Cold-start Solution to Location-based Entity Shop Recommender Systems Using Online Sales Records

Yichen Yao[1], Zhongjie Li[2]

[1]Department of Engineering Mechanics, Tsinghua University, Beijing, China
yaoyichen@aliyun.com
[2]Department of Thermal Engineering, Tsinghua University, Beijing, China
lizhongjie1989@163.com

Abstract. Cold start solution to location-based entity shop recommender system is discussed with dataset from real business scenario 'Koubei' platform. Severe cold start problem is encountered as for the rapid accumulation of new customers and new merchants. Test dataset is classified into three groups based on the amount of user information, and different recommend strategy is implemented for each user group. Purchase probability of old users is predicted using neural network with features in respect of user, time and merchant characteristics. For new customers, user-based collaborative filtering is applied with records from online retail platform Alibaba Group.

Keywords: recommender system • cold-start • collaborative filtering

1 Introduction

The rapid development of wireless communication and the ubiquitous usage of mobile device enable easy acquisition of location information. The performance of recommender system can be enhanced with adequate appliance of these location-based data. Zheng *et al.* [1] and Bao *et al.* [2] used information from user's geo-position history to help construct social networking. Liao [3] and Zheng *et al.* [4] learned human behavior pattern based on GPS location data, and human activity was predicted and local services were recommended.

Cold-start is one of the most common and challenging problems in recommender system. Extensive efforts have been made by previous researchers to deal with such

problems. Zhou *et al.* [5] predicted the new user preference by using decision tree with user-based collaborative filtering, and Liu *et al.* [6] applied linear combination of existing users to approximate the behavior of a new customer. However, with the availability of user information from other sources, the exploitation of relation between different information sources will be very helpful for dealing with cold-start problem. Lin *et al.* [7] addressed the cold-start problem for App recommendation with user information from social network. Additionally, latent factor model is applied to deal with the cold start problem. User features are projected into a reduced space that finds the most effective representation of user similarity [8]. The applicable latent factor model that is suitable for cold-start solution includes principal component analysis [9], restricted Boltzmann machine [10] and singular vector decomposition [11].

In this paper, we focus on a real business scenario with entity merchant sales log from 'Koubei' platform, which is a startup business with only five month sales log. With the rapid emergence of both new customers and new merchants, the cold-start problem should be coped with special cautions. We generated diverse recommend strategies for different user groups according to the amount of available information. Entity shopping behavior can be learned directly from past experience, or can be inferred from their online shopping logs. Typical features of the dataset are discussed and analyzed in the paper, and the corresponding algorithm is proposed.

2 Problem Statement

The problem under discussion is provided by Alibaba Tianchi big data contest platform, and the subject of this contest is "Brick-and-Mortar Store Recommendation with Budget Constraints" (http://click.aliyun.com/m/4383/). The contest focuses on location based nearby store recommendation on mobile terminals. As mobile devices become ubiquitous in our daily life, recommendation demands of entity shops like restaurant and retail stores on these terminals expand rapidly. The entity shops sales log from July to November of 2015 is provided as training set, and participates are expected to give recommendation of December 2015, and the F1 score is adopted as evaluation score. The contest aims at recommending solutions to two major issues, the cold-start problem and the supply constraint in the entity-store sales.

Firstly, location-based recommender system and the related mobile apps are blooming these years. The increasing rate of new customers and new merchants is consequentially very high at the early stage of recommender platform. A group of new

customers show up every day, and it will be a difficult task to give appropriate recommendation without previous shopping logs to speculate user's characteristic and shopping preference. Similar situation is also encountered in the app 'Koubei', which provides the recommendation and payment services in scope of this contest. Luckily, the online sale log of Alibaba Group, the largest online retail platforms in China, is provided to give insight of user's characteristic and help solve the cold-start problem.

Secondly, constraints are imposed on purchase user number of each merchant for the predicting month. This sales budget restrictive condition is in accordance with real business scenario, like the limited service capacity in a restaurant every day. For merchants whose sales performance is likely to be restricted by the budget, our systems should only recommend those popular merchants to the users that are most likely to visit.

