Proceedings Paper:
Guo, B and Sismeiro, C (2020) Between Click and Purchase: Predicting Purchase Decisions Using Clickstream Data. In: Advances in Consumer Research. 50th Annual Conference of the Association for Consumer Research (ACR 2019), 17-20 Oct 2019, Atlanta, Georgia, USA. Association for Consumer Research, pp. 608-609. ISSN: 0098-9258.

Between Click and Purchase: Predicting Purchase Decisions Using Clickstream Data

Abstract

We develop innovative approaches of using customers' online navigation-path data to predict purchases. We describe customers' navigation paths as sequence of browsing behaviours. We predict viewing behaviours and use this predicted viewing behaviour to predict purchases. Our approach improves the prediction accuracy compared to existing modelling alternatives.

1. Introduction

Clickstream data provides us with the record of a site visitor's journey on a website (Montgomery, 2001). This navigation path can allow researchers to uncover rich information about each individual customer such as goals (Pirolli & Card, 1999), types of visits (Moe, 2003) and purchase tendency. Previous research has indeed noted that path information, both within and across websites, has been underutilised (Hui, Fader & Bradlow, 2009). The complex trajectory of individual browsing behaviours, and the sheer exponential number of possible paths, brings significant challenges to the fully understanding and prediction of customers' entire path on a website. The inadequate understanding of customers' entire path can lead to significant managerial inefficiencies (Anderson & Cheng, 2017; Li & Kannan, 2014; Moe, 2003; Montgomery et al., 2004).

Among the few past studies on path information, most describe paths as a sequence of webpages viewed by individual site visitors in accordance with a specific typology of webpages (home page, product information pages, check out pages, etc.). Although past results suggest that customers' sequence of viewed page categories can be good indicators of customers' goals and purchase tendencies (Moe, 2003), the sequence of page categories has limitations in marketing practice. One of the major limitations is that the approach of using sequence of page categories lacks detailed information and insights about what exactly customers are doing when viewing a certain type of page. Lack of the detailed information about viewing behaviour can lead to biased prediction of purchase tendency. For example, in our studies of customers' search for airline tickets, we find that purchase tendencies are significantly different between customers who repeatedly search for a different route and customers who repeatedly search for different departure/arrival dates, although both groups of customers view a wide number of the same type of page.

In order to deal with this disadvantage and improve the approaches of using path data to predict purchases, we propose a new approach of describing a customer's path within a website and use this path information to predict purchases. Our proposed approach relies on information of the customer's actions during a sequence of page views. We care not only about the type of page (page category) but also what the customer does at each page.

In this paper, we develop our concept of sequence of online actions as a sequence of viewing behaviours. This conceptualization of browsing paths allows us to describe not just how much activity people engage in at the website (e.g., how many pages are viewed) but also what customers were doing while browsing (e.g., filtering for products vs. exploring for additional products). We propose that in accounting for the detailed actions customers are engaged in can help the prediction of purchases. We then conducted multiple analysis to test whether sequences of viewing behaviours can better indicate purchase tendencies. We finally developed an innovative approach of predicting the next viewing behaviour and using the predicted viewing behaviour to predict purchases.

2. Data and Conceptualisation of Sequences

We collected customers' page-by-page clickstream data from a leading European online travel agency (OTA). The data was generated when customers were searching for airline tickets of twelve domestic routes. We kept only the data generated by the customers who registered with the OTA and were given users' names, which allows us to recognise the multiple page requests by the same individual.

The clickstream data records a "search" when a customer inputs the following information and click the search button: departure and arrival cities; departure and return dates; number of adults, children, infants; the arrival and departure airports. The website will return a page showing this customer the list of available flights. The page of search result is equivalent to the "category page" in Moe's research (Moe 2003) where customers can view a list of

available products. Customers can filter the returned results according to these aspects: departure and arrival airports; morning/afternoon/evening departure and arrival hours; direct or connected flight; airline companies. When a customer filter the searched results, the website will return another page showing only the flight options that meet the filtering requirements. The clickstream data record it as another page request. We linked each individual customer's page requests into this customer's sequence of browsing behaviours according to the time series along which each page request happened.

