A Recommender System Model based on Commodity-Purchase-Cycle Classification

Meina Song, Xue Zhou, Haihong E, Zhonghong Ou
Beijing University of Posts and Telecommunications
Beijing, China
{mnsong, zhouxue160, ehaihong, zhonghong.ou}@bupt.edu.cn

ABSTRACT

Recommender systems have been widely used in e-commerce platforms, such as Amazon and Taobao. Among the available recommendation algorithms, Item Collaborative Filtering (ItemCF) Algorithm and Content Filtering Algorithm have gained wide adoption because of various strengths. For example, hidden interests can be digged so as to get fresh recommendations, and highly individual recommendations can be made. Despite their strengths and wide adoption, there are still some weaknesses associated with them. One representative weakness is the existence of duplicated, and outdated recommendations due to the lack of purchasing cycles, e.g., weekly or seasonal, of goods. We name such cycles Commodity Purchase Cycle (CPC), and propose a new recommendation algorithm based on CPC in this paper. We leverage CPC attributes to modify the collaborative filtering output rating matrix acquired by the ItemCF Algorithm, and take into consideration both user behaviors and commodity characteristics to make timely recommendations. We utilize a realistic dataset from Taobao to verify the performance of the proposed algorithm. Experimental results demonstrate good performance of CPC algorithm. Specifically, from the perspective of Root Mean Square Error (RMSE), the CPC Algorithm promotes the recommendation accuracy by 15%-20%, compared with the state-of-the-art ItemCF Algorithm.

Keywords

Recommender System, ItemCF, Content Filtering, Commodity-Purchase-Cycle

1. INTRODUCTION

In the era of big data, as the number of things connected on the Internet explodes, the value of each individual shrinks accordingly. It becomes more challenging and time-consuming to dig value from the vastness of information. In the context of e-commerce platforms, recommender systems have been widely used to help end users choose the most appropriate goods [13]. For example, Taobao, Amazon, Jingdong, Netflix, and Douban, all have their own specialized recommender system. According to VentureBeat statistics, Amazon recommender system has brought in 20% - 30% more sales for Amazon [15]. Netflix also once provided millions of dollars to reward the winners in their Recommendation Algorithm Competitions [9]. From these data, we can clearly see the value of recommender systems.

In the recommender systems of e-commerce, the base algorithms widely used include Collaborative Filtering Algorithm [17] and Content based Algorithm [16]. In general, these algorithms demonstrate good performance. But in different application scenarios, there are still some problems. For example, after browsing a goods from an e-commerce platform, regardless of purchasing or not, the goods (or relevant goods) will be recommended to the end user for a long time. While some goods do not have strict time attribute, i.e., suitable throughout the year, some others might have purchasing cycles. For example, the purchasing cycle of mobile phone recharge service is one month, especially at the beginning of each month.

Keep this in mind, a proper recommender system needs to fully consider different application scenarios and different CPC attributes. After a comprehensive survey on academic research work, including Sun et al. [8], Zhao et al. [11] and He et al. [24], and representative commercial platforms, e.g., Taobao, Amazon, Jingdong, and Douban, we conclude that the CPC characteristics have not been fully utilized in the state-of-the-art recommender systems.

In this paper, we propose a recommendation algorithm based on CPC attributes. It is based on the mixture of ItemCF Algorithm and Content based Recommendation Algorithm. The overall system works as follows. Firstly, we use user logs to create a rating matrix. Secondly, we use ItemCF to filter initial rating matrix in order to get a filtered matrix as the input for the next step. Thirdly, we use the CPC classification factors obtained by ID3 Decision Tree Algorithm [3] and item lists to modify the filtered ratings. Finally, we acquire the rating results. To verify the performance of the proposed CPC algorithm, we leverage a realistic dataset from Taobao for experiments. Experimental results demonstrate the outperformance of CPC algorithm. Specifically, compared with the state-of-the-art ItemCF algorithm, the CPC Algorithm improves the recommendation accuracy by 15%-20%.

The rest of the paper is structured as follows. In Section II, we present the related work. In Section III, we introduce

MOBIMEDIA 2016, June 18-19, Xi'an, People's Republic of China

DOI 10.4108/eai.18-6-2016.2264232

the design and implementation details of the CPC system model. In Section IV, we focus on experimental evaluation. In Section V, we discuss the CPC Algorithm. Finally, in Section VI, we conclude the whole paper.

