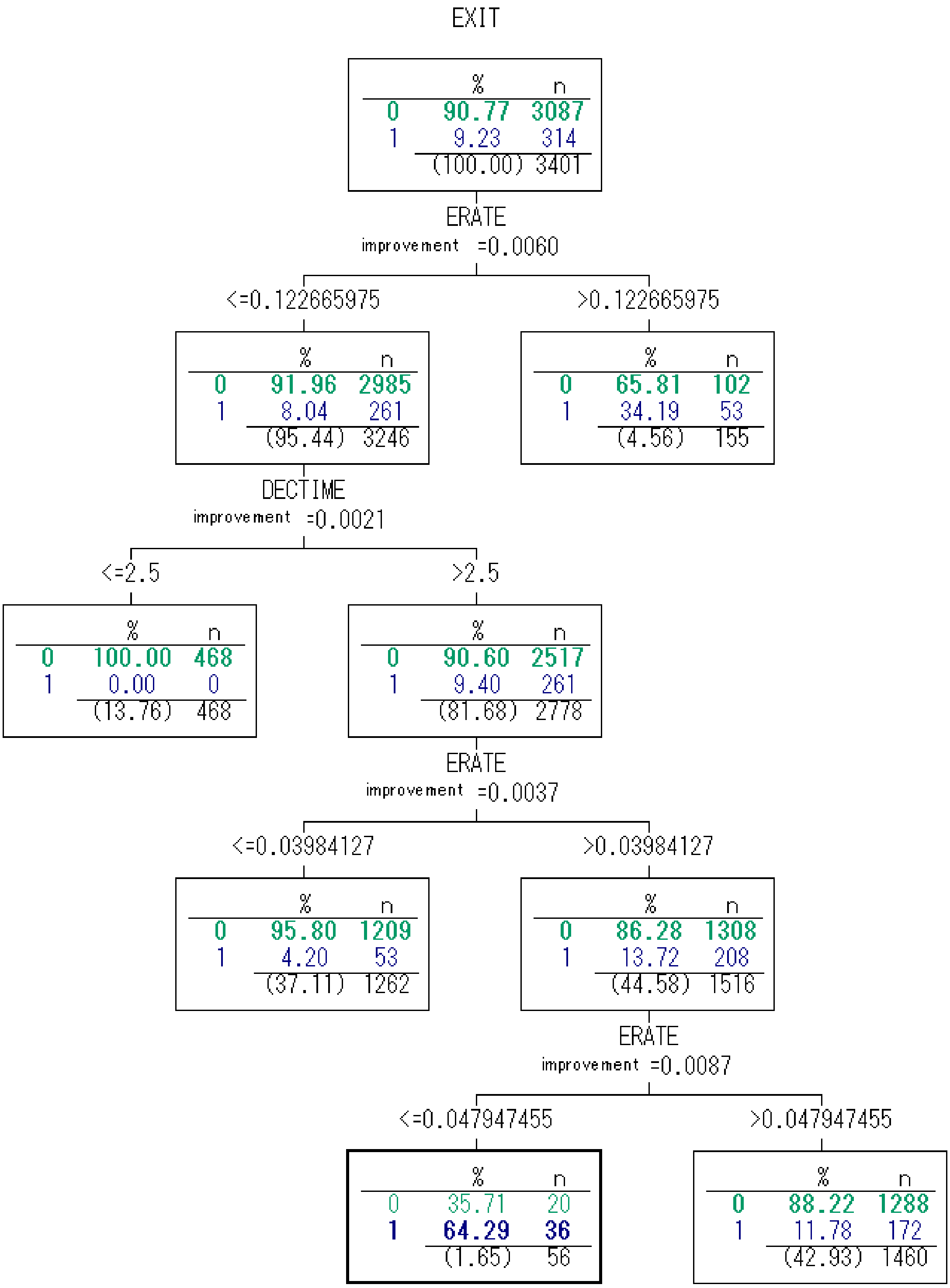


Figure 2: Decision Tree of the Second Stage

The rules by decision tree suggests a site exit if the cumulative time (DECTIME) is greater than 2.5 seconds and the probability of exit predicted in the first stage (ERATE) is between 0.040 and 0.048, which is rather unintuitive.


EXIT
% n
0 90.77 3087
1 9.23 314
(100.00) 3401
ERATE
improvement =0.0060
<=0.122665975
% n
0 91.96 2985
1 8.04 261
(95.44) 3246
>0.122665975
% n
0 65.81 102
1 34.19 53
(4.56) 155
DECTIME
improvement =0.0021
<=2.5
% n
0 100.00 468
1 0.00 0
(13.76) 468
>2.5
% n
0 90.60 2517
1 9.40 261
(81.68) 2778
ERATE
improvement =0.0037
<=0.03984127
% n
0 95.80 1209
1 4.20 53
(37.11) 1262
>0.03984127
% n
0 86.28 1308
1 13.72 208
(44.58) 1516
ERATE
improvement =0.0087
<=0.047947455
% n
0 35.71 20
1 64.29 36
(1.65) 56
>0.047947455
% n
0 88.22 1288
1 11.78 172
(42.93) 1460