3 Data Analysis

As the several features of the current dataset stated above distinguish from those in the traditional recommender system, a detailed analysis of the dataset will necessarily help solve the problem.

In the first place, the recommender system under discussion 'Koubei' is a start-up enterprise which owns only 5 months sales log, and the cold-start problem should be treated with caution. Figure 1 presents the weekly sales volume and number of customers within the 5 months. Apparently, the business is growing expansively and drawing a group of new users every month, and it is challenging to make precise recommendation to these new-user group. Moreover, the average number of merchants that users patronize is less than 1.75 as shown in figure 1, meaning that we might only partially infer user preference from this biased data.

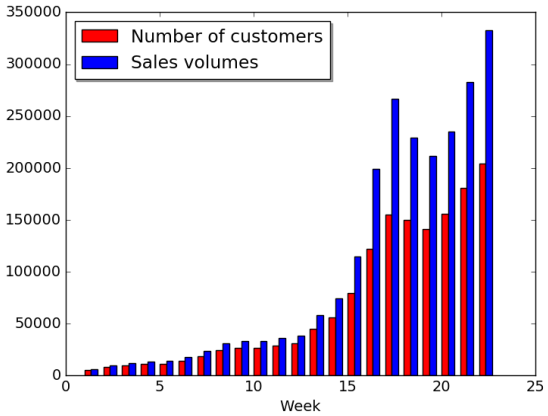

Fig. 1. Weekly sale volume and customer number over time of entity shops from Koubei platform.

To cope with the cold-start problem, the online sales log of Alibaba Group is provided, and the corresponding weekly sales volume and customer number are presented in Figure 2. Due to both business and noise concerns, the data in the great promotion period during first 3 weeks of November is not provided in the dataset. In contrast with then entity shop sales statistics, the online sales log is stable with little fluctuation over weeks, and this fully developed trading platform might give more hints about user preferences of entity shops at a given location.

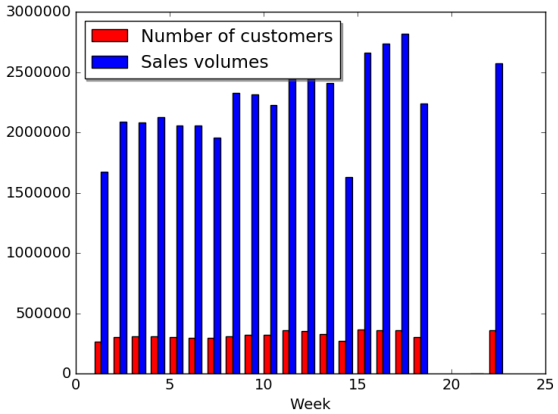

Fig. 2. Weekly sale volume and customer number over time of online retail sale platform Alibaba

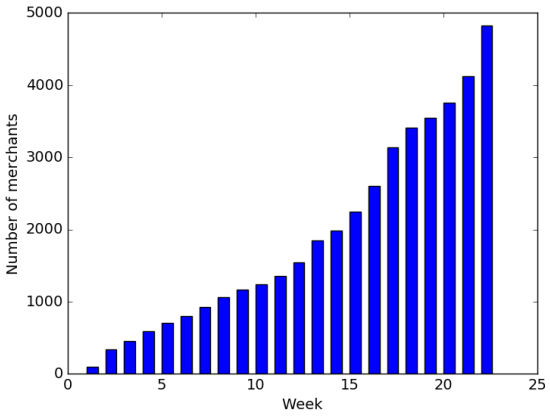

Fig. 3. Weekly growth in the number of entity shops

Meanwhile, cold-start problem is also encountered in respect of the growth of merchant number. The total number of merchants every week is shown in Figure 3, and it is indicated new merchants join into the 'Koubei' platform gradually. For the merchants that only owns a few days of trading log, it will be tricky to predict the sale volume of the next month. Also, the lack of enough historical data means that only simple predicted model should be adopted for the time-series prediction, otherwise over-fitting problem might become serious and reduce the prediction accuracy of the test dataset.

