*Journal of Computing and Electronic Information Management*

ISSN: 2413-1660

# Personalized recommendation algorithm based on user behavior

Yi Zhang [a], Jin Deng, Huang He, Shuo Yang

School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, China

[a]1839583689@qq.com

**Abstract:** The traditional recommendation algorithm blindly emphasizing the accuracy and neglecting the diversity of the recommended table which cause the problem of over-one-sidedness and homogeneity, this paper presents a personalized recommendation algorithm based on user behavior. First of all, extract user characteristics trough learning their history data, build a real user behavior model. Then, using the Pearson similarity algorithm to calculating the similarity of the users, forming the high-clustering of the target user. Finally, using association rules combine with the interest of user to mining and generate top-N recommendation list to complete the recommendation process. The results shows the diversity of recommendation turned to be good on the condition of the accuracy is guaranteed and the similarity between the target user is high, that when the target users share the high similarity , the diversity of the users can be obtained with the assurance of accuracy, and the performance of the algorithm is higher than the traditional ones.

**Keywords:** Diversity, Accuracy, User Behavior, Association Rules, Recommendation Algorithm.

## 1. Introduction

With the accelerated development of electronic commerce, information overload of e-commerce websites has been caused, which has brought difficulties to users. In order to facilitate users to complete online shopping, save users online shopping time and effort, while increasing the user shopping experience, shorten the purchase decision cycle, improve the browsing click conversion rate [1-2]. In recent years, various recommendation systems have emerged. They have received unanimous attention from the academic community and businesses.

At present, there are various methods for recommending research directions and certain results have been achieved. After a lot of investigation and analysis, there are still many problems. In recent years, most of the recommendation algorithms aim to improve the accuracy of recommendations, thereby ignoring the diversity of the recommended algorithms, resulting in the unilateralization and homogenization of recommendation results [3-4].

Blindly emphasizing accuracy is unwise because it can lead users to obtain a series of "accurate recommendations" with a zero amount of information. Hu Rong et al. studied the relationship between the diversity of recommendations and user satisfaction. The results show that the diversity of recommendations has an important impact on user satisfaction [5]. For this reason, this article conducts research on this issue.

## 2. Research status

In recent years, some scholars have conducted related research and exploration and achieved certain results. Research can be broadly divided into four categories: information physics, secondary optimization, social networking, and time perception [6].

In terms of information physics methods: Relevant investigators introduced the theory of material diffusion and heat conduction in physics into the recommended research, constructed a two-part network of item-items or user-items, and then used its ideas to calculate the heat of items in two networks [7- 8]. If Jambor T considers that the collaboration tag contains rich information about the preferences and the content of the project, a substance diffusion algorithm based on three parts of the user project tag is proposed [9].

In terms of the secondary optimization method: In order to make the candidate list of candidates highly diverse, most researchers have adopted a heuristic strategy to re-optimize the recommendation list. At present, there are many optimization strategies in this area, such as objective function optimization, user similarity power-law adjustment, and user model separation [10-12]. For example, Zhang Z K et al by defining the diversity and utility values, measure the similarity between items in the candidate set with valid values, and then convert it into an integer programming solution with constraints [13].

Social network methods: Researchers establish user-character models through sociological methods. They emphasized that social relations are more important, and they draw user social activities from various social networks or platforms to explore user relationships [14]. Time perception: Researchers in this area regard time as an important situation, consider the impact of time on user behavior, use a dynamic thinking to study user behavior, and implement accurate and diversified recommendation strategies [15]. According to the way of using time, it can be divided into continuous time perception, classification time perception and adaptive time perception method.

The analysis finds that the degree of network node needs to be calculated repeatedly in the information physics method, and feature label extraction is difficult. In the second optimization method, the recommended list obtained by the traditional algorithm is mostly optimized twice. The diversity of the algorithm is still essentially that of the traditional. The algorithm decides that social network method needs to acquire a large number of diverse data in the construction of social relationship model, but it is difficult to acquire and the adaptability of the algorithm is poor. In time perception method, the time context dynamically tracks the user interest, but the user behavior is constrained by many factors. . Therefore, the current research on diversity algorithms is still not ideal. In addition, some studies require the participation of multiple users, which makes this part of the study lack user independence.

## 3. Method of this article

Association rule is a simple and commonly used diversity recommendation algorithm. It uses the three indicators of support, confidence and promotion to describe the rules. First, find all item sets whose frequency is greater than the preset threshold of minimum support degree from the data set and call these itemsets frequent item sets; then, generate association rules from frequent item sets and calculate the confidence degree of the rules. The retention confidence is greater than Minimum Confidence Preset threshold rules.