2. RELATED WORK

There are several representative recommendation algorithms used in commercial recommender systems, including Content based Recommendation Algorithm [10], Collaborative Filtering based Recommendation Algorithm [12], Association Rules based Recommendation Algorithm [22], Knowledge based Recommendation Algorithm [7], and Hybrid Recommendation Algorithm [5].

Amazon is a global e-commerce platform, and has its own specialized recommender system. Its fundamental recommendation algorithm is Collaborative Filtering based Recommendation Algorithm [17]. It is a widely used algorithm, and can be divided into User Collaborative Filtering algorithm (UserCF) and Item Collaborative Filtering algorithm (ItemCF) [1]. According to Zhao [20], Amazon mainly makes use of the ItemCF algorithm. The main idea of ItemCF is "feather flock together, people in groups". Namely, among behavior characteristics from a large number of users, the most similar users can be selected. Based on these user's behavior preferences, recommendations can be made for the target user. ItemCF has various strengths, including independence of background knowledges, capable of digging hidden interests, being able to acquire fresh recommendations [4]. Thus, ItemCF has been widely used in e-commerce recommender systems. However, it also has several weaknesses, such as the recommended items are not intuitive, the recommended reasons are not easy to explain, as well as the overlooking of the influence on user behaviors [18].

Taobao is a e-commerce platform in China, whose trading volume reached 10 Trillion in 2013 [23]. Its basic recommendation algorithm is Content based Recommendation Algorithm (CRA). Compared with ItemCF, CRA takes no consideration of the similarity between user behaviors, but pays more attention to the candidate goods themselves. The main idea of CRA is that the user interested in one item may be interested in some similar items. The obvious advantages of CRA include higher level of personalization and better interpretability of the recommendation results. Nevertheless, it also has several weaknesses, including limitation from background content references [14], not able to dig user's hidden interests, not able to get fresh or creative recommendations, and existence of repeated recommendations due to overlooking of user behaviors.

In [8], Sun et al. constructed a time series based network, which improved the recognition accuracy of the largest influence neighbor set for the current user (product). In [11], Zhao et al. proposed to use time as the context information, which is combined to the progress of collaborative filtering recommendation. It then utilizes user behavior logs to gain similarities among users, and makes use of the similarities and time attenuation factors to get the recommendation results. Nevertheless, it is unreasonable to filter out or weaken recommendation results simply based on the timestamp of the user, because some items may have purchasing cycles. For example, mobile phone recharge service has a purchasing cycle of one month. At the beginning of each month, it is the most appropriate time to recommend.

In [24], He et al. proposed a method to improve the Collaborative Filtering Algorithm based on user's purchasing records. To avoid getting too many duplicated recommendations, it lowers the rating of the items which have been purchased already. Nevertheless, filtering out the items simply because they have been purchased is not reasonable, and may bring errors. Because even for the same type of goods, a user may buy it again within a short period of time.

Based on the work mentioned above, we take advantage of both ItemCF and Content-based Algorithm, and utilize CPC Classification factors to improve the recommendation results. The CPC-based algorithm clearly differentiates this work from the state-of-the-art.

3. DESIGN AND IMPLEMENTATION

In this section, we first introduce the overall algorithm flow and system model of CPC. We then explain the basic principles and implementation details of the Item Collaborative Filtering Algorithm. Finally, we propose the classification method of CPC algorithm based on ID3 Algorithm.

3.1 Algorithm Flow and System Model

CPC algorithm takes the Taobao Logs as input of the overall system. We first construct an initial rating matrix based on the log files, then we execute the ItemCF Algorithm to gain a filtered rating matrix. After that, we classify the candidate items in the matrix using ID3 Algorithm, and get a series of cycle-classify factors. By multiplying the filtered rating matrix and the cycle-classify factors, we can acquire a modified rating matrix, which is the final recommendation result. The detailed algorithm flow is shown in Algorithm 1.