Recommender systems have its strength in coping with information overload problems, in which decision makers have difficulty in fully understanding the details of each choice due to the presence of too much information. Therefore, recommender systems make personalized suggestions to consumers, and great achievements have been made in recommending online sales product, like books and movies. Item-based collaborative filtering approaches predict the rating of a user *u* for an item *i* based on the ratings of *u* for items similar to *i*, and is always very successful in dealing with information overload issues in recommendation problem. In addition, item-based filtering is much less computational demanding than user-based approaches due to the relatively smaller item group size compared to number of users.

However, the information overload problem is much less severe for the location-based entity shops scenario under discussion. Figure 4 presents the histogram of merchant number at each location. There are totally 426 locations in the datasets, and 75% locations have merchant number less than 20. Therefore, at most of the locations,

customers only need to have a browse of all the merchants name and make a decision. Moreover, item-based collaborating filtering makes predictions based on similar items, which requires large item sets and existing items with high similarity. Obviously, in the current dataset, users do not encounter information overload at most of the locations, therefore item-based filtering is not very suitable in this recommender system. Moreover, loyal customers are more likely to patronize merchants more than once, which differs from the conventional book recommender system that users are less likely to buy the same book more than once.

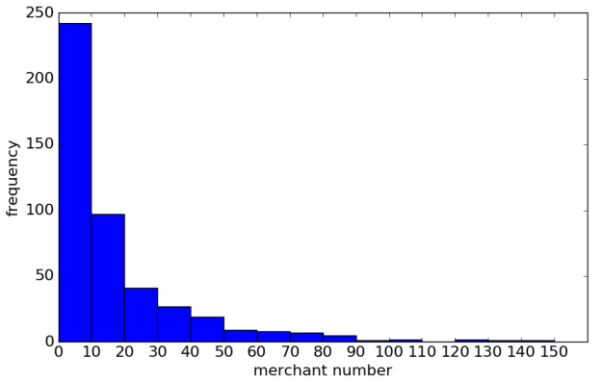

Fig. 4. Histogram of merchant number at each location

Finally, sales constraint imposed on the merchants requires us to recommend shops to the user groups that are most likely to patronize. Merchants that are likely to hit the constraint limit line are always very popular shops and the purchase probability is comparatively high. If the purchase probability is overestimated, then the early hitting of constraint line may leads to the potential loss of high-score users. In the contrast, if the purchase probability is underestimated, then the redundant recommend suggestion to users will decrease the prediction precision. Therefore, it requires giving accurate prediction of the purchase probability of every merchant at a given location.

4 Algorithm and Discussion

Based on the above analysis, the recommendation strategy we adopt should deal with cold-start problem for both the new merchants and new customers. Based on the amount of user information, the test dataset is classified into three groups, respectively old user sets, new users with online log sets and the new users without online log sets.

For old user sets, we can obtain their shopping preferences from the consumption record during the last five months. For new users with online shopping logs, we may infer that users with similar online shopping behavior will also tend to patronize analogous entity shops. For the last bunch of test sets, we have neither the online or entity shopping log, therefore the absence of user information make the recommendation depended only on the information from merchants. The dataset size for each of the three groups of people is presented in figure 5, old user account for only 24.5% of the whole test dataset and the rest are all new users. Meanwhile, we can obtain the online trade record for most of the new users, and only 16.6% of the total datasets lack any information in respect of the user. The recommendation strategy for these three group sets differs and the details are discussed below.

