6-18-2022

IMPUTING OR SMOOTHING? MODELLING THE MISSING ONLINE CUSTOMER JOURNEY TRANSITIONS FOR PURCHASE PREDICTION

Jihao Lin
Loughborough University, j.lin@lboro.ac.uk

Christopher P. Holland
Loughborough University, c.p.holland@lboro.ac.uk

Nikolaos Argyris
Loughborough University, n.argyris@lboro.ac.uk

IMPUTING OR SMOOTHING? MODELLING THE MISSING ONLINE CUSTOMER JOURNEY TRANSITIONS FOR PURCHASE PREDICTION

Research in Progress

Jihao Lin, Loughborough University, J.Lin@lboro.ac.uk

Christopher P. Holland, Loughborough University, C.P.Holland@lboro.ac.uk

Nikolaos Argyris, Loughborough University, N.Argyris@lboro.ac.uk

Abstract

Online customer journeys are at the core of e-commerce systems and it is therefore important to model and understand this online customer behaviour. Clickstream data from online journeys can be modelled using Markov Chains. This study investigates two different approaches to handle missing transition probabilities in constructing Markov Chain models for purchase prediction. Imputing the transition probabilities by using Chapman-Kolmogorov (CK) equation addresses this issue and achieves high prediction accuracy by approximating them with one step ahead probability. However, it comes with the problem of a high computational burden and some probabilities remaining zero after imputation. An alternative approach is to smooth the transition probabilities using Bayesian techniques. This ensures non-zero probabilities but this approach has been criticized for not being as accurate as the CK method, though this has not been fully evaluated in the literature using realistic, commercial data. We compare the accuracy of the purchase prediction of the CK and Bayesian methods, and evaluate them based on commercial web server data from a major European airline.

Keywords: Online Customer Journey, E-commerce, Clickstream Data, Markov Chain, Purchase Prediction.

1 Introduction

Web server tracking software captures clickstream data that provides rich insights into online customer behaviour (Koehn, et al., 2020, Montgomery, et al. 2004, Moe and Fader 2004), which can be used to improve digital marketing strategies in areas such as natural and paid search, website design, affiliate marketing, conversion of searchers to buyers, and improved e-service to increase customer retention. Clickstream data is therefore a valuable resource for all e-commerce companies. In this paper, some novel ideas on purchase prediction are proposed, which are based on research in the European airline industry, and builds on earlier research that used Markov Chain theory to model online customer behaviour (Lin, et al., 2021).

Clickstream data represents a website user's online footprint and contains information such as pages visited, device type, time duration, the sequence of page visits and purchase behaviour. There are two types of clickstream data, site-centric and user-centric. Site-centric data are collected within one single website, and has very detailed information about all aspects of website users' behaviour on a single website (Bucklin and Sismeiro 2003, 2009). Adobe and Google Analytics provide useful insights but these tend to be based on average behaviour of users or very simple statistics such as the level and growth of unique visitors, bounce rates, and the average time spend on pages. User-centric data tracks a

panel of users across multiple websites and therefore provides insights into how users move across websites, e.g., between search and social media, or between online travel agents and multiple airline websites. Commercial examples of Internet panels include comScore and Nielsen Netratings and can be used for competitor analysis and the evaluation of broad search and buying behaviour, e.g., to measure the online consideration set (Holland, et al., 2020).

This paper develops a technique for modelling online customer behaviour that can be applied to both site-centric and user-centrick clickstream data. An important concept in clickstream data is a *session*, which is defined as a group of user interactions within a website that take place within a given time frame. If marketers can predict a customer's purchase intent during their session, positive interventions can be made to facilitate their online search and improve the likelihood of a purchase. Previous work in this area has shown some promise but has received very little attention (Antonellis, et al. 2009).