There are three sub-modules in the CPC recommendation model: information input module, processing module and result output module, We mainly focus on the processing module in the middle. The basic commodity trading database can be used in the execution of ItemCF Algorithm. The user database, goods database, and history database can be used to classify the candidate items by purchasing cycles. The output of ItemCF and the result of classification are combined to calculate the final ratings. The detailed CPC recommendation model is illustrated in Fig. 1.

Algorithm 1 Cycle-based-recommendation

Input: Taobao user's behaviour logs;

1: $Rate_{ij} = Viewtimes_{ij} * W_1 + Collect_{ij} * W_2 + Cart_{ij} * W3 + Bought_{ij} * W_4$;
2: $M = createMatrix(Rate)$;
3: $M_c = itemCF(M)$;
4: $ItemRate = classfyID3(M_c)$;
5: $M_r = M_c * ItemRate$;

Output: Revised-recommendation-matrix;

3.2 Acquisition of ItemCF Rating Matrix

There are two key steps when using the ItemCF algorithm to acquire the rating matrix [6]: (1) finding the nearest user and calculating the similarities among items; and (2) employing the similarities between items to predict the ratings and getting the recommendation result [19][25].

3.2.1 Finding the Nearest Users

There are many methods to calculate similarity, such as Cosine, Adjusted Cosine, Pearson and so on. The Mean

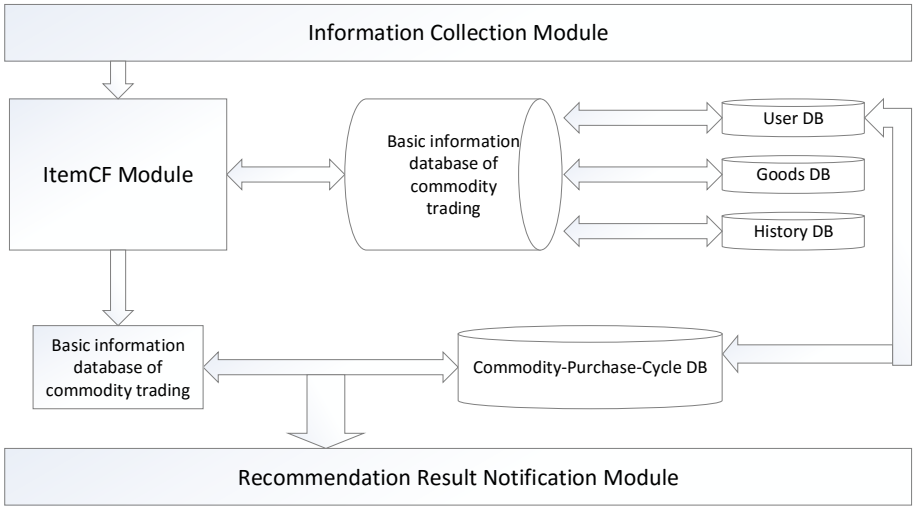

Figure 1: The CPC recommendation system model.

Absolute Error (MAE) is the mean of deviation's absolute value from single reality value and the arithmetic mean of them [19]. Under the four scales of datasets (100, 200, 500, 1000), our experiment makes use of Cosine, Adjusted Cosine, Pearson to calculate the MAE, respectively, in order to choose the least error measuring method. The detailed calculation of MAE is shown in Equation 1.

$$MAE = \frac{\sum_{i=1}^{n} | A_i - \widehat{A_i} |}{n} \quad (1)$$

$\widehat{A_i}$ is the reality score, A_i is the estimated value gained from the model [19]. The comparison is illustrated in Fig. 2. From the figure, we can see that as the item numbers increase, the MAE demonstrates a downward trend. Comparing the three methods, we can conclude that the MAE of Pearson Measurement is the smallest. Thus, we choose the Pearson Measurement as the similarity measurement in the rest of the paper.

Utilizing the Pearson Similarity Measurement, we can acquire the similarity matrix among items. It will be used as inputs of step 2.

3.2.2 *Acquisition of User-Item Rating Matrix*

In order to calculate a user's score to an item, the main principle is as follows: the more likely between the current item and historical items that the user is interested about, the more probably the current item will be ranked at the front. Equation 2 can be used to calculate the rating from user u to item j [21]:

$$P_{uj} = \sum_{i \in N(u) \cap S(j,k)} W_{ji} R_{ui} \quad (2)$$

In this equation, P_{uj} indicates the score from user u to item j, $N(u)$ indicates the set of items that the user is interested in (i is one of the items user likes), $S(i,k)$indicates the set of K items which are the closest to item i (j is an item of the set), W_{ji} indicates the similarity of item j and item i, R_{ui} indicates the rating from user u to item i [15].