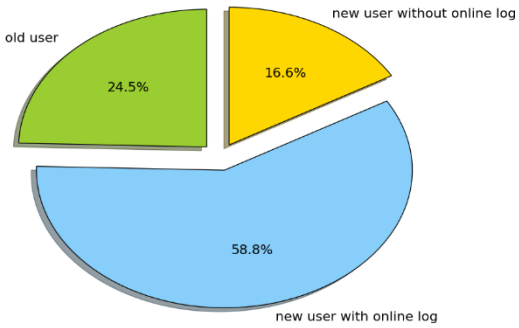

Fig. 5. Users are categorized into 3 groups according to the users' information

4.1 New user without online shopping record

For these test sets, we have neither online nor entity shopping records. Therefore the absence of user information results in that the recommendation is only depended on the merchant information. Besides, the cold-start problem of merchants should also be taken into consideration. Simple predictive model is more favorable due to the potential over-fitting problem. Currently linear regression with L1 regularization is adopted to predict the sale volume for the first week of December, and the sales volume of the last three weeks is chosen as inputs. For the training data, the output is the sales volume of a given week and the input is the value from the previous three weeks. The normalized coefficient for the previous 3 weeks in the linear model are [0.0000, -0.0772, 1.0773]

respectively, indicating that the sales behavior is only highly dependent on the previous week. The negative coefficient for the week before last week also demonstrates the cold-start severity. In the method, the total sale volume can be predicted for every merchant at each location, and the probability of a given merchant is calculated as sales volume over the total amount. In this way, the new users purchase probability is generated. However, the predicting reliability is low due to the absence of user information.

4.2 Old Customer

For old user sets, their entity shop consumption record of the last five months is provided. However, the item-based collaborating filtering is not suitable in this scenario since the merchant quantity is small at most of the locations. The merchant similarity matrix is also important information considering the complex mutual exclusion or mutual supporting relation between merchants. In our strategy, neural network is applied to predict the probability of user behavior. The features for each user-location-merchant pair comes from three main aspects, respectively related to user characteristics, time characteristics and merchant characteristics. The user characteristics for the user-location-merchant include the following features: whether patronize before or not, total patronize amount, patronize probability, online purchase amount. The time characteristics include last patronize time, whether the last patronize merchant, first patronize time. The merchant characteristics includes the similarity coefficient of the two most analogical merchants, whether patronized before or not and predicted patronize probability, the merchant amount at a location, and the predicted purchase probability from 4.1. The neural network includes 2 hidden layers and each layer includes 10 units. The cost function is defined as the cross-entropy. Using the neural network method with selected features described above, the averaged predictive precision for these old customer groups can reach 78.8%. It can be deduced that, as the old customer percentage increases, the recommender system will become stable, and the recommending precision will gradually approach this value.

4.3 New user with online shopping record

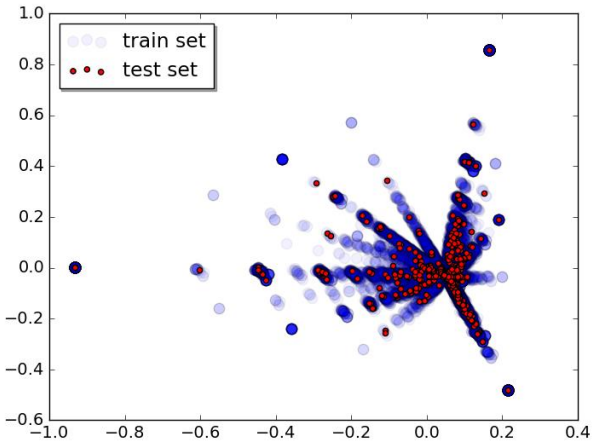

Fig. 6. User clustering with PCA from online sales records

To some extent, user's shopping behavior at entity shops share similarity with their online records. Therefore, it is reasonable to infer user shopping preference from their online logs. In our strategy, user-based collaborating filtering is applied for each user in this group set, and the top-N user is selected based on the Euclid distance of condensed user-category matrix. Figure 6 presents the diagram of first two components after the principle component analysis of the normalized user-category matrix, in which light blue circles correspond to the user from train set, while small red dots is user from test set. It is obvious that there exists blue circles in the vicinity of every red dot, and then the top-N user in the train set can be founded. The appliance of PCA enables dataset to be projected into the plane of maximum variance, and the user distance calculated from the condensed matrix becomes more precise. However, the purchase probability from the above user-based CF still requires correction due to the difference between the online merchants and entity merchants. Therefore the data from November is chosen as validation sample to obtain the correction function, which is expressed as a function dependent on averaged user distance.