In the data mining literature, Markov Chain (MC) models are mainly used for link prediction, i.e., predicting what a user is going to click next (Liu and Lü, 2010, Jayalal, et al. 2007). The order of the MC model indicates the length of dependency of the future state on the past history. For example, considering all the pages in a website as the state space of a Markov Chain, a 2nd order MC is used to model the probability of the next page a customer is going to visit, then we make the assumption that this probability only depends on the previous two pages that they visit. The issue arises when some particular transitions are not observed, therefore, these probabilities will be estimated as zero using maximum likelihood estimator. This will cause the probability of the occurrence for some unique sequences that have not been seen in the training set to be zero if we deploy the model for prediction. Two main approaches can be used to tackle this problem, imputation and smoothing. On the one hand, imputation fills in these zero probabilities with substituted values of choice; on the other hand, probability smoothing assigns non-zero probability to events that were not observed in the training data (Hiemstra, 2009). In a Markov Chain model, imputation can be done through the CK equation, while smoothing is done via imposing a prior distribution on the parameters of interest. The advantage of using the CK equation is the potential for higher accuracy, but this comes with higher computational cost and some transitions remaining zero after imputation. The philosophy behind smoothing is the assumption that all events are plausible before we have seen the data. We will focus on comparing the predictive performance of two main approaches, imputation and smoothing, to solve this missing value problem.

2 Related Work

Markov models have been used to predict Internet browsing behaviour since the 1980s (Papoulis & Saunders, 1989) and more recently to reveal user navigation patterns. For example, Cadez et al., 2003 used the mixture of first order Markov Chains to cluster the user sessions for a news website. Research in the airline industry used Markov Chain transition probabilities to represent customers' decision processes and identified data-driven online market segments using cluster analsysis (Lin, et al., 2021). These academic studies illustrate the utility and power of MC theory to model clickstream data and to identify behavioural patterns that have important managerial implications, e.g., market segmentation. However, they do not attempt to predict user behaviour at a granular level.

Examples from data mining (Zhu, et al., 2002; Sarukkai, 2000) attempt to predict the next link that a user will choose. Lower order MC models are easy to train but have lower accuracy than higher order models. For example, Pirolli & Pitkow, 1999 evaluated the predictive power of different order MC models from first-order to ninth-order, and showed that that strong, longer path dependencies in higher order models lead to better prediction. One can also train different orders of the models and use all of them in prediction, which is referred to as using an all k-th-order Markov Model (Pitkow & Pirolli, 1999). However, as the order increases, the model requires a much larger state space and a low coverage issue.

To tackle this problem, an algorithm to prune the unnecessary states was proposed by Deshpande & Karypis, 2004, which pruned the states based on their frequency, prediction confidence and errors. A further innovation was demonstrated in web streaming data, which used sliding windows of the sequence, (Bernhard, et al., 2016). Based on these MC studies, which can be grouped together as next-

click prediction models, Lakshminarayan, et al., 2016 used a first-order MC to predict the conversion and impute the missing probabilities by the Chapman-Kolmogorov equation. Our work is an extension of this study, where we will be comparing the prediction accuracy between imputation using the CK equation and on the other hand smoothing using a Dirichlet prior for the estimation of transition probabilities.

3 Data Set and Methodology

The clickstream data we will be used in our study is from an airline website. The data set contains 7.8 million search and buying sessions of January 2020. The data is in tabular format, where each row contains the following information: session ID, click order, name of the page and a label that indicates whether it is a buying session or not. We try to build a classifier to distinguish whether a visitior is going to purchase, based on their navigation path. Table 1 shows a sample of a user session for our data set.

Session ID	Page Name	Click Order	Label
100145	Homepage	1	Non-buy
100145	Flight Selection	2	Non-buy
100146	Item Confirmation	1	Buy
100146	Passenger Details	2	Buy
100146	Payment Details	3	Buy
100146	Order Confirmation	4	Buy

Table 1 An example of Web Log Data

We can stitch the pages a user goes through by the session ID and click order, which is also called sessionisation. For example, after sessionisation, the data in Table 1 will have the format:

Session ID	
100145	[Homepage, Flight Selection]
100146	[Item Confirmation, Passenger Details, Payment Details, Order Confirmation]

Each session consists of a sequence of pages that a visitor goes through. Depending on the order of the Markov Chain, each sequence is decomposed into $(n+1)$-grams, i.e all the consecutive sequences of $n+1$ elements from the overall sequence. For example, given a session with 5 pages

$$Sequence = (P_1, P_2, P_3, P_4, P_5),$$

the corresponding decomposition of this sequence for a second order Markov Chain will be a trigram, $(P_1, P_2, P_3), (P_2, P_3, P_4), (P_3, P_4, P_5)$, since the transition probability, p_{ijk} is determined by the total counts of occurance of the trigram (i, j, k).