Further studies have found that when looking at frequent item sets in a centralized manner for the entire transaction, the overall volume of the data sets formed by the entire transaction sets is large, which makes searching extremely difficult. In addition, there are users in the data set that have low similarity or interest with the target user. The injection of such user transaction data will dilute the effective transaction data and ultimately affect the accuracy of the recommendation.The research recommendation list found that when the similarity between items in the list is high, the recommended diversity performance is not good.For this reason, this article studies this and improves it.

The user history data truly records the online shopping user behavior information. Analyze user history data and extract data that can objectively characterize users. Then it is subjected to statistical discrete quantization to obtain a set of user eigenvalue sequences, then the eigenvalue sequence identifies the user.

Definition Let user U be identified by eigenvalues, and eigenvalue $x_i$ is a eigenvalue of user U. Then user U can be identified by the vector formed by this eigenvalue, denoted as .

$$\vec{U} = (x_1, x_2, \cdots, x_n)$$

Based on the above model, with the target user as the center, the similarity with other user behaviors is calculated. A user who has selected a similarity not less than the preset threshold is added to a high-cluster cluster formed centered on the target user. This article uses Pearson correlation coefficient method to measure the similarity between two users, the formula is as follows:

$$\rho_{u1u2} = \frac{Cov(u_1, u_2)}{\sqrt{D(u_1)}\sqrt{D(u_2)}} = \frac{E((u_1 - E(u_1))(u_2 - E(u_2)))}{\sqrt{D(u_1)}\sqrt{D(u_2)}}$$

Algorithm Description:

Input: Target users and other user transaction data

Output: Target user's product recommendation list

(1) Clean user transaction data, and perform data format conversion (in order to facilitate feature extraction as a guideline storage).

(2) Iterate through the transaction database and calculate the characteristics of all users based on the feature extraction formula.

(3) According to the distribution status, determine the level of quantitative standards, build a user feature vector model.

(4) Calculate the similarity between the target user and other users based on the eigenvector Sim. If Sim is greater than the preset threshold, join the target user class. Otherwise, discard the user and do not join.

(5) The search data set is formed by the target user class, and the frequent itemsets are searched from the data set with the support degree as a screening condition; association rules are generated accordingly, and the retention rules are screened with confidence and degree of promotion.

(6) In combination with the user's degree of interest in the item, the retention rule item constitutes the candidate recommendation list.

(7) Using the cosine similarity method to calculate the similarity among candidate recommended list items, and sort them by similarity degree ascending power, and finally select the top-N items as recommendation list.

## 4. Experiment

### 4.1 Data description

In order to verify the theory proposed in this paper, we applied for the raw data from Taobao's API interface. One type of data records the transaction data of 10,000 users in the month from November 18, 2014 to December 18, 2014, totaling 12 million records. Among them, the volume of goods is as high as 2.8 million, and the category of commodities is 8916. The data sample is as follows:

| user_id | item_id | item_category | behavior_type | user_geohash | time |
|---|---|---|---|---|---|
| 98047837 | 232431562 | 4245 | 1 | | 2014/12/6 02 |
| 97726136 | 383583590 | 5894 | 1 | | 2014/12/9 20 |
| 98607707 | 64749712 | 2883 | 1 | | 2014/12/18 11 |
| 98662432 | 320593836 | 6562 | 1 | 96nn52n | 2014/12/6 10 |
| 98145908 | 290208520 | 13926 | 1 | | 2014/12/16 21 |
| 93784494 | 337869048 | 3979 | 1 | | 2014/12/3 20 |
| 94832743 | 105749725 | 9559 | 1 | | 2014/12/13 20 |
| 95290487 | 76866650 | 10875 | 1 | | 2014/11/27 16 |
| 96610296 | 161166643 | 3064 | 1 | | 2014/12/11 23 |
| 100684618 | 21751142 | 2158 | 3 | | 2014/12/5 23 |
| 100509623 | 266020206 | 4923 | 3 | tfvomgk | 2014/12/8 17 |
| 101781721 | 139144131 | 3424 | 1 | 9rgt162 | 2014/12/13 21 |
| 103891828 | 255365467 | 552 | 1 | 96sjmho | 2014/12/1 13 |
| 101260672 | 212072908 | 10984 | 1 | 95q0is4 | 2014/12/12 11 |

Note: The above data is the original data after desensitization processing, and the recording time is accurate to hours.

