RESEARCH ON PERSONALIZED RECOMMENDATION ALGORITHM OF CROSS-BORDER E-COMMERCE UNDER LARGE DATA BACKGROUND

Shujun Ji
Zhejiang Fashion Institute of Technology
No 495, Fenghua Road, Jiangbei District
Ningbo, Zhejiang, 315211
China
shujunj66@sina.com

Abstract. With the expansion of the E-commerce industry, the coverage of goods is becoming increasingly wider. Moreover, foreign E-commerce industries have gradually expanded to the Chinese market, resulting in higher requirements of domestic consumers on the safety, variety and cost performance of foreign products. Therefore, how to make the sales of cross-border E-commerce more stable and keep the balance of cross-border E-commerce inventory and sales is an urgent problem to be solved. As traditional modes are not suitable for foreign E-commerce industries, a personalized recommendation system with favorable big data processing capacity is needed to address the problem. This paper introduced a common personalized recommendation system and applied the collaborative filtering algorithm as the main algorithm of the system to solve problems in practice. The results showed that the improved collaborative filtering recommendation system could meet the requirement of the times and was worth being promoted.

Keywords: large data, cross-border E-commerce, personalized recommendation system.

Introduction

Cross-border E-commerce refers to an international commercial activity through which transaction bodies of different countries reach deals, make payments and cross - border deliveries via e-commerce platform. Though the development and popularization of computer technology and information technology has brought convenience to peoples life in recent years, it has also caused some problems due to the imperfect cross - border E-commerce development system and the shortage of professional staffs, which have been studied by some experts. Wang W [1] believed that cross-border E-commerce was an important way to export Chinese products, an effective method to solve the logistic problems should be sought, and government investment and information exchange should be enhanced. He also proposed a new E-commerce operation mode to reduce human labor and realize the rational use of labor resources [2]. Holding that the traditional recommendation algorithm had high requirement on accuracy and was not easy to implement, Song S et al. [3] put forward a user-based Slope One algorithm.

Zhou X et al. [4] argued that in a society where information was exploded, it was difficult for users to find information of interest to them, and it was difficult for users to obtain information with low page views, which were common problems in the development of the E-commerce industry, among which information overload problem was the most important and most urgent problem to be solved. Therefore, the personalized recommendation technology was introduced to help improve users shopping experience, increase user stickiness, and promote the E-commerce site sales.

1. Cross-border E-commerce under the big data background

With the advent of the Internet age and the continuous accumulation of application system data, there are more and more digital products and Internet costumers, which produces rich big data resources [5]. In this study, the Hadoop architecture was applied to analyze massive amounts of data. Besides, the core problems of big data recommendation were considered.

At present, the personalized recommendation system is divided into collaborative filtering recommendation, content-based recommendation and recommendation based on association rules [6]. Collaborative filtering recommendation is the earliest and most widely used personalized recommendation technology [7] which mainly includes user-based collaborative filtering and item-based collaborative filtering. In addition, there are also content-based personalized recommendation system [8] and rule-based personalized recommendation system [9]. In this study, the improved user-based recommendation algorithm was used to solve the problems. The Hadoop distributed computer was used to sort and store user commodity log. Cross-border E-commerce personalized recommendation services include recommendation of the products which customers may like, new products and relevance of commodities. The specific process is as follows:

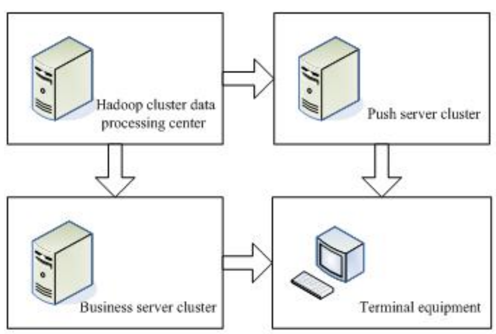

Figure 1: The architecture of the personalized recommendation system

As shown in Figure 1, Hadoop cluster data processing center is responsible for storage and processing of user feedback information; Business server cluster is the main equipment for the business management of the whole system which periodically transmits the data content of the terminal equipment to the Hadoop processing center via the network; Push server cluster transmits recommendation and private messages to the target user at regular intervals.

