LETTER

Improving Purchase Behavior Prediction with Most Popular Items

Chen CHEN[†], Member, Jiakun XIAO[†], Chunyan HOU^{††a)}, and Xiaojie YUAN[†], Nonmembers

SUMMARY Purchase behavior prediction is one of the most important issues to promote both e-commerce companies' sales and the consumers' satisfaction. The prediction usually uses features based on the statistics of items. This kind of features can lead to the loss of detailed information of items. While all items are included, a large number of features has the negative impact on the efficiency of learning the predictive model. In this study, we propose to use the most popular items for improving the prediction. Experiments on the real-world dataset have demonstrated the effectiveness and the efficiency of our proposed method. We also analyze the reason for the performance of the most popular items. In addition, our work also reveals if interactions among most popular items are taken into account, the further significant improvement can be achieved. One possible explanation is that online retailers usually use a variety of sales promotion methods and the interactions can help to predict the purchase behavior.

key words: recommender system, behavior analysis, prediction, e-commerce, session

1. Introduction

As online shopping becomes popular, the behavior analysis of e-commerce customers has become an increasingly important business tool for promoting sales. Many e-commerce companies, such as Amazon in USA and Taobao in China, are looking for all possible approaches to detect customers' intent and predict their purchase behaviors. Both the consumers and online businesses benefit from this kind of prediction technique. For example, a user, who is not registered and a temporary visitor, can be attracted by recommending him the discounted item if the user's preference for this item is determined by his short-term browsing information. The ability to predict what a customer will purchase is also valuable for e-commerce companies. Exact predictions for the next purchased products make e-commerce companies to implement a product recommendation system, determine the positions of products in the result of a customer's search query, optimize the collection of products to be displayed or suggest products for customer's shopping basket.

The purchase prediction becomes a challenge when there is only the http session which consists of a list of clicks in a short time. The prediction with the short-term history

becomes an area of growing research and commercial interest in recent years. Especially, RecSys Challenge 2015* is associated with such problem. A history of user's click behavior during a browsing session at a website of online retailer is given, and the goal is to predict whether a user will purchase at the end of this session. If a session ends with the purchase behavior, this session is called a buy session.

Because the information in a session is really limited, it becomes essential to make full use of session information to help purchase behavior prediction. Participants of RecSys Challenge 2015 use a variety of statistical features based on the short-term session. Most of them ignore the detailed information of items clicked in the session. For example, the number of items clicked in a session is defined as a feature, but we do not know which items are clicked. On the other hand, if we include all items that are clicked in a session, a large number of features has a negative impact for machine learning algorithms to train the predictive model efficiently.

Our study reveals that it is more appropriate to include just most popular items than all items. If the most popular items are used by indicator features, both the effectiveness and efficiency can be enhanced for purchase behavior prediction. In addition, we find that if the interactions among most popular items can be used by the machine learning algorithms, the further improvement can be achieved. We also investigate why the interactions play an important role based on the prices of bought items. There is a majority of bought items whose prices are zero or varied. One possible explanation is that retailers usually use sales promotion methods such as price reduction sales and "Buy one. Get one free" deals. Those promotion activities are associated with a group of items, which make the bought items affect other other, so the interactions among those items are effective in purchase behavior prediction.

The rest of paper is organized as follows. Section 2 introduces related works. Section 3 describes the problem and our approach. Experiment setup and results are discussed in Sect. 4. In Sect. 5, we conclude our paper.

2. Related Work

Session-based recommendation has been researched in the music recommendation. Park et al. [7] develop a session-based collaborative filtering, which can use the session information to capture sequence and repetitiveness in the lis-

Manuscript received August 8, 2016.

Manuscript revised October 3, 2016.

Manuscript publicized November 7, 2016.

[†]The authors are with the College of Computer and Control Engineering, Nankai University, Tianjin, China.

^{††}The author is with the School of Computer and Communication Engineering, Tianjin University of Technology, Tianjin, China.

a) E-mail: houchunyan@tjut.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2016EDL8169

*<http://2015.recsyschallenge.com/>

tening process. Zheleva et al. [14] develop a session-based hierarchical graphical model using latent dirichlet allocation and show that their model could facilitate playlist completion based on previous listening sessions or songs that the user has just listened to. Dios et al. [4] explore the usage of temporal context and session diversity in session-based collaborative filtering techniques for music recommendation. Although there have been some studies in developing recommendation techniques based on the session, few work has been done to research purchase behavior prediction based on e-commerce sessions.