We find that although customers browsing behaviours are highly diverse in terms of the searched information and filtering requirements, customers' browsing behaviours can be described according to four aspects: whether the customer repeatedly views a group of options; whether the customer filters the options, whether the customer searches for a different departure/arrival date and whether the customer searches for a different departure/arrival city. According to these four aspects, a customer's browsing can be described as one of the four Viewing Behaviours: repeated-viewing(R), filtering(F), search-for-a-different-date(DD) and search-for-a-different-route(DR). We thus can describe customers' entire journey on a shopping website as the sequence of Viewing Behaviours.

We removed the data of those customers who made only one search which cannot be described as any of these Viewing Behaviours and the data of customers who made only two searches. Our final dataset has 4,020 sequences (i.e., done by 4,020 individual customers) with a total of 22,353 actions. Each sequence of actions provides a picture of an individual customer's path on this website within the 32-day period observation. The final data set has an on-average 7.34 page requests per customer, ranging from a minimum of two page requests to a maximum of 54 page requests. Among the 22,353 actions, we have 4,020 initial searches, 3,643 (16%) repeated viewing, 2,568 (11%) filtering, 6,891 (31%) searches for a different date, and 5,231 (23%) searched for a different route.

We have 1368 distinctive sequences in the 4020 sequences of viewing behaviours. The most frequent sequence of viewing behaviours is "initial(first) search, followed by a search for a deferent date, followed by another search for another different date", which takes 6.94% of the total of 4020 sequences. Figure 1 shows the top five most frequent sequences in our data set. We have more than one thousand unique sequences that appeared only once and was generated by one customer in this entire 1368 sequences of viewing behaviours within the 32-day observation. It shows that customers are highly varied according to their sequence of viewing behaviours.

3. Methodology

We adopted multiple quantitative methods to answer a series of correlated questions. Table 1 summarises our aim of each studies and the research results in 4 steps. We describe the details of each step in this conference paper.

3.1. Grouping the sequence of viewing behaviours

Customers' sequences of viewing behaviours are highly diverse. In order to know whether sequences of viewing behaviours indicate different purchase tendencies, we group them using sequence analysis. We used a normalised longest-common-prefix (LCP) method (Elzinga, 2007; Gabadinho et al., 2011) to calculate the distances (differences) between every pair of sequences. The distance between a pair of sequences (X and Y) is D(X, Y):

$$D(X,Y) = 1 - \frac{AP(x,y)}{\sqrt{|x| * |y|}} \qquad (1)$$

$|x|$ and $|y|$ denote the length of Sequence X and Sequence Y. For example, if X is: Initial-Search_DD_DD_DD, its length is 4. $AP(x,y)$ is the length of the longest common prefix of X and Y. If Y is Initial-Search_DD_DD_DR, the $AP(x,y)$ is 3. We calculated the distances between every pair of the 4020 sequences and made a 4020*4020 matrix of distances. We conducted a two-dimensional classical multidimensional scaling on this matrix of distance. We found that the sequences of viewing behaviours can be clustered into three groups. Table 2 shows the key statistics of each cluster.

Table 2 shows that Cluster 1 has significantly more search-for-a-different-route(DR) than other three viewing behaviours, while Cluster 2 has more frequent repeated-viewing (R). Search-for-a-different-date(DD) is the most frequent viewing behaviour in Cluster 3.
According to Moe (2003), the number of repeated viewing and the number of different product categories viewed are the most significant variables distinguishing customers with strongest purchase tendency from customers whose primary purchase is hedonic browsing or gathering information. Customers who have a bigger number of repeated viewing tend to show a stronger purchase tendency while customers who viewed more different category pages show stronger tendency of browsing (Moe, 2003). We assume that Cluster 2 should have a higher purchase rate than the other two clusters, since Cluster 2 has more repeated-viewings than the other two clusters, while the other two clusters request a wider variety of category pages. We summarised the purchase rate and found that Cluster 2 has a page-level average purchase rate of 0.32 in all page requests except the initial request, higher than Cluster 1 and Cluster 3. We found, however, the page-level average purchase rate in Cluster 1 (0.31) is not dramatically lower than Cluster 2, while it significantly higher than the purchase rate in Cluster 3 (0.21). Though both Cluster 1 and Cluster 3 indicate the behaviour of viewing a wide variety of category pages because of the repeated searches for a different route or date, these two clusters, however, show different levels of purchase tendency. This result shows that customers who make more searches for a different travel dates are more flexible, or having a weaker purchase tendency, than the customers who search for a wider variety of routes, though both groups of customers view a wide variety of category pages. This result also shows that our approach of using customers' viewing behaviours to predict purchases can yield a more detailed and accurate insight of purchase tendencies, compared with the approaches of using page categories viewed.