5 Discussion on budget constraint

As for the budget constraint imposed on merchants stipulated in the evaluation score, the recommender systems should only recommend the merchant that the users are most likely to visit. Using the method discussed in section 3.2, the purchase probability of each user-location-merchant record can be predicted. The probability is sorted in descending order, and the submission record is generated from all the records above the probability criterion. The criterion is chosen to obtain the highest F1 score, with a balance between both the precision and recall. Meanwhile, the predictive accumulated budget cost is calculated for every merchant. Once the budget cost reaches the constraint, the record submitted subsequently is removed in case of redundant recommendation.

6 Conclusions

This paper focuses on the recommend strategy based on real business scenario. The 'Koubei' platform is a startup recommender system that offers location-based suggestion to customers on mobile terminals, and severe cold-start problem is encountered with the rapid accumulation of new customers and new merchants to this platform. Also, the information-overload problem in respect of items is much less severe than the online sales scenario, since there are only less than 20 merchants at most of the locations. The test dataset is separated into three groups based on the user information. For new users without online shopping record, the purchase probability can only be calculated by predicting the sales volume, simple linear regression is applied to prevent over-fitting. For old customers, neural network is applied with features from three main aspects, namely user, time and merchants. For new users with online shopping, their shopping preference on entity shops is predicted by user-based collaborative filtering with the similarity calculated from their online shopping records. The F1 evaluation score based on our strategy is 0.4665, with precision 0.5863 and recall 0.3874. However, the discussion above presents preliminary results based on the analysis of the current dataset. Further efforts should be made to improve the overall recommend performance.

References

1. Zheng, Y., Zhang, L., Ma, Z., Xie, X., & Ma, W. Y.: Recommending friends and locations based on individual location history. ACM Transactions on the Web (TWEB), 5(1), 5.(2011)
2. Bao, J., Zheng, Y., & Mokbel, M. F.: Location-based and preference-aware recommendation using sparse geo-social networking data. In Proceedings of the 20th international conference on advances in geographic information systems, ACM, pp. 199-208 (2012)
3. Liao, L.: Location-based activity recognition, Doctoral dissertation, University of Washington (2006)
4. Zheng, V. W., Zheng, Y., Xie, X., & Yang, Q.: Collaborative location and activity recommendations with GPS history data. In Proceedings of the 19th international conference on World wide web, pp. 1029-1038. ACM. (2010)
5. Zhou, K., Yang, S. H., & Zha, H.: Functional matrix factorizations for cold-start recommendation. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 315-324. ACM. (2011)
6. Liu, N. N., Meng, X., Liu, C., & Yang, Q.: Wisdom of the better few: cold start recommendation via representative based rating elicitation. In Proceedings of the fifth ACM conference on Recommender systems, pp. 37-44. ACM. (2011)
7. Lin, J., Sugiyama, K., Kan, M. Y., & Chua, T. S.: Addressing cold-start in app recommendation: latent user models constructed from twitter followers. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp. 283-292. ACM. (2013)
8. Koren, Y., & Bell, R.: Advances in collaborative filtering. In Recommender systems handbook, pp. 145-186. Springer US. (2011)
9. Kim, D., & Yum, B. J.: Collaborative filtering based on iterative principal component analysis. Expert Systems with Applications, 28(4), 823-830 (2005)
10. Salakhutdinov, R., Mnih, A., & Hinton, G.: Restricted Boltzmann machines for collaborative filtering. In Proceedings of the 24th international conference on Machine learning, pp. 791-798. ACM.(2007)
11. Takács, G., Pilászy, I., Németh, B., & Tikk, D.: Matrix factorization and neighbor based algorithms for the netflix prize problem. In Proceedings of the 2008 ACM conference on Recommender systems, pp. 267-274. ACM. (2008)