With respect to purchase behavior prediction, Kooti et al. [6] build a model to predict when customers are most likely to make a purchase and how much will they spend based on their emails and demographic information. Tanaka et al. [10] proposes a method for inferring the factors that trigger purchases and predicting purchases. They use the histories of item purchases and media advertisement views for each individual. Zhang et al. [13] present a system for predicting a user's purchase behaviors on e-commerce websites from the user's social media profile. Chen et al. [3] predict purchase behaviors in e-commerce based on the product impression and click behavior of advertisement. Those studies are based on the demographic information, social media profiles, or media advertisements. Compared with those researches, our work is based on a list of clicks during a browsing session at an e-commerce website. It means that short-term browsing history is just available and the prediction is required highly temporal. The most related works come from researchers [8]–[10] who participate RecSys 2015 Challenge. Our study is inspired by their idea. Moreover, we analyze the reason for the performance of the most popular items and find that the interactions among features are effective in the purchase prediction.

3. Purchase Behavior Prediction

3.1 Problem Definition

The dataset includes train and test dataset. Train and test dataset consist of sets of sessions S_{train} and S_{test} respectively. Each session s is represented as a click stream

$$c(s) = (c_1(s), c_2(s), \dots, c_{n(s)}(s)) \quad (1)$$

$$c_j(s) = (i_j(s), t_j(s), y_j(s)), j \in \{1, \dots, n(s)\} \quad (2)$$

where $n(s)$ is the number of clicks in session s , $i_j(s)$ denotes the j -th clicked item in session s , $t_j(s)$ is the time when the item $i_j(s)$ is clicked, $y_j(s)$ is one when item $i_j(s)$ is purchased at least once, and zero otherwise. A session s has the label $y(s)$ defined as

$$y(s) = \begin{cases} 1 & \exists j : y_j(s) = 1 \\ 0 & \forall j : y_j(s) = 0 \end{cases} \quad (3)$$

If $y(s) = 1$, session s is a buy session. We are given sets of purchased items for session $s \in S_{train}$, and are required to predict $y(s)$ for session $s \in S_{test}$.

3.2 Indicator Features

There are many unique items in the dataset. We regard the number of clicks on an item as the measure of its popularity, rank unique items by their number of clicks in descending order, and extract top items in different proportions. Suppose we extract top M items, that is, the most popular M items. For a session s , let $V(s)$ be a $1 \times M$ vector and the m -th element of $V(s)$ is defined as

$$V_m(s) = \begin{cases} 1 & I_m \in i(s) \\ 0 & I_m \notin i(s) \end{cases} \quad (4)$$

where $i(s)$ is the set of items clicked in session s , and I_m is the item which corresponds to the m -th element of $V(s)$.

In this way, we can create a $1 \times M$ feature vector for each session which consists of M indicator features. The indicator features enable the machine learning algorithms to item-aware parameters and learn interactions between groups of clicked items and buy sessions.

3.3 Machine Learning Algorithms

In order to validation the effect of the most popular items, we choose three typical algorithms which are Logistic Regression (LR), Factorization Machines (FM) [8] and Gradient Boosting decision tree (GB) [5]. LR can be seen as a special case of generalized linear model and the interactions among features are not taken into account. Let w_i and x_i be the i -th element of weight vector w and feature vector x . The LR model can be expressed as

$$y(x) = \sigma \left(w_0 + \sum_{i=1}^p w_i x_i \right) \quad (5)$$

FM is a generic approach for using factorization models to estimate pairwise interactions in the machine learning task. The FM model of order $d = 2$ is defined as:

$$y(x) = \sigma \left(w_0 + \sum_{i=1}^p w_i x_i + \sum_{i=1}^p \sum_{j=i+1}^p \langle v_i, v_j \rangle x_i x_j \right) \quad (6)$$

GB is a powerful model for both regression and classification problems. The model consists of a number of individual decision trees and each tree is trained by the residual of previous trees. It usually produces competitive, highly robust, interpretable model, which is especially appropriate for mining from noisy data and containing high order interactions among features. The GB model is formulated as:

$$y(x) = \sigma \left(\sum_{i=1}^M T_i(x) \right) \quad (7)$$

where T_i is the i -th decision tree.

The reasons of selecting these algorithms are as follows. First, their capacity to model interactions among features becomes gradually stronger from LR, FM to GB. Second, compared with SVM etc., these algorithms are appropriate for the large-scale dataset and usually used by the

teams in RecSys 2015 Challenge.