## 4.2 Feature extraction

According to the mastered data combined with the research objectives, select five features: user's weekly traffic, midnight average, search accuracy, purchase index and click conversion rate. Use C language to connect SQL sever programming to achieve feature extraction. Among them, weekly clicks reflect the user visit frequency characteristics, purchase index reflects the user's expected purchase expectations, midnight averages generally reflect the user's online shopping time tendencies, search accuracy reflects the characteristics of user shopping screening behavior, click conversion rate is certain The degree reflects the characteristics of the user's purchase decision. Therefore, this paper selects these five features.

1.Weekly traffic The weekly visits are counted on a weekly basis, and the weekly traffic average is calculated accordingly.

$$Ew = \frac{1}{n}\sum\left(\sum_{i=1}^{7} D_i\right)$$

Where $D_i$ is the number of hits on the $i$-th days of the week

2. Midnight than midday time: 10 - 14h, night time: 18 - 24h. Count the number of visitors during the midnight and night time segments in days, and then calculate the average number of visits, and finally get the average ratio of midnight.

$$NI = \frac{1}{k}\sum_{i=1}^{k}\left(N_i / E_i\right)$$

Where $N_i$ and $E_i$ are daily lunch and night traffic

3. Search Accuracy The comparison screening is generally performed before the user purchases, and the search accuracy can reflect this feature of the user. Firstly, each search accuracy is obtained, that is, the number of visits per week before each single order; then, the expectation of the user's search accuracy is calculated.

$$SA=\frac{1}{k}\sum_{i=1}^{k}Et_i$$

Where $Et_i$ is the $i$-th single search precision

Purchase index From the data, it can be seen that there are four kinds of behaviors for the user, namely, clicking, collecting, adding to the shopping cart and purchasing. According to the objectives of this study, the four types of behavior are given weight distributions with $\alpha_1=0.1,\alpha_2=0.2,\alpha_3=0.3,\alpha_4=0.4$ respectively. Then calculate the user's weekly purchase index.

$$BI=\frac{1}{n}\sum\left(\sum_{i=1}^{7}\left(\alpha_1C_{1i}+\alpha_2C_{2i}+\alpha_3C_{3i}+\alpha_4C_{4i}\right)\right)$$

5. Click conversion rate Users click to visit, but do not necessarily translate into purchases. Such visits are ineffective for e-commerce. For this reason, it is of great value to study the click conversion rate. According to the importance of collecting, adding shopping carts, and purchasing to e-commerce, weight distribution can be given to $\beta_1=0.2,\beta_2=0.3,\beta_3=0.5$.

$$CR=\frac{1}{n}\sum\left(\sum_{j=1}^{7}\left(\beta_1C_{2j}+\beta_2C_{3j}+\beta_3C_{4j}\right)\Big/\sum_{j=1}^{7}C_{1j}\right)$$

Among them, c1, c2, c3, and c4 indicate the number of hits, collections, shopping carts, and purchased visits per day.

## 4.3 Evaluation standard

### 4.3.1 Accuracy

At present, there are many evaluation criteria for the evaluation of recommendation accuracy. Mainly include accuracy rate, recall rate and so on. This article selects the accuracy rate as the evaluation index, and the accuracy rate can be subdivided into the single user accuracy rate and the system accuracy rate. The calculation formula for the accuracy of single user and system is: $P(Lu)=(Lu\cap Bu)/Lu$ and $P_L=\frac{1}{n}\sum P(Lu)$.

### 4.3.2 Diversity

Diversity includes both individual and system diversity. The individual diversity was evaluated by the ILS method proposed by Ziegler et al.: The individual diversity was determined by the similarity between individual items on the list of individual users, $ILS(L(u))=\sum_{i,j\in Lu,i\neq j}Sim(i,j)\Big/(|Lu|(|Lu|)-1)$. The system diversity uses the Entropy evaluation index, which measures the distribution of articles recommended to all users and reflects the uniformity of all recommended articles, $Diversity=-\sum_{i=1}^{n}\left(\frac{rec(i)}{total}\right)\ln\left(\frac{rec(i)}{total}\right)$.

Explanation of symbols:

| Symbol | significance | Symbol | significance |
|---|---|---|---|
| U | User set | I | Item set |
| Lu | user u recommendation list | Bu | user u positive feedback item |
| rec($i$) | The number of users who get the recommended item $i$ | | |
| n | the total number of candidate items that can be recommended | | |
| total | Total number of systems recommended to all users top-N | | |

### 4.4 Result analysis

In order to reduce the influence of measurement error and calculation of rounding error on the experiment, this paper discretely quantizes the user's characteristic distribution based on the study and the related data. The grading standards are shown in Table 1.

