Fall 9-30-2008

Dynamic Prediction of retail Website Visitors' Intentions

Pawel Kalczynski

Sylvain Senecal, Ph.D.

Marc Fredette

Dynamic Prediction of Retail Website Visitors' Intentions

Pawel Kalczynski, CSU, Fullerton, USA, pkalczynski@fullerton.edu
Sylvain Senecal, Marc Fredette, HEC, Montreal, Canada, {sylvain.senecal, marc.fredette}@hec.ca

Abstract

This paper presents a model for identifying general intentions of consumers visiting a retail website. When visiting a transactional website, consumers have various intentions such as browsing (i.e., no purchase intention), purchasing a product in the near future, or purchasing a particular product during their current visit. By predicting these intentions early in the visit, online merchants could personalize their offer to better fulfill the needs of consumers. We propose a simple model which enables classifying visitors according to their intentions after only four traversals (clicks). The model is based solely on navigation patterns which can be automatically extracted from clickstream. The results are presented and extensions of the model are proposed.

Keywords: Clickstream analysis, consumer behavior, e-commerce

1. Introduction

The classical consumer decision-making process suggests that consumers go through a series of steps (i.e., problem recognition, information search, evaluation of alternatives, intention, purchase, and post-purchase behavior [4]) while making consumption decisions. Furthermore, consumer behavior research suggests that, even during a particular decision-making step, consumers may have various intentions. For instance, Bloch, Sherrell, and Ridgway [1] suggest that during the information search step consumers perform pre-purchase (for the current purchase decision) search or ongoing search activities (for future purchase decisions). Thus, consumers who visit a retailer's website may not be at the same level of purchase readiness and may have different intentions when entering this website. For instance, Moe [9] suggests that visitors can be categorized into various intention groups based on their clickstream. In this paper we emphasize the importance of dynamic identification and following the evolution of consumers' intentions in order to provide personalized content which, in turn, might improve website effectiveness.

It is widely reported that most consumers visiting a transactional website do not complete a purchase during their visit [14]. As a result, website conversion rates (percentage of visitors who make a purchase) are very low, cart abandonment is frequent, and many consumers are dissatisfied with transactional websites [8]. If one could help retailers identify the intentions of visitors before they abandon the site, the websites could personalize their offers and consumers would be better served which, in turn, would improve customer satisfaction and conversion rates.

Marketing researchers are increasingly using clickstream (web usage) data, which was originally collected for website performance analyses. Clickstream data has been used to investigate consumer behaviors across websites (user-centric scenarios) [5] and within specific websites (site-centric scenarios) [14]. In the latter category, some studies focused on single visits to a given website [8], some dealt with multiple visits [3], while others investigated visits of both types [11]. Researchers engaged in this type of work have traditionally focused on such issues as: (1) clickstream or website-related variables that identify the goals which consumers are pursuing while they are navigating [11], (2) information search and usage [7], (3) the question of why consumers continue navigating through a website [3], (4) on-line decision-making processes [13], and (5) identifying which visitors are likely to make a purchase [8].

In marketing and e-commerce literature, clickstream analysis typically employs content-independent data, i.e., no information about the content of pages visited is analyzed. However, some studies included general types of pages consulted by consumers. For instance, it was found that consumers who enter a website with a clear purchase

intention visit fewer product category pages and more often repeat viewings of product pages than consumers with no clear purchase intention [10].

Others, in an effort to combine clickstream with rich content-dependent website data, have used concurrent verbal protocols [12]; consumers were asked to verbalize their thoughts while performing online tasks. However, as most qualitative research methods, verbal protocol analysis is very resource-consuming. Because of that, small sample sizes have been used and this may have limited the generalizability of the findings.

The proposed model distinguishes itself from previous approaches in the following two ways: (1) it provides results early in the session, thus enabling taking appropriate actions before the visitor abandons the website and (2) no additional knowledge about the consumer, other than the navigational pattern automatically extracted from clickstream, is assumed. These two properties enable practical applications of the proposed model to sessions which can not be handled by traditional recommender agents, in particular, when the identification of the visitor is not possible (e.g. the visitor did not log on) or when the identified customer exhibits atypical behavior (e.g. he or she is shopping for someone else). In addition, the proposed model contributes to the theory of e-commerce by demonstrating how much information can be extracted from the content-independent weblog without employing sophisticated and resource-consuming systems.

2. Methodology

2.1 Data Collection

The user data for this research project were collected in a laboratory setting at a major North-American university. The subjects were recruited from a group of actual consumers who responded to the invitation to participate in the study. The experiment required the participants to visit the particular website (an on-line music retailer). All participants were given an electronic gift certificate redeemable on this website in exchange for their participation.