4. Experiments

4.1 Dataset

RecSys 2015 Challenge provides a large number of sessions from an online e-commerce retailer [1]. Our experiment is based on this dataset. As shown in Table 1, the dataset is large-scale and extremely imbalanced. There are 33 million sessions and only 5.5% of those session is buy sessions. The average number of clicks per session is about 4. It means that the click information per session is less and just 4 clicks in average is used to predict whether a session is a buy session. About 38% of all unique items are bought at least once. More than a half of all items are not purchased at all, so it is not necessary to include all items in the prediction. We randomly select 75% of sessions in the train dataset as the train-part dataset while the 25% sessions are kept as the valid-part dataset. There are two reasons why we make such split and do not conduct the cross validation on the train dataset. Firstly, we find that major improvements on the valid-part dataset are always translated to the test set. The fixed validation split is enough for the parameter tuning of models. Secondly, the train dataset is large-scale and our split can avoid the time-consuming cross validation on the large dataset. Thus, for all experiments, we use the train-part dataset to learn a model and tune all parameters by the valid-part dataset.

4.2 Setup

We have two groups of features. One group describes basic session features and the other is features with most popular items. The basic features used in the experiments are listed in Table 2. Those basic features are usually used by participants of RecSys 2015 Challenge and proved to be effective for the purchase prediction. The other group is based

on the most popular items. There are 52,739 unique items in train dataset, so up to additional 52,739 items are included in features. We regard the number of clicks on an item as the measure of its popularity. All unique items are ranked by their number of clicks in descending order. The top 1%, 5%, 10%, 20%, 50% and 100% of ranked items are selected as the most popular items.

We use the precision, recall and F1-score to evaluate the prediction models. As an average of the precision and recall, F1-score is used as the evaluation metric to tune parameters on the valid-part dataset. Normalization is used to scale the range of features' value and make the converge faster. We use two-sided paired approximate randomization tests to assess the statistical significance of the difference between two F1-scores with a significance level of $p = 0.05$.

For LR, we use LibLinear[†] which is a library for the linear classification. The standard $L2$ regularization is used in the experiments. For FM, we use LibFM^{††} which is an implementation of FM. Markov Chain Monte Carlo is used to learn the model. For GB, we use the XGBoost^{†††} library. We concentrate on the binary classification with the logistic loss function, and use tree boosters and $L2$ regularization term on weights.

4.3 Result and Discussion

As shown in Table 3, the performance can be enhanced when the most popular items are included with different machine learning algorithms. For a machine learning algorithm, the F1-score with more stars is significantly better than that with less stars. The improvement with most popular items is more significant than without them. When the top 5% of the most popular items are added, the best effectiveness is achieved. When more items are included, the improvement is not significant, but the more time and memory are required to train the model. The top 5% is the recommended proportion based on both the effectiveness and the efficiency. For machine learning algorithms, GB achieve the best performance and the FM outperforms LR. We would like to know why most popular items can enhance the performance. The reason may be that the most popular items tend to be bought, and the sessions which include the most popular items are more likely to be buy session.

The performances become better from LR, FM to GB. LR does not consider the interactions among features while FM is able to estimate pairwise interactions and GB can model the high order interactions. In order to explain why interactions play a significant role in the prediction, we investigate the prices of bought items. It is found that when at least two items are bought simultaneously in some sessions, one of them is bought at the zero price. There are 275,407 sessions which consist of an item whose price is zero and account for about 54% out of all buy sessions. The promotion activities (e.g. "Buy one. Get one free" deals or free

Table 1 Statistics of dataset.

dataset	click	session	item
train	33,003,944	9,249,729	52,739
test	8,251,791	2,312,432	42,155
buy	1,150,753	509,696	19,949
train-part	24,754,609	6,937,297	50,678
valid-part	8,249,335	2,312,432	41,994

Table 2 Basic session features.

feature	description
click	#(clicks in the session)
max click	#(max clicks on an item in the session)
time span	time span of the session
item	#(unique items clicked)
category	#(unique categories clicked)
day	day when the last item is clicked
month	month when the last item is clicked
day of week	day of week when the last item is clicked
hour	hour when the last item is clicked

[†]<https://www.csie.ntu.edu.tw/~cjlin/liblinear/>

^{††}<http://www.libfm.org/>

^{†††}<https://github.com/dmlc/xgboost/>

Table 3 Experimental results.

metric	model	0%	1%	5%	10%	20%	50%	100%
Precision	LR	0.193	0.199	0.219	0.219	0.215	0.221	0.217
	FM	0.185	0.258	0.265	0.266	0.269	0.270	0.269
	GB	0.192	0.250	0.268	0.272	0.272	0.271	0.273
Recall	LR	0.353	0.414	0.397	0.413	0.428	0.419	0.427
	FM	0.415	0.325	0.376	0.382	0.382	0.375	0.359
	GB	0.359	0.402	0.448	0.456	0.454	0.456	0.456
F1-score	LR	0.249	0.269*	0.282**	0.286**	0.287**	0.290**	0.288**
	FM	0.257	0.288*	0.311**	0.313**	0.316**	0.314**	0.308**
	GB	0.241	0.309*	0.335**	0.340**	0.340**	0.340**	0.341**

gifts) possibly happen in those sessions. In addition, we find that the prices of some bought items vary in different sessions, but do not equal zero. There are 106,604 sessions which consist of items whose price are reduced and account for about 21% out of all buy sessions. In most cases, the price reduction is due to discount activities such as “20% off once purchase up to 200” sales or the coupons. Retailers usually use sales promotion methods such as discount sales and “three for the price of two” deals. Therefore, the clicked items in a session affect each other and the interactions also determine the buy behavior. The interaction can be learn by FM and GB, rather than LR. Thus, GB and FM outperform LR. GB is better than FM because GB can model high order interactions among features.