Following Lakshminarayan, et al., 2016, we split the data into two categories, buyer and non-buyer, labelled as $y_i \in \{1,0\}$ for every session i. Then the data set is divided into training set and testing set, two transition matrices are created in parallel, one for buyers and the other for non-buyers, namely TM_1 and TM_0 . The decision rule is given by comparing the class conditional probabilities, if $\frac{P(session\ i|TM_1)}{P(session\ i|TM_0)} > 1$, then it is classified as buyer, vice versa.

The problem with using the maximum likelihood estimator for training the model is that some transition probabilities can be zero, due to the lack of specific page transitions in the training set. If the training set contains no occurrence of a user jumping from page a to b then the associated transition probability estimate is zero. To solve this, Lakshminarayan, et al., 2016 proposed the idea of imputing these probabilities by using the Chapman Kolmogorov equation. The equation states that given the state space $S = \{S_1, S_2, S_3, \dots\}$,

$$P(X_{m+n} = j|X_0 = i) = \sum_{s\in S} P(X_{m+n} = j|X_m = s)\, P(X_m = s|X_0 = i),$$
$$for\ i, j \in S\ and\ m, n \geq 1$$

This is used to impute the first order Markov Chain transition probability by setting

$$\begin{aligned} & m = n = 1, \\ p_{ij} &= P(X_1 = j|X_0 = i) \\ &\approx P(X_2 = j|X_0 = i) \\ &= \sum_{s\in S} P(X_2 = j|X_1 = s)P(X_m = s|S_0 = i) = \sum_s p_{is}p_{sj} \end{aligned}$$

Similarly we can state the generalized form of imputing any order of Markov Chain transition probabilities. For a k-th order Markov Chain this is as follows:

$$\begin{aligned} p_{ij\ldots mnr} &= P(X_{t+c} = r|X_{t+c-1} = n, X_{t+c-2} = m, \ldots, X_{t+c-k+1} = n, X_{t+c-k} = i) \\ &\approx P(X_{t+c+1} = r|X_{t+c-1} = n, X_{t+c-2} = m, \ldots, X_{t+c-k+1} = n, X_{t+c-k} = i) \\ &= \textstyle\sum_{s\in S} P(X_{t+c+1} = r, X_{t+c} = s|X_{t+c-1} = n, X_{t+c-2} = m, \ldots, X_{t+c-k+1} = n, X_{t+c-k} = i) \\ &= \textstyle\sum_{s\in S} P(X_{t+c+1} = r|X_{t+c} = s, X_{t+c-1} = n, X_{t+c-2} = m, \ldots, X_{t+c-k+1} = n) \times \\ &\qquad P(X_{t+c} = s|X_{t+c-1} = n, X_{t+c-2} = m, \ldots, X_{t+c-k+1} = n, X_{t+c-k} = i) \\ &= \textstyle\sum_{s\in S} p_{n\ldots mnsr}\, p_{in\ldots mns} \end{aligned}$$

Importanlty, despite imputing the missing probabilities with this equation, some may still remain zero. Therefore, we may also consider an alternative method to smooth these probabilities. The maximum likelihood estimator used for Markov Chain model is based on the assumption that each row of the transition count matrix is a multinominal distribution. Under a Bayesian setting, we may impose a prior distribution for these multinomial paramters to avoid them to be zero. There are many probability distributions can be chosen as a prior for multinomial distribution, we choose Dirichlet distribution because of the conjugacy and computational convenience.

For example, suppose $n_i = (n_{i1}, n_{i2}, n_{i3}, \ldots n_{iM})$ are the transition counts of the ith state in a first order Markov Chain with M states. Then

$$n_i \sim Multi(p_{i1}, p_{i2}, p_{i3}, \ldots, p_{iM}).$$

Now we impose a Dirichlet prior on $p_i = (p_{i1}, p_{i2}, p_{i3}, \ldots p_{iM})$,

$$p_i \sim D(\alpha_{i1}, \alpha_{i2}, \alpha_{i3}, \ldots, \alpha_{iM}),$$

and then the posterior distribution can be derived by

$$\begin{aligned} P(p_i|Data) &\propto \textstyle\prod_{k=1}^{M} p_{ik}^{\alpha_k - 1} \prod_{k=1}^{M} p_{ik}^{n_{ik}} \\ &= \textstyle\prod_{k=1}^{M} p_{ik}^{n_{ik}+\alpha_k - 1} \\ &\sim D(n_{i1} + \alpha_{i1}, n_{i2} + \alpha_{i2}, \ldots, n_{iM} + \alpha_{iM}). \end{aligned}$$