The following data were collected for each individual traversal (click): session (user) id, time stamp, visited node (web page). The collected clickstream dataset consists of 138 sessions (2,798 traversals) performed on the website by 138 different users. Only 13 participants completed the purchase. The sessions were divided into two separate tasks:

TASK A (46 sessions) was a browsing task; the participants were asked to familiarize themselves with the website and were given an electronic gift certificate once they completed this task.

TASK B (96 sessions) was a shopping task; the participants were asked to shop for a CD and buy it with the electronic gift certificate if they found what they wanted.

The participants were surveyed before and after the experiment. The shopping sessions were recorded and played back to the subjects in order to give them an opportunity to explain their choices and evaluate the outcomes of the navigation process. The explanations and evaluation data were linked with clickstream data.

The average number of clicks was about 17 per session for TASK A (browsing) and about 22 for TASK B (shopping); the number of clicks per session ranged from 4 to 105. Figure 1 shows the relative frequency distributions of the number of traversals (clicks) per session for tasks A and B respectively. One can observe that most sessions consisted of 10 to 30 clicks.

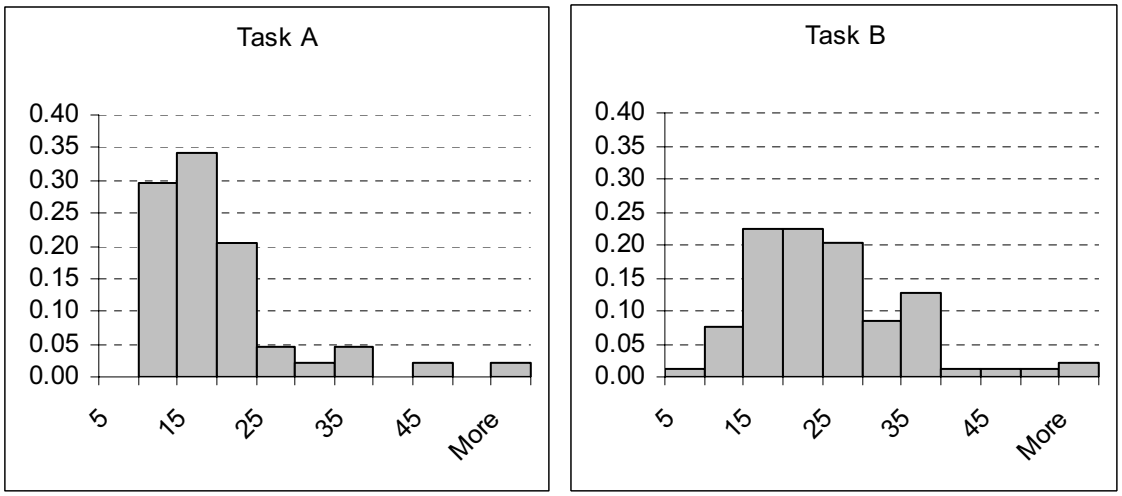

Figure 1. Relative frequency distributions of the number of traversals per session

2.2 Data Transformation

For the purpose of this project, each traversal was marked as forward (F), backward (B), or search (S). The forward traversal results when the visitor chooses a hyperlink leading to the new (previously unvisited) content, whereas a backward traversal indicates a re-visited page. The search traversal results when the consumer chooses to use the search engine, thus bypassing the navigational structure. Figure 2 shows sample clickstream data tagged as F, B, or S. The numbers in parentheses indicate the number of times a given node (page) was revisited. For instance, the first clickstream sequence indicates that the participant visited two new pages, then used the search engine, then moved back to the previously-visited page, then used the search engine again, etc.

FFSB(2)SFSSSSFFSFB(2)FFFFFFFFFF
FFFFFB(2)FFFB(2)FFFB(3)FFFFFFF
FFFFFB(2)FB(2)B(2)SB(3)B(2)B(3)
FFSFFFFSFSFB(2)FB(3)FB(4)FSSSFSSSS

Figure 2. Sample clickstream data tagged as forward, backward, or search traversals

The proposed transformation resulted in a simple representation of the navigational paths taken by the visitors to accomplish their goals.

2.3 The Model

In order to classify visitors as either browsers or shoppers, we assumed that the differences in navigational patterns between these two groups can be measured using the number of times the search engine was used and the time spent viewing content pages prior to moving forward or backward in the website. If the intention remains constant one can expect that shoppers will use the search engine more often and spend more time reading the content of each webpage [9] early in the session. On the other hand, browsers are likely to use the search engine less often and spend less time reading the content of the pages visited [9] early in the session. Below we describe a model capable of classifying visitors as browsers or shoppers using clickstream data collected after k traversals.