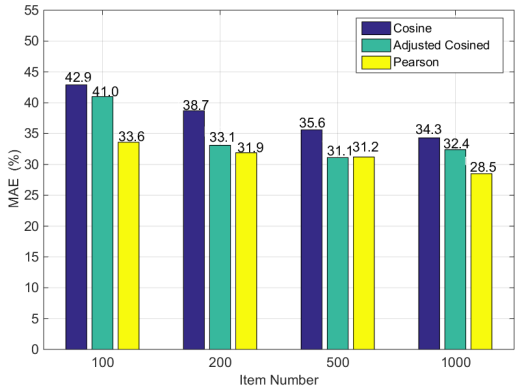

Figure 2: Mean Absolute Error (MAE) comparison from Cosine, Adjusted Cosine, Pearson Similarity Measurement under the 100, 200, 500, 1000 four scales of datasets.

3.3 Classification by Commodity-Purchase-Cycle

3.3.1 *Commodity Purchase Cycle*

In the scenarios described above, Item Collaborative Filtering Algorithm has several weaknesses. Thus, a Commodity-Purchase-Cycle (CPC) factor has been introduced. For some costly high-end goods, such as household electric appliances and digital products, when people purchase it once, they may not purchase it again during a relatively long period of time. So this kind of item is durable purchasing goods. For some middle-price goods such as clothes, a user who once purchases clothes of brand C and is interested in brand C, he might purchase it again after a period of time. So this kind of item is middle purchasing cycle goods. For low-price goods such as food and daily necessities, it may be purchased many times in a short time period, so this kind of item is

short purchasing cycle goods.

3.3.2 Classifying Goods by Purchase Cycle

The classify method based on purchasing cycles is realized by ID3 Decision Tree Algorithm. A decision tree is a simple but widely used classifier. Creating a decision tree via training datasets can give the unclassified data a classification effectively. Decision tree is a predictive model, which represents a mapping relationship between instance attributes and instance classes [2]. Generally speaking, we can build a decision tree from top to bottom. The first node created is called the root node, as it has no parent node. After that, intermediate nodes should be added corresponding to the feature of a partition of the samples' set. When using the decision tree to perform classification work, we should start from the root node of the tree, testing the root node according to an instance attribute. Then according to the tested attribute value, we move to one branch of the tree, and test the root node of the sub tree at a lower level. This procedure continues recursively until we encounter the leaf node, we then give the leaf node a class to the test instance.

In the scene described above, we utilize the price of the goods, whether it is purchased or not, and goods' ordinary classifications as three main attributes to predict the CPC. We then make use of the training dataset in Table 1 to train the decision tree until the expected classification results can be obtained. After that, we utilize the decision tree to make classification for the testing dataset (Table 1 is part of the actual training data set).

Table 1: Part of Reality Training Dataset

Price	BoughtOrNot	Class	BuyCycle
high	yes	digital	long
high	no	electric	short
high	yes	cloth	middle
high	no	office	short
high	yes	makeups	middle
high	no	food	short
low	yes	food	short
low	no	jewelry	short
low	yes	carmakes	middle
low	no	bookradio	short
low	no	bookradio	short
low	yes	cloth	short
middle	yes	addmoney	short
middle	no	makeups	short
middle	yes	baggift	middle
middle	no	mum	short
middle	yes	furniture	long
middle	no	sports	short

According to the classification result, the accuracy rate is 97.78%, which is considered to be acceptable.

Using the design and implementation introduced above, we will perform experimental evaluation through Taobao dataset in the next section.

4. EXPERIMENTAL EVALUATION

In this section, we first introduce the experimental dataset used. We then explain the Root Mean Square Error (RMSE) evaluation indicator. Finally, we show the final experimental results and give explanation for the results.