3.2. Predicting the next viewing behaviour and purchases

We find that there is persistence in viewing behaviours, which should enable us to predict the next viewing behaviour at the next page request. We assume that the persistence is due to the unobserved individual-specific heterogeneity that leads to similar behaviours/decisions. The predicted viewing behaviour at the next page request can be used to predict the purchase decision at this page request, since different viewing behaviours indicate different levels of purchase tendencies.

We assume that a site visitor i's choice of the t^{th} viewing behaviour is a joint consequence of the unobserved individual-specific heterogeneity (denoted as M_i) and the information this site visitor observed at the $(t-1)^{th}$ page request. The information consists of the average price, denoted as w_{it-1}, across all available options on this page and the number of available options, denoted as x_{it-1}. We include the distance d_{it-1} between the departure and arrival cities to account for the possible heterogeneity across routes. We also assume that the number of options observed on $(t-1)^{th}$ page request is influenced by the site visitor's $(t-1)^{th}$ viewing behaviour.

We adopt a structural equation modelling approach for model estimation and prediction. The viewing behaviour that leads to each page request, denoted as y_{it}, is the endogenous variable in the structural model. Our modelling approach combines a multinomial

logit model (of the choice of viewing behaviour y_{it}) and a Poisson regression model (of the number of options observed on (t-1)th page request). We hypothesize that the persistence in different viewing behaviours is influenced by different unobserved factors. Accordingly, we divided M_i into three latent variables M_i^2, M_i^3 and M_i^4. M_i^2 is the latent variable influencing the choice of a filtering (F); M_i^3 influences the decision of searching-for-a-different-date (DD) and M_i^4 influence searching-for-a-different-route (DR) and the choice of a repeated-viewing behaviour is set as reference level.

We set the repeated-viewing, denoted as 1, as the reference category; filtering as 2; searching-for-a-different-date as 3 and searching-for-a-different-route as 4. Equation (1) denotes the latent utility of site visitor *i* to choose a viewing behaviour *n* at the tth page request. The probability of y_{it} =n, n=2, 3, 4 is shown in Equation (2). The probability of y_{it} =1 is shown in Equation (3).

$$V_{it}^n = M_i^n + \beta_1 w_{it-1} + \beta_2 x_{it-1} + \beta_3 d_{it-1} + \beta_0 \quad (1)$$

$$P(y_{it} = n) = \frac{\exp(V_{it}^n)}{1+\sum_{h=2}^{4} \exp(V_{it}^h)} \quad (2)$$

$$P(y_{it} = 1) = \frac{1}{1+\sum_{h=2}^{4} \exp(V_{it}^h)} \quad (3)$$

We model the sequentially exogenous variable of number of flight option, x_{it-1}, as a Poisson regression model on the proceeding viewing behaviour y_{it-1}, the (instrumental variable of) price of this choice set w_{it-1}, the distance d_{it-1} and an individual-level random intercept L_i which accounts for the unobserved, site-visitor level effect on the probability of observing a certain number of flight options. We adopt a ML-SEM approach to fit the two models simultaneously. We use Moral-Benito's (2013) and Allison and co-authors' method (Allison, Williams & Moral-Benito, 2017; Moral-Benito, 2013) for model identification.