Let m be the total number of sessions recorded. Let n_i denote the number of pages visited by the visitor in the i-th session (including the starting page). Let t_{ij} be the time spent by the visitor viewing the j-th page ($j = 1,\ldots, n_i$) accessed in the i-th session ($i = 1,...,m$).

For each individual session *i*, and each individual traversal $k < n_i$, one can compute the total time spent on the website after the *k*-th traversal: $T_i^k = \sum_{j=1}^{k} t_{ij}$. Further, let TF_i^k denote the total page-viewing time before moving forward computed after the *k*-th traversal in the *i*-th session. Similarly, let TB_i^k be the total page-viewing time before moving backward computed after the *k*-th traversal in the *i*-th session. Also, let CS_i^k denote the number of times the search engine was used after the *k*-th traversal in the *i*-th session. For example, TB_1^5 =67 indicates that, after five traversals, the visitor in session 1 spent a total of 67 seconds viewing pages from which he or she moved "backward" in the website.

We propose the following binary logistic regression model to classify visitors as shoppers or browsers after the *k*-th traversal:

$$\pi_i = 1 / \left(1 + e^{-\left(b_0 + b_1 \sqrt{TB_i^k} + b_2 \sqrt{TF_i^k} + b_3 CS_i^k \right)} \right) \tag{1}$$

where and b_0, b_1, b_2 are the coefficients of the model, *e* is Euler's number, and π_i denotes the probability that the *i*-th session is a shopping session.

3. Results

We estimated the parameters of the model for different values of *k* (ranging from 2 to 10) using the available data. Smaller values of *k* indicate the "early in the session" period, which is the most interesting from a practical standpoint. Table 1 presents the summary of the results of fitting of the proposed model. One can observe that the model works after only four traversals and seems to improve as the number of traversals increases.

Table 1. Predictive model summary

K=	2*	3*	4	5	6	7	8	9	10
Hosmer-Lemeshow	0.18	0.21	0.46	0.32	0.20	0.78	0.81	0.90	0.88
Nagelkerke R^2	0.30	0.31	0.30	0.36	0.40	0.41	0.41	0.41	0.46
Odds ratio ($p<0.05$)									
B1 (TB)	>10000	1.18	1.373	1.53	1.571	1.368	1.294	1.264	1.204**
B2 (TF)	1.401	1.257	1.328	1.38	1.35	1.385	1.365	1.354	1.494
B3 (CS)	>10000	>10000	11.441	9.145	8.003	7.881	7.322	7.789	9.435
B0 (intercept)	-1.5564	-1.3155	-2.2845	-3.1875	-3.5307	-3.8974	-3.9522	-4.1078	-5.2814
Percentage of Shoppers									
Naïve Rate	68.1	68.1	68.1	67.9	67.9	67.9	67.9	71.7	71.7
Percentage of Correct Classification									
Browser	61.4	50.0	63.6	52.3	54.5	51.2	59.5	66.7	79.4
Shopper	78.7	84.0	76.6	86.0	87.1	90.1	80.9	79.8	79.1
Overall	73.2	73.2	72.5	75.2	76.6	77.6	74.0	75.8	79.2
Cutpoint	57%	54%	59%	50%	47%	45%	57%	58%	63%

*) Fit is questionable; some odds ratios are probably infinite
**) p-value 10%

The cutpoints were chosen to maximize the percentage of correct classification. For cases in which multiple cutpoints resulted in almost the same percentage of correct classification (a difference of less than 1%), the cutpoint maximizing the lowest percentage between browsers and shoppers was chosen.

The results are generally consistent for all values of *k*. The model has a positive lift and the values of the coefficients of the model indicate that an increase of either TF, TB, or CF, increases the odds that the actual shoppers will be classified as shoppers by the model.

4. Conclusions and Future Research

In order to dynamically personalize offers during a consumer's visit, an online retailer needs to identify the visitor's intention. Once the intention is identified, it is then possible to communicate personalized and, thus, relevant information to the consumer.

This paper demonstrates that clickstream can be used to effectively classify visitors according to their intentions related to purchasing a product or service in the current session. The output of this model could serve as input to a recommender agent thus enabling better recommendations.

The data necessary for the proposed model (session ID, time stamp, node ID) are collected in real-time by both IIS and Apache Web servers, thus, the model could be implemented as a back-office process, i.e., the process running on the Web server.