Recommended algorithm module

1.1 User-Based collaborative filtering algorithm

The user-based collaborative filtering algorithm is the earliest, most widely used and successful recommendation algorithm. In this algorithm, users who have similar interest as user A are found firstly. Then, the items which these users like and are not known to user A are recommended to user A:

(1) Find the user cluster which has the similar interest as the target user.

(2) Recommend items which have been evaluated by users with similar interests to the target user, expressed as:

$$p(u,m) = \sum_{v \in S(u,K) \cap N(m)},$$

where $S(u,K)$ refers to K users who have similar interest as user $u, N(m)$ refers to the user cluster which evaluates item m, w_{uv} refers to the interest similar degree between user u and user v, and r_{vm} refers to the interest of user v to item m.

1.2 Item-Based collaborative filtering algorithm

The item-based collaborative filtering algorithm is based on the assumption that the items similar to the ones favored by a user will also be favored by the user. The algorithm has the similar recommendation procedures as the user-based recommendation algorithm, including similarity degree calculation and score prediction. There are three commonly used predictive models for the item-based collaborative filtering algorithm:

(1) Weight similarity calculation method

$$P_{ui} = \frac{\sum_{j \in I_{neighbor}} sim(i,j) \cdot R_{uj}}{\sum_{j \in I_{neighbor}} sim(i,j)}.$$

(2) Return model prediction method

$$\overline{R}_j = \alpha \overline{R}_i + \beta + \varepsilon.$$

(3) Park adopted prediction method

$$P_{ui} = \overline{R}_i + \frac{\sum_{j \in I_{neighbor}} sim(i,j) \cdot (R_{uj-\overline{R}_j})}{\sum_{j \in I_{neighbor}} (|sim(i,j)|)},$$

where P_{ui} refers to the predicted score of user u to item i, $I_{neighbor}$ refers to item i's neighborhood item cluster, R_{uj} refers to the score of user u to item j, and R_i and R_j refers to the average values of the score clusters of item i and j.

1.3 Hybrid recommendation algorithm

Many mathematical algorithms can solve some professional problems in life [10, 11]. Hybrid recommendation algorithm is introduced in order to solve the cold start and data sparse problem in the traditional collaborative filtering method. It combines user-based algorithm with item-based algorithm and can calculate similarity more accurately. When predicting scores, it considers both control factor and balance factor and carries out recommendation based on the comprehensive results. To obtain more accurate predicting results, the control factor $\lambda(0 \leq \lambda \leq 1)$ was combined with the balance factors m_u and m_i.

The calculation formula of the balance factor m_u is

$$m_u = \sum_{u_m \in N(u)} \frac{(sim(u_m, u))^2}{\sum_{u_m \in N(u)} sim(u_m, u)}.$$

The calculation formula of the balance factor m_i is

$$m_i = \frac{(sim(i_n, i))^2}{\sum_{i_n \in N(i)} sim(i_n, i)}.$$

Based on the combination of the balance factors m_u and m_i, and the control factor λ, parameters a_u and a_i were added. The definitions of the two parameters were:

$$a_u = \frac{m_u \times \lambda}{m_u \times l + m_i \times (1 - \lambda)},$$
$$a_i = \frac{m_i \times (1 - l)}{m_u \times \lambda + m_i \times (1 - l)}.$$

It can be known from the above formulas that $a_u + a_i = 1$.

When neither the user neighbor cluster $N(u)$ nor the item-based neighbor cluster is an empty set, the score is:

$$P(r_{u,i}) = t_u \times P_u(r_u, i) + P_i(r_u, i).$$

1.4 Evaluation indicators of the recommendation system

To determine whether a recommendation system meets the requirements, normally three indicators, i.e. precision indicator, recall indicator and MAE (mean absolute error) indicator, are applied [12]. The precision indicator is a basic and commonly used indicator to evaluate the recommendation system currently.