We predicted the probability of the four viewing behaviours for each individual site visitor at every page request, except the initial page request. We use the viewing behaviour with the biggest probability as the predicted outcome. We use the data of half the customers for model estimation and another half of the customers for validation. We successfully predicted 65% of the viewing behaviours in-sample and 64% out-of-sample. Table 3 shows coefficients estimated and the in-sample and out-of-sample rates of correct prediction.

We adopt a logit model with individual-specific intercept to predict each site visitor's purchase decisions. We assume that a site visitor's purchase decision at tth page request will be influenced by the site visitor's purchase tendency which is indicated by the viewing behaviour, an individual-specific latent variable that captures other unobserved factors, price and the number of available flight options observed at this page request and the interaction between decision time limit and number of options. We correctly predicted 58% purchases in-sample and 57% out-of-sample, which is higher than Montgomery and co authors' (2004) correct prediction rate that used the sequence of webpage categories as predictor.

4. Conclusion

We aim to improve the approaches of using path data to predict purchases by overcoming the disadvantages of existing approaches of predicting purchases using sequence of page categories. To achieve this aim, we first develop the concept of sequence of viewing behaviours. We find that the viewing behaviour of "repeated-viewing" indicates the strongest tendency to make a purchase, while the behaviour of "search-for-a-different-date" indicates the strongest tendency of non-purchase. In order to use viewing behaviour to predict purchase

probability of the next search request, we develop the modelling approaches of predicting the next viewing behaviour and using this predicted viewing behaviour to predict the purchase probability. Our method achieves higher rate of correct prediction of purchases than past research using sequence of page categories.

Our research has multiple contributions. First, we contribute to the research on customers' path information by developing the concepts of sequences of customers' viewing behaviours to describe customers' paths. We have shown that customers' sequence of browsing behaviours can uncover more detailed information on purchase tendencies than sequences of page categories. Besides, our approach of summarising path information into sequences of browsing behaviours is easy to adept to websites of varied structures and page categories. Second, we develop a modelling approach of using sequence of browsing behaviours to predict purchases. Our modelling approach involves a reduced computation load, while achieve a higher correct prediction rate than models using sequence of viewed page categories in past studies.

5. Tables and Figures

Table 1: Summary of Results

Aim	Result
1. Develop a new approach of describing customers' navigation paths in order to account for the detailed actions customers are engaged in at the website	We describe a customer's browsing behaviour as one of the four Viewing Behaviours: repeated-viewing, filtering, search-for-a-different-date and search-for-a-different-route. We can describe customers' entire journey on a shopping website as the sequence of Viewing Behaviours.
2. To understand whether sequences of viewing behaviours can better indicate different purchase tendencies, compared with sequences of page categories viewed.	We find that sequences of viewing behaviours can better indicate different purchase tendencies, compared with sequences of page categories in the past studies.
3. To develop a modelling approach of predicting the next viewing behaviour	We adopt a multinomial logit model of a customer's choice of the viewing behaviour on the unobserved individual-specific latent variable and the information this site visitor observed at the previous page request. We successfully predicted 65% of the viewing behaviours in-sample and 64% out-of-sample.
4. To develop a modelling approach of predicting purchase decision using predicted viewing behaviour	We adopt a logit model of a site visitor's purchase decision on the predicted viewing behaviour that captures purchase tendency, an individual-specific latent variable, price and the number of available flight options observed at this page request and the interaction between decision time limit and number of options. We correctly predicted 58% purchases in-sample and 57% purchases out-of-sample.