Using the mean of the Dirichlet distribution to estimate the transition probabilities will give us:

$$\hat{p}_{ij} = \frac{n_{ij}+\alpha_{ij}}{\sum_{k=1}^{M} n_{ik}+\alpha_{ik}}.$$

By setting $\alpha_{i1}, \alpha_{i2}, .., \alpha_{iM} > 0$, we can ensure that the transition probability will never be zero even if there is no observed transition from i to j (because it will simply be estimated as $\frac{\alpha_{ij}}{\sum_{k=1}^{M} n_{ik}+\alpha_{ik}}$). By choosing $\alpha_{i1} = \alpha_{i2} = \cdots = \alpha_{iM} = 1$, we have the uniform prior for these probabilities.

These are two very different approaches to tackle the missing transition problem, one uses one step ahead transition probability to approximate the original probability, whereas under Bayesian setting, it is believed that every transition is possible by imposing a prior distribution.

4 Experiment and Result

Based on the decision rule given above, we examine the results of imputation and smoothing on the testing set as a function of partial session lengths, specifically for lengths of 5, 10, 15, 20, 25 and 30. The metrics reported here are: false positive rate (FPR)=$\frac{fp}{fp+tn}$ and F1-score=$\frac{2\times precision\times recall}{precision+recall}$, where precision=$\frac{tp}{tp+fp}$, recall=$\frac{tp}{tp+fn}$, and *tp*, *fp*, *tn*, *fn* are true positives, false positives, true negatives and false negatives, respectively. The following figures show the F1-score and FPR of predicting whether a session is a purchase session, under imputing and smoothing methods.

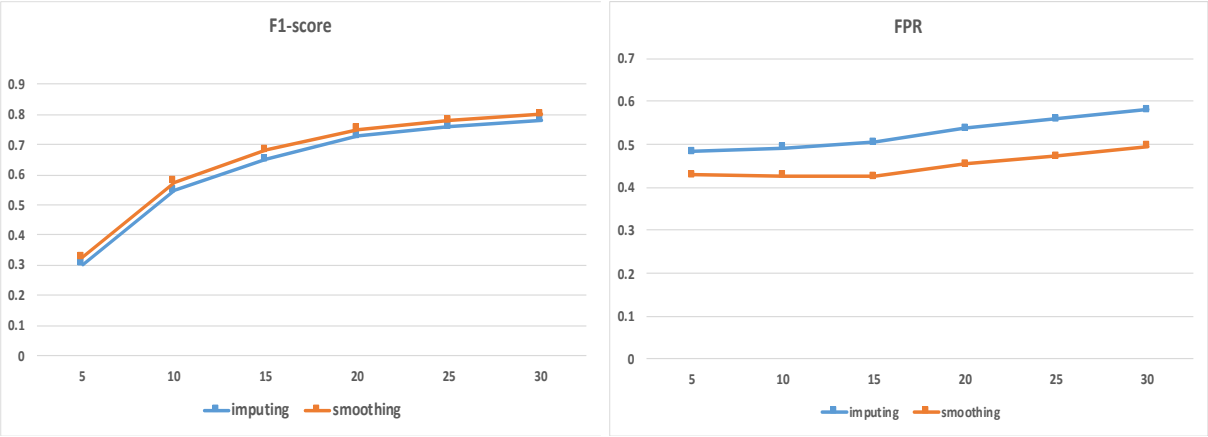

Figure 1: F1-score and FPR for a 1st order Markov Chain for purchase prediction

From Figure 1, it can be seen that as the session length increases, the F1-score increases as well, for either method. This is expected as customers interact more with the websites, we have more information regarding whether they are going to purchase or not. The FPR increases as the session lengths increase for imputing method, however, it only increases in the range of session length from 15 to 30 for smoothing. Comparing the imputing and smoothing, we notice that their F-scores are very close, with higher value for smoothing method. The gap is larger for the FPR, with consistently higher values for the imputing method, which suggests that it has falsely classified more non-purchased sessions. Overall, smoothing method is only marginally better than imputing for F1-score, but noticeably better in terms of FPR.