At this stage we are unable to confirm whether this model works for other websites. We expect that the way in which the content is presented affects the time spent by visitors on each individual webpage and the number of times the search engine is used. Further research is required to test this approach on different types of websites. If the model proves applicable to most websites, future research will focus on understanding the detailed intention of visitors to the website, i.e., the product, service, or piece of information that the visitor is interested in. This, however, will require incorporating semantic information into the available content-independent clickstream data.

To accomplish that, we will use website ontologies. An ontology can be defined as a "formal explicit specification of a shared conceptualization" [16] or an "explicit specification of an abstract, simplified view of a world we desire to represent [6]." Our ontology will consist of terms (keywords and phrases), and relationships among these terms. Terms will be assigned to ontological categories (conceptual types) organized in the form of a lattice [15 p. 72]. A hybrid of the deductive and inductive approaches [6] to website ontology design will be used in this project. Therefore, in addition to clickstream data, we will use language constructs extracted from the website to dynamically identify consumers' visit intention and predict the likelihood of purchase.

One can think of the resulting mechanism as of a software equivalent of a pro-active and attentive shopping assistant. We expect that this mechanism will provide some insight into inherent cognitive processes which make visitors choose certain hyperlinks over others. It is also expected to help identify general and detailed intentions of visitors and make better recommendations by the website's recommender agent. This dynamic adaptation of the website offerings [2] could lead to an increase in the purchase incidence.

References

[1] Bloch, P.H., Sherrell, D.L., and Ridgeway, N.M. "Consumer Search: An Extended Framework," *Journal of Consumer Research*, 1986, 13(1), 119-126.

[2] Brusilovsky, P. & Nejdl, W. "Adaptive Hypermedia and Adaptive Web," In M.P. Singh (Ed.), *Adaptive Hypermedia and Adaptive Web*, CRC Press, Boca Raton, Florida, 2004.

[3] Bucklin, R.E. & Sismeiro, C. "A Model of Web Site Browsing Behavior Estimated on Clickstream Data," *Journal of Marketing Research*, 2003, 40(3), 249-267.

[4] Engel, J.F., Kollat, D.T., and Blackwell, R.D. *Consumer Behavior*, Rinehart & Winston, Holt, 1973.

[5] Goldfarb, A. "Analyzing Website Choice Using Clickstream Data," In M. Baye (Ed.), *Analyzing Website Choice Using Clickstream Data*, 209-230, Elsevier Science Ltd., 2002.

[6] Holsapple, C.W. & Joshi, K.D. "A Collaborative Approach to Ontology Design," *Communications of the ACM*, 2002, 45(2), 42-47.

[7] Johnson, E.J., Moe, W.W., Fader, P.S., Bellman, S., and Lohse, J. "On the Depth and Dynamics of World Wide Web Shopping Behavior," *Management Science*, 2004, 50(3), 299-308.

[8] Kalczynski, P.J., Senecal, S., and Nantel, J. "Predicting Online Task Completion with Clickstream Complexity Measures: A Graph-Based Approach," *International Journal of Electronic Commerce (IJEC)*, 2006, 10(3), 121-141.

[9] Moe, W.W. "Buying, Searching, or Browsing: Differentiating between Online Shoppers Using In-Store Navigational Clickstream," *Journal of Consumer Psychology*, 2003, 13(1&2), 29-40.

[10] Moe, W.W. & Fader, P.S. "Uncovering Patterns in CyberShopping," *California Management Review*, 2001, 43(4), 106-117.

[11] Moe, W.W. & Fader, P.S. "Capturing Evolving Visit Behavior in Clickstream Data," *Journal of Interactive Marketing*, 2004, 18(1), 5-19.

[12] Nantel, J. & Senecal, S. "The Effect of Counterproductive Time on Online Task Completion," In M. Craig-Lees, G. Gregory and T. Davis Eds.), *The Effect of Counterproductive Time on Online Task Completion*, vol. 7, 2007.

[13] Senecal, S., Kalczynski, P.J., and Nantel, J. "Consumers' Decision-Making Process and Their Online Shopping Behavior: A Clickstream Analysis," *Journal of Business Research*, 2005, 58(11), 1599-1608.

[14] Sismeiro, C. & Bucklin, R.E. "Modeling Purchase Behavior at an E-Commerce Web Site: A Task-Completion Approach," *Journal of Marketing Research.*, 2004, XLI(August 2004), 306-323.

[15] Sowa, J.F. *Knowledge representation : logical, philosophical, and computational foundations*, Brooks/Cole, Pacific Grove, 2000.

[16] Studer, R., Benjamins, R., and Fensel, D. "Knowledge Engineering: Principles and Methods," *Data Knowledge Engineering*, 1998, 25(1-2), 161-197.