Table 2: Statistics of Sequences of Viewing Behaviours

	Number of sequence	Length of sequence	Number of viewing behaviour*	Frequency: repeated viewing	Frequency: filtering	Frequency: search for different date	Frequency: search for different route
Cluster 1	1218	3	6366	604	573	1263	2708
Cluster 2	1252	4	7355	2219	1496	1268	1120
Cluster 3	1550	3	8632	820	499	4360	1403

*this number of viewing behaviours includes the number of initial search

Table 3: Estimated Coefficients and In-sample and Out-of-sample Hit Rate

Variable	1(R)	2(F)	3(DD)	4(DR)
Price (w_{it-1})	(Base outcome)	-0.015***	0.006***	0.005***
Number of flight option (x_{it-1})		0.3***	0.354***	0.311***
Distance(d_{it-1})		0.0005**	-0.00004	-0.00009
Constant		-0.58**	-2.12***	-1.85***
Var(M^2_i)		1.66		
Var(M^3_i)			2.457	
Var(M^4_i)				1.979
Hit rate: in-sample: 65% (odd customer id) out-of-sample: 65% (even customer id)				
* p<0.1 **p<0.05 ***p<0.001				

Figure 1: Top Five Most Frequent Sequences of Viewing Behaviour

Sequence	Number of Sequence	Percentage in Total Number of Sequences
Start Search → DR → DR	279	6.94%
Start Search → DD → DD	236	5.87%
Start Search → DR → DD	150	3.73%
Start Search → DD → DR	123	3.06%
Start Search → DD → DD → DD	111	2.76%

6. References

Allenby, G.M. & Rossi, P.E. 1998, "Marketing models of consumer heterogeneity", Journal of Econometrics, vol. 89, no. 1-2, pp. 57-78.

Allison, P.D., Williams, R. & Moral-Benito, E. 2017, "Maximum Likelihood for Cross-lagged Panel Models with Fixed Effects", Socius, vol. 3, pp. 2378023117710578.

Anderl, E., Becker, I., Wangenheim, F.V. & Schumann, J.H. 2014, "Mapping the customer journey: A graph-based framework for online attribution modeling".

Anderson, C.K. & Cheng, M. 2017, "Multi-Click Attribution in Sponsored Search Advertising: An Empirical Study in Hospitality Industry", Cornell Hospitality Quarterly, vol. 58, no. 3, pp. 253-262.

Elzinga, C.H. 2006, "Sequence analysis: Metric representations of categorical time series", Sociological methods and research.

Gabadinho, A., Ritschard, G., Mueller, N.S. & Studer, M. 2011, "Analyzing and visualizing state sequences in R with TraMineR", Journal of Statistical Software, vol. 40, no. 4, pp. 1-37.

Hui, S.K., Fader, P.S. & Bradlow, E.T. 2009, "Path data in marketing: An integrative framework and prospectus for model building", Marketing Science, vol. 28, no. 2, pp. 320-335.

Li, H., Kannan, P.K., 2014. Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment. Journal of Marketing Research, 51(1), pp.40-56.

Moe, W.W. & Fader, P.S. 2004, "Dynamic conversion behavior at e-commerce sites", Management Science, vol. 50, no. 3, pp. 326-335.

Moe, W.W. 2003, Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-Store Navigational Clickstream.

Montgomery, A.L. 2001, "Applying quantitative marketing techniques to the internet", Interfaces, vol. 31, no. 2, pp. 90-108.

Montgomery, A.L., Li, S., Srinivasan, K. & Liechty, J.C. 2004, "Modeling online browsing and path analysis using clickstream data", Marketing science, vol. 23, no. 4, pp. 579-595.

Moral-Benito, E. 2013, "Likelihood-based estimation of dynamic panels with predetermined regressors", Journal of Business & Economic Statistics, vol. 31, no. 4, pp. 451-472.

Moral-Benito, E., Allison, P.D. & Williams, R.A. 2017, "Dynamic panel data modelling using maximum likelihood: an alternative to Arellano-Bond".

Pirolli, P. & Card, S. 1999, "Information foraging.", Psychological review, vol. 106, no. 4, pp. 643.

Sismeiro, C. & Bucklin, R.E. 2004, "Modeling purchase behavior at an e-commerce web site: A task-completion approach", Journal of Marketing Research, vol. 41, no. 3, pp. 306-323.

Srinivasan, S., Rutz, O.J. & Pauwels, K. 2016, "Paths to and off purchase: quantifying the impact of traditional marketing and online consumer activity", Journal of the Academy of Marketing Science, vol. 44, no. 4, pp. 440-453