5. Conclusions

In this paper, we explore to use most popular items for purchase behavior prediction. Experimental results show that indicator features based on the most popular items are a promising way to improve the prediction with different machine learning algorithms. This is a supplement of traditional features. We give the explanation for the performance of the most popular items. In addition, we reveal that the interactions among most popular items are effective for predicting the purchase behavior. One possible explanation is that online retailers usually use sales promotion methods and the interactions can help to predict the purchase behavior.

In the future, it is interesting to explore other features for the purchase prediction. We also wish to validate our work on more datasets.

Acknowledgments

Research was partially supported by National Natural Science Foundation of China (No. 61170184, 61402242, 61402243 and 61402333), the Tianjin Municipal Science and Technology Commission (No. 13ZCZDGX02200, 13ZCZDGX01098), the Research Foundation of Ministry of Education and China Mobile Communications Corporation (No. MCM20150507).

References

- [1] D. Ben-Shimon, A. Tsikinovsky, M. Friedmann, B. Shapira, L. Rokach, and J. Hoerle, “RecSys challenge 2015 and the YOO-CHOOSE dataset,” *Proc. RecSys Challenge*, pp.357–358, 2015.
- [2] V. Bogina, T. Kuflik, and O. Mokryn, “Learning item temporal dynamics for predicting buying sessions,” *Proc. 21st International Conference on Intelligent User Interfaces*, pp.251–255, New York, NY, USA, 2016.
- [3] C. Chen, C. Hou, J. Xiao, and X. Yuan, “Purchase behavior prediction in e-commerce with factorization machines,” *IEICE Trans. Inf. & Syst.*, vol.E99-D, no.1, pp.270–274, Jan. 2016.
- [4] R. Dias and M.J. Fonseca, “Improving music recommendation in session-based collaborative filtering by using temporal context,” *Proc. IEEE 25th International Conference on Tools with Artificial Intelligence*, pp.783–788, 2013.
- [5] J.H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, vol.29, no.5, pp.1189–1232, 2001.
- [6] F. Kooti, K. Lerman, L.M. Aiello, M. Grbovic, N. Djuric, and V. Radosavljevic, “Portrait of an online shopper: Understanding and predicting consumer behavior,” *Proc. WSDM*, pp.205–214, 2016.
- [7] S.E. Park, S. Lee, and S.-G. Lee, “Session-based collaborative filtering for predicting the next song,” *Proc. 1st ACIS/JNU International Conference on Computers, Networks, Systems and Industrial Engineering*, pp.353–358, 2011.
- [8] S. Rendle, “Factorization machines with libFM,” *ACM Trans. Intelligent Systems and Technology*, vol.3, no.3, pp.1–22, 2012.
- [9] P. Romov and E. Sokolov, “RecSys challenge 2015: Ensemble learning with categorical features,” *Proc. 2015 International ACM Recommender Systems Challenge*, Article No. 1, 2015.
- [10] Y. Tanaka, T. Kurashima, Y. Fujiwara, T. Iwata, and H. Sawada, “Inferring latent triggers of purchases with consideration of social effects and media advertisements,” *Proc. 9th ACM International Conference on WSDM*, pp.543–552, New York, NY, USA, 2016.
- [11] M. Volkovs, “Two-stage approach to item recommendation from user sessions,” *Proc. 2015 International ACM Recommender Systems Challenge*, Article No. 3, New York, NY, USA, 2015.
- [12] P. Yan, X. Zhou, and Y. Duan, “E-commerce item recommendation based on field-aware factorization machine,” *Proc. International ACM Recommender Systems Challenge*, Article No. 2, 2015.
- [13] Y. Zhang and M. Pennacchiotti, “Predicting purchase behaviors from social media,” *Proc. 22nd International Conference on WWW*, pp.1521–1532, 2013.
- [14] E. Zheleva, J. Guiver, E.M. Rodrigues, and N. Milić-Frayling, “Statistical models of music-listening sessions in social media,” *Proc. 19th International Conference on WWW*, pp.1019–1028, 2010.