The precision rate of the Top-N recommendation system can be determined by precision and recall indicators [13], and its definition was:

$$Precision = \frac{\sum_{u\in U} |Re(u) \cap Te(u)|}{\sum_{u\in U} |Re(u)|},$$

where $Re(u)$ refers to the linked list recommended based on the user training set, and $Te(u)$ refers to the linked list recommended based on the user test set

$$Recall = \frac{\sum_{u\in U} |Re(u) \cap Te(u)|}{\sum_{u\in U} |Te(u)|}.$$

The above two formulas suggested that there is mutual effect between precision and recall, one rises and the other decreases. In practical situation, a comprehensive evaluation indicator which combines both indicators should be applied.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall},$$

$$Emeasure = \frac{1}{\alpha(1/Precision) + (1-\alpha)(1/Recall)}.$$

The larger the $F1$, the better the recommendation effect.

MAE evaluates the precision of the predication score based on the size of difference between the prediction score and the actual score. For example, the scoring item number of user x in the test item set is T_x, the actual scoring set is $\{x_1, x_2, x_3, \ldots, x_n\}$, and the prediction scoring set is $\{p_1, p_2, p_3, \ldots, p_{r_u}\}$; then the calculation formula for MAEx is:

$$MAE_x = \frac{\sum_{i=1}^{T_x} |u_i - p_i|}{T_x}.$$

For the recommendation system, the MAE calculation result of all users in the system is:

$$MAE = \frac{\sum_{u=1}^{M} MAE_u}{M}.$$

According to the above equation, the greater the deviation of the prediction score from the actual score, the greater the MAE value.

2. Application of personalized algorithms

2.1 The improved collaborative filtering algorithm

In the multi-personalized recommendation system, the collaborative filtering algorithm is one of the most widely used algorithms. The traditional user-based algorithm recommends products to users based on the interest of similar uses on products. The application of the traditional user-based algorithm will greatly reduce the precision because of the multiple categories and wide coverage of

User/Item	Item1 (home products)	Item2 (home products)	Item3 (digital products)	Item4 (digital products)	Item5 (digital products)	Item6 (home products)
User1	3	1	1	3	3	4
User2	3	1	1	3	3	4
User3	3	1	1	3	3	4
User4	1	3	4	3	1	1
User5	2	2	4	3	2	1
User6	3	1	4	3	2	1
User7	3	2	4	4	2	**

Table 1. Cross-border E-commerce scoring table

cross-border E-commerce products. In Table 1, there are 7 users and 6 products. Item 1, 2 and 6 are home products while item 3, 4 and 5 are digital products. Firstly, the traditional collaborative filtering algorithm was applied to predict $R_{7,6}$. The users which were similar to user 7 were selected, i.e. {User4, User5, User6}. Then the interest of user 7 to item 6 was determined based on the interest of these users to item 6, $R_{7,6} = 1$.

The prediction of the interest of user 7 to home products was the target; however the similarity of interest between these users and user 7 obtained before was on digital products, which might lead to the inaccuracy of the prediction values.

To avoid this problem, a user-based multi-interest collaborative filtering algorithm was proposed. Based on the interest of a user on a product, it was known that the similar users were {User1, User2, User3}, and then $R_{7,6} = 4$ was obtained. The proposed algorithm considered the direct correlation between users and products based on different attributes of products, suggesting higher accuracy and practicability.

2.2 The application of the improved algorithm in cross-border E-commerce

2.2.1 Characteristics of cross-border E-commerce

Compared to the traditional E-commerce, cross-border E-commerce requires more effective product recommendation [14] because consumers have higher requirements on products with clearer aim. The consumer group of cross-border E-commerce is global, and the anonymous browsing of products by users can increase difficulty to information collection [15]. Therefore, the traditional E-commerce personalized recommendation method and algorithm are no longer applicable to the environment in which cross-border E-commerce is located.

2.2.2 Cross-border E-commerce user group

The user group of cross-border E-commerce is large, which includes cross-border online shopping users, pragmatism users, fashion users, entry users and potential users. The consumption levels of those users are shown in Figure 2.