Table 1 Classification Standards

| Type | Min level | IF and Low frequency | High frequency | Others |
|---|---|---|---|---|
| Ew | 0 | Grade spacing 100 | > 2000 | |
| NI | 0 | Grade spacing 1 | > 20 | ∞ visit in the afternoon |
| SA | 0 | Grade spacing 50 | > 1200 | ∞ has no purchase |
| BI | 0 | Grade spacing 10 | > 200 | |
| CR | 0 | Grade spacing 0.05 | > 1.6 | |

#### 4.4.1 K Nearness Similarity and Accuracy

In this experiment, the relationship between the target user clusters formed by different User Similarity and the impact on recommendation accuracy is mainly studied. In the experiment, the above five characteristics were selected to construct the user model, and the Pearson similarity method was used to calculate the user similarity. The experimental results are shown in Figure 1.

As can be seen from Figure 1, the accuracy increases approximately with the increase in user similarity in the target user cluster. When the similarity increases rapidly around 0.4 - 0.7, the growth slows down to a steady level. This shows that better accuracy can be obtained under higher similarity, which proves that our algorithm can guarantee better recommendation accuracy.

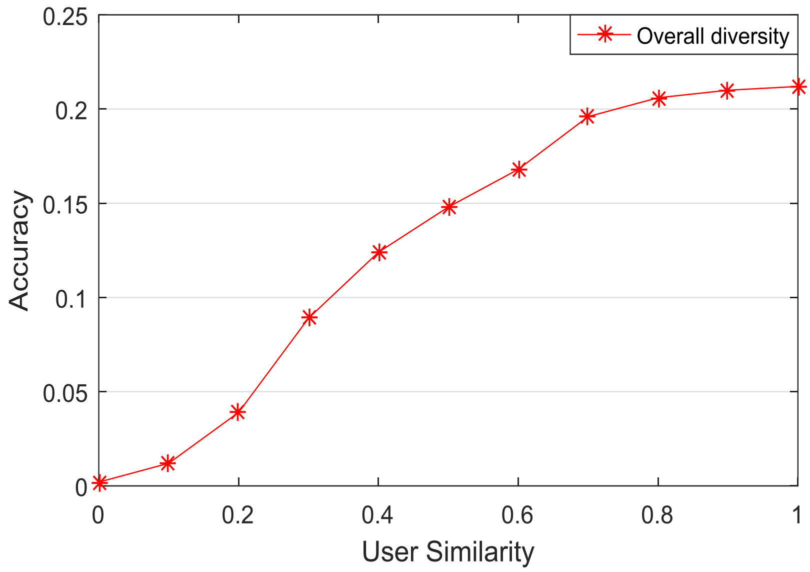


Figure 1 Cluster user similarity and accuracy diagram

### 4.4.2 Diversity and accuracy

In this experiment, we mainly study the relationship between the diversity and accuracy of the algorithm when the target cluster user similarity is different, and compared with the Apriori algorithm. The experimental results are shown in Figure 2.

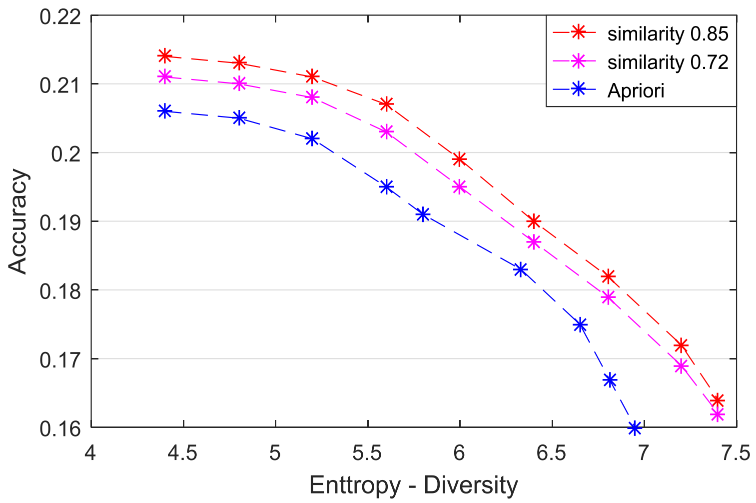


Figure 2 Relationship between diversity and accuracy

From Figure 2, we can see that the accuracy of the recommendation decreases with the increase of diversity, that is, there is a contradiction between the two. The research also shows that when the diversity index is consistent, higher user similarity can obtain better recommendation accuracy, and the recommendation accuracy and diversity performance are superior to the traditional Apriori algorithm when the user similarity is high.

4.4.3 Relationship between recommendation list length and diversity

In this experiment, the candidate candidate set is obtained under the condition that the user similarity is high, and then the length of different recommended table columns is selected to study its influence on the algorithm diversity. The experimental results are shown in Figure. 3 .