4.1 Experimental Setup

Pretreated ąřlog dataset from 200 days' Taobao user's behaviors and ąřTaobao product categoryąś dataset have been used in testing CPC recommendation system model. There are 660 thousands of behavior logs in the Taobao log dataset from 110 thousands of users in 200 days, from which we can get user history behaviors as well as commodity attributes. In this paper, records from 100 active users have been selected to constitute the initial rating matrix, and the initial matrix's fill rate is 6.2%. The ąřTaobao product categoryąś dataset is a list of classifications for Taobao goods, including electric, cloth, makeups, i.e., 15 classifications in all, which can be used in CPC Classifying. The influence attributes of commodity purchase cycle include: ordinary classifications of commodity, commodity prices, and whether purchasing it or not. Taking these attributes as the basises of classification, commodities can be further divided into long purchase cycle, middle purchase cycle and short purchase cycle commodities, using the classification factors to revise the collaborative filtering result in order to get the final ratings.

4.2 Evaluation Metrics

Root Mean Square Error (RMSE), which is the square root of the ratio of the square of the predicted value and the true value, can be used to evaluate the accuracy of recommendation result [6]. The calculation formula is as Formula 3.

$$RMSE(T) = \sqrt{\frac{\sum_{(u,m)\in T} (\widehat{r_{u,m}} - r_{u,m})^2}{|T|}} \quad (3)$$

From which T stands for the test dataset, $r_{u,m}$ m indicates the predictive score of item i to user u, $\widehat{r_{u,m}}$ indicates the actual score of item m from user u [14].

4.3 Experimental Results

For final goods rating matrix M_r from CPC Algorithm and output matrix M_c from ItemCF, calculating the RMSE under the item numbers of 100, 200, 500, 1000 respectively, the contrast of M_r and M_c are as follows in Fig. 3.

As it can be seen from the Table 3, with the increase of the item numbers, the RMSE showed a gradual downward trend overall. Compared with the result of ItemCF, the result of CPC has a reduction from 15% to 20% in RMSE. The main reason for this reduction is that, after obtaining the score results from ItemCF, classification factors are calculated according to the purchase cycle, which has been used to revise the original rating result, and then the final score result can be more close to the actual score. Therefore, the CPC model gains higher recommendation accuracy in the applications of e-business, which greatly solve the problems due to lack of considering for user behavior and commodity characteristic which leading to continually appearance for many repetitive, inappropriate recommendations.

5. DISCUSSION

Although the CPC Algorithm has taken into consideration both user behaviors and commodity characteristics to make timely recommendations and has improved 15%-20%

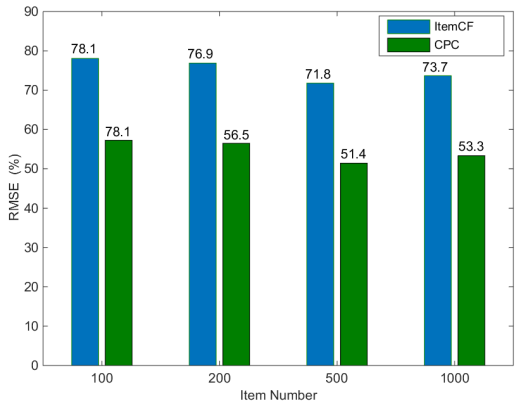

Figure 3: Root Mean Square Error (RMSE) comparison of ItemCF and CPC Algorithm under the 100, 200, 500, 1000 four scales of datasets.

recommendation accuracy in the experiment of Taobao logs, there are some specail or unexpected conditions that not fit for the CPC Algorithm. For example, after you bought a expensive mobile phone for yourself, you want to buy another for someone else, but you will never get the recommendation of this mobile phone anymore.

6. CONCLUSION

According to the characteristics of e-commerce platforms, we proposed a CPC system model based on ItemCF Algorithm, which introduces the Commodity-Purchase- Cycle Classification attributes creatively. By considering user behaviors and commodity purchasing cycle characteristic, improved the practical value of Collaborative Filtering Algorithm in the application of e-commence. After experimental testing by dataset from Taobao, in the aspect of RMSE evaluation indicator, the CPC recommendation model has shown a better predictive accuracy compared with the state-of-the-art Item Collaborative Filtering Algorithm. Thus, it has great practical significance in the application of e-commerce. In the next stage, we should utilize more decision attributes to train the classification decision tree or try some other classifiers in order to give each commodity a more reasonable and detailed purchasing cycle classification result.