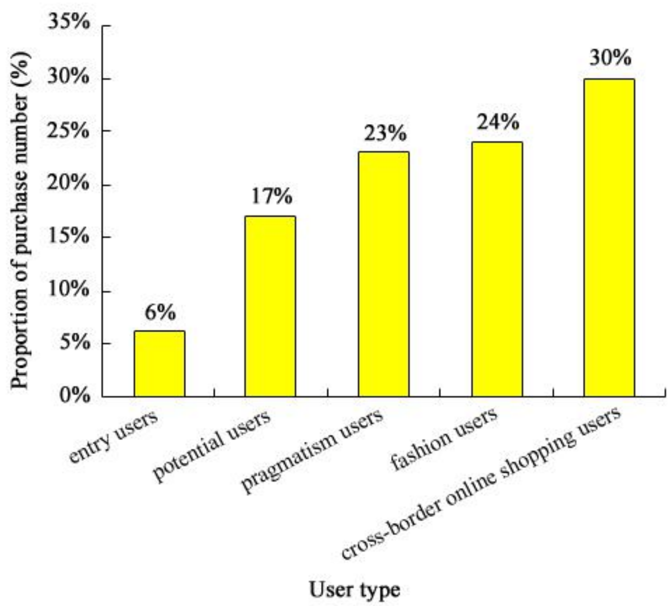

Figure 2: Proportion of consumption number of different user groups

As shown in Figure 3, the contribution of cross-border online shopping users is the highest, which is the reason why mainstream websites take them as key consumers. Besides, pragmatism users rank the second and fashion users the third. Nevertheless, potential users, though have the largest scale, have the smallest contribution. For the prevention of the cold start problem [16], the system will allow customers to choose their own interested category of goods when they log in for the first time, known as explicit interest. In addition, merchants can judge users interest on a product based on the information such as the browsing, forwarding and collecting information of the product of the users. According to different behavior of users, evaluation is made, which is called hidden interest.

To effectively determine the interests of users, we combined explicit interest with hidden interest and obtained a comprehensive calculation formula:

$$R_{u,i} = \lambda Re_{u,i} + \beta Ri_{u,i},$$

where $Re_{u,i}$ is explicit interest and $Ri_{u,i}$ is hidden interest; the range of $Re_{u,i}$ and $Ri_{u,i}$ is between 0-5 level, and the comprehensive interest range of a product is also between 0-5 level.

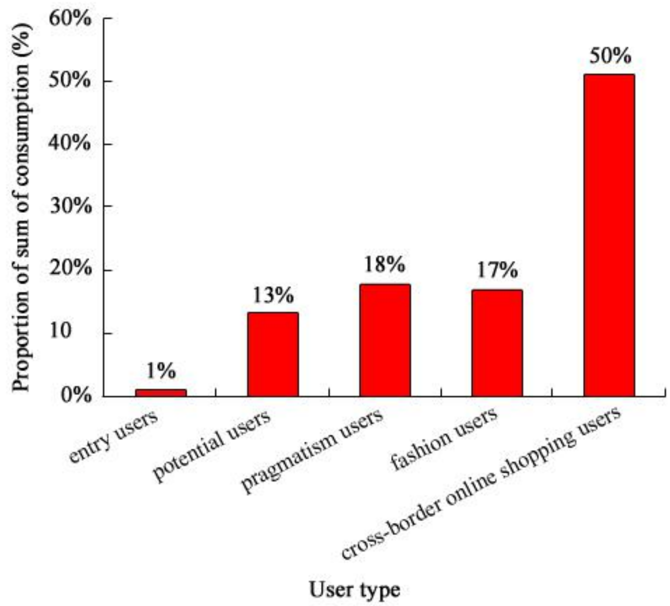

Figure 3: Proportion of sum of consumption of different user groups

According to the formula, recommendation results could be obtained. Firstly, the category of product was determined. Then, all the products belonging to the category were listed. Afterwards, the Top-N user was selected from neighbors by calculating neighbor y that has the similar interest as user u. Then, the interest of user u on the product was calculated:

$$P_{u,j} = \overline{R}_u + \sum_{i \in y} w(u,i)(R_{i,j} - \overline{R}),$$

where $w(u,i)$ refers to the similarity between user u and i. Ranking is performed based on the size of similarity and the Top-N product is recommended to user u.

2.2.3 Effect analysis of improved algorithm

In this study, the data of four months released by an E-commerce corporate was taken as the research subjects.

After processing the data in Table 2, Table 3 was obtained.

Then, the offline results of the data were tested using the aforementioned recall and precision formulas.

As shown in Table 4, the recommendation results of the improved algorithm were better than the results of the traditional algorithm.