5 Conclusions

In this research, we give a more general form for imputing the missing probabilities for any order of Markov Chain models using the Chapman Kolmogorov equation. In addition, we explore an alternative approach to dealing with missing transitions, that uses a Bayesian approach to estimate transition probabilities. The advantage of using the Bayesian method is that computational time is saved as we only need to initialise the transition count matrix with our prior belief. However, it is shown previously that imputation increases the prediction accuracy, without comparing it to the smoothing method (Lakshminarayan, et al., 2016).

We run the experiment on the clickstream data from an airline website for the duration of one-month, with 7.8 million search and buying sessions. We conclude that the imputation method does not perform better than smoothing method on our data set, which is validated by both F1-scores and false positive rates. The purpose is to compare and better understand differences in prediction power of the two methods of dealing with missing transitions in discrete Markov Chain models, and to explore the commercial potential and managerial implications of these new data analytics methods, underpinned by Markov theory, in terms of insights and implications from detailed, and live, analyses of the online customer journey.

A further enhancement to both models is to consider the use of ensemble algorithms that combine the predictive power of different orders of MC models after imputation or smoothing. Another is to impose different prior distributions or combine expert opinions on setting the prior parameters for smoothing method.

References

Antonellis, P., Makris, C. and Tsirakis, N., 2009. *Algorithms for clustering clickstream data. Information Processing Letters*, 109(8), pp.381-385.

Bernhard, S., Leung, C., Reimer, V. & Westlake, J., 2016. *Clickstream prediction using sequential stream mining techniques with Markov chains.* s.l., 20th International Database Engineering & Applications Symposium.

Bucklin, R.E. and Sismeiro, C., 2003. *A model of web site browsing behavior estimated on clickstream data*. Journal of marketing research, 40(3), pp.249-267.

Bucklin, R.E. and Sismeiro, C., 2009. *Click here for Internet insight: Advances in clickstream data analysis in marketing*. Journal of Interactive marketing, 23(1), pp.35-48.

Deshpande, M. & Karypis, G., 2004. *Selective markov models for predicting web page accesses*. *Transactions on Internet Technology*, pp. pp.163-184.

Hiemstra, D., 2009. *Probability Smoothing*. In: *Encyclopedia of Database Systems.* s.l.:s.n.

Holland, C., Thornton, S. & and Naudé, P., 2020. B2B analytics in the airline market: Harnessing the power of consumer big data. *Industrial Marketing Management,* pp. pp.52-64..

Jayalal, S., Hawksley, C. and Brereton, P., 2007. *Website link prediction using a Markov chain model based on multiple time periods*. International journal of Web engineering and technology, 3(3), pp.271-287.

Koehn, D., Lessmann, S. and Schaal, M., 2020. *Predicting online shopping behaviour from clickstream data using deep learning*. Expert Systems with Applications, 150, p.113342.

Lakshminarayan, C., Kosuru, R. & Hsu, M., 2016. *Modeling complex clickstream data by stochastic models: Theory and methods.* s.l., 25th International Conference Companion on World Wide Web.

Lin, J. et al., 2021. A Machine Learning Approach to Online Market Segmentation. *Social Science Research Network.*

Liu, W. and Lü, L., 2010. *Link prediction based on local random walk*. EPL (Europhysics Letters), 89(5), p.58007.

Moe, W.W. and Fader, P.S., 2004. *Dynamic conversion behavior at e-commerce sites*. Management Science, 50(3), pp.326-335.

Montgomery, A.L., Li, S., Srinivasan, K. and Liechty, J.C., 2004. *Modeling online browsing and path analysis using clickstream data*. Marketing science, 23(4), pp.579-595.

Papoulis, A. & & Saunders, H., 1989. *Probability, random variables and stochastic processes*.

Pirolli, P. & Pitkow, J., 1999. *Distributions of surfers' paths through the World Wide Web: Empirical characterizations*. s.l., World Wide Web.

Pitkow, J. & Pirolli, P., 1999. *Mining longest repeating subsequences to predict world wide web surfing.* s.l., USENIX Symposium on Internet Technologies & Systems.

Sarukkai, R., 2000. *Link prediction and path analysis using Markov chains*. *Computer Networks* , Volume 33, pp. 377-386.

Zhu, J., Hong, J. & Hughes, J., 2002. *Using markov chains for link prediction in adaptive web sites*. s.l., Soft Issues in the Design, Development, and Operation of Computing Systems .