7. ACKNOWLEDGEMENT

This work is supported by the National Key Project of Scientific and Technical Supporting Programs of China (Grant No.2014BAK15B01); Industry Cosponsored Project (S2016025); Cosponsored Project of Beijing Committee of Education; Engineering Research Center of Information Networks, Ministry of Education, China.

8. REFERENCES

[1] A. B. Barragáns-Martínez, E. Costa-Montenegro, J. C. Burguillo, M. Rey-López, F. A. Mikic-Fonte, and A. Peleteiro. A hybrid content-based and item-based collaborative filtering approach to recommend tv programs enhanced with singular value decomposition. *Information Sciences*, 180(22):4290–4311, 2010.

[2] R. C. Barros, A. C. de Carvalho, and A. A. Freitas. *Automatic design of decision-tree induction algorithms*. Springer, 2015.

[3] S. Bashir, U. Qamar, F. H. Khan, and M. Y. Javed. An Efficient Rule-Based Classification of Diabetes Using ID3 C4.5 and CART Ensembles. In *12th International Conference on Frontiers of Information Technology (FIT)*, pages 226–231, 2014.

[4] H. E. Bo, Y. Chen, H. Wang, and S. Dong. E-commerce collaborative recommendation system based on agent. *Computer Engineering*, 33(9):216–218, 2007.

[5] I. Cantador, A. Bellogín, and D. Vallet. Content-based recommendation in social tagging systems. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 237–240. ACM, 2010.

[6] T. Chai and R. R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3):1247–1250, 2014.

[7] J. F. Recommender system[m]. *Beijing: Beijing University of Telecommunication Press*, pages 80–92, 2013.

[8] G. Sun, L. Wu, Q. Liu . Collaborative filtering recommendation algorithm based on time sequence behavior. *Journal of Software*, (11):2721–2733, 2013.

[9] H. Chai. Research of hybrid advertising recommendation technology based on collaborative filtering and content filtering. *Beijing University of Posts and Telecommunications*, 2015.

[10] H. Xu, X. Wu, X. Li . Comparative study of internet recommendation system. *Journal of Software*, 20(2):350–362, 2009.

[11] H. Zhao, J. Hou, Q. Chen. Collaborative filtering recommendation algorithm based on time weight and trust relation. *Application Research of Computers*, 32(12):3565–3568, 2015.

[12] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[13] A. G. Lalita Sharma. A survey of recommendation system Research challenges. *International Journal of Engineering Trends and Technology*, 4(5), 2013.

[14] L. C. Leng YJ, Lu Q. Review on collaborative filtering recommendation technology. *Pattern Recognition and Artificial Intelligence*, (08):720–734, 2014.

[15] J. Liu, T. Zhou, and B. Wang. Research progress of personalized recommendation system. *Progress in Natural Science*, 19(1):1–15, 2009.

[16] M. J. Pazzani and D. Billsus. Content-based recommendation systems. In *The adaptive web*, pages 325–341. Springer, 2007.

[17] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295. ACM, 2001.

[18] J.-H. Su, B.-W. Wang, C.-Y. Hsiao, and V. S. Tseng. Personalized rough-set-based recommendation by integrating multiple contents and collaborative

information. *Information Sciences*, 180(1):113–131, 2010.

[19] T. Yao. Research on personalized recommendation based on collaborative filtering algorithm. *Beijing University of Technology*, 2015.

[20] W. Zhao. Research for personalized recommendation method based on user behavior analysis and hybrid recommendation strategy. *Beijing University of Technology*, 2014.

[21] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 448–456. ACM, 2011.

[22] H. Wang, N. Wang, and D.-Y. Yeung. Collaborative deep learning for recommender systems. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1235–1244. ACM, 2015.

[23] X. Zhang. Taobao ten years: Turnover from 34 million to 1 trillion[eb/ol]. 2013.

[24] Y. He, C. Song. Improved collaborative filtering recommendation based on user purchase records. *Computer Engineering and Design*, (9):3091–3094, 2014.

[25] Z. Wu. Research and application on collaborative filtering algorithm in web recommender system. *East China Normal University*, 2014.