	User-id	Brand-id	type	Visit-datetime
1	10944750	13451	0	6.4
2	10944750	13451	2	6.4
3	10944750	13451	2	6.4
4	10944750	13451	0	6.4
5	10944750	13451	0	6.4
6	10944750	13451	0	6.4
7	10944750	13451	0	6.4
8	10944750	13451	0	6.4
9	10944750	13451	0	6.4
10	10944750	21110	0	6.7
11	10944750	1131	0	7.23
12	10944750	1131	0	7.23
13	10944750	8689	0	5.2
14	10944750	8689	2	5.2

Table 2. User behavior data

	User-id	Brand-id	click	fav	addcart	buy
1	10944750	21110	5	0	1	0
2	10944750	25687	6	0	1	3
3	10944750	25372	1	1	0	1

Table 3. Processed data

3. Conclusion

This paper systematically introduced the classification of personalized recommendation systems and analyzed several collaborative filtering algorithms and their evaluation indexes. Besides, the improved algorithm was applied to the product recommendation of cross-border E-commerce. Nevertheless, there are rooms for further improvement and research, which is expected to be realized in future studies.

References

[1] W. Wang, *Chinese cross-border electricity supplier logistics development analysis*, Modern Economy, 7 (2016), 875-880.

[2] W. Wang, *Comparative research on states of cross-border electricity supplier logistics*, Ibusiness, 6 (2016), 726-730.

[3] S. Song, K. Wu, *A creative personalized recommendation algorithm-User-based Slope One algorithm*, International Conference on Systems and Informatics, IEEE, 2012, 2203-2207.

[4] X. Zhou, J. He, G. Huang et al., *A personalized recommendation algorithm based on approximating the singular value decomposition (ApproSVD)*,

Algorithm	Precision rate	Recall rate
Traditional User-based	5.35%	5.45%
Improved User-based	5.75%	5.80%

Table 4. Results of two algorithms

IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, IEEE, 2012, 458-464.

[5] D. Feinleib, *Big data resources*, Big Data Bootcamp, Apress, 2014, 205-208.

[6] S.J. Gong, *Personalized recommendation system based on association rules mining and collaborative filtering*, Applied Mechanics & Materials, 39 (2011), 540-544.

[7] L. Zhang, X. Deng, D. Lei, *Collaborative filtering recommendation algorithm based on user interest characteristics and item category*, Journal of Computational Information Systems, 9 (2013), 5973-5986.

[8] S. Gong, *Learning user interest model for content-based filtering in personalized recommendation system*, International Journal of Digital Content Technology & Its Application, 6 (2012), 155-162.

[9] Z.D. Burgos, *Meta-Mender: a meta-rule based recommendation system for educational applications*, Procedia Computer Science, 1 (2010), 2877-2882.

[10] A. Freihat, M. AL-Smadi, *A new reliable algorithm using the generalized differential transform method for the numeric analytic solution of fractional-order Liu chaotic and hyperchaotic systems*, Pensee Journal, 75 (2013), 263-276.

[11] O.A. Arqub, M. Al-Smadi, *Numerical algorithm for solving two-point, second-order periodic boundary value problems for mixed integro-differential equations*, Applied Mathematics and Computation, 243 (2014), 911-922.

[12] C. Porcel, A. Tejeda-Lorente, M.A. Martinez et al., *A hybrid recommender system for the selective dissemination of research resources in a Technology Transfer Office*, Information Sciences, 184 (2012), 1-19.

[13] P. Cremonesi, R. Turrin, R. Turrin, *Performance of recommender algorithms on top-n recommendation tasks*, ACM Conference on Recommender Systems, ACM, 2010, 39-46.

[14] P. Chen, *Four major bottlenecks in cross-border electricity suppliers need to be resolved*, 3 (2015), 668.

[15] W. Wang, *Analysis and evaluation of chinese cross-border electricity supplier logistics*, Open Journal of Business & Management, 04 (2016), 500-504.

[16] J. Bobadilla, F. Ortega, A. Hernando et al., *A collaborative filtering approach to mitigate the new user cold start problem*, Knowledge-Based Systems, 26 (2012), 225-238.

Accepted: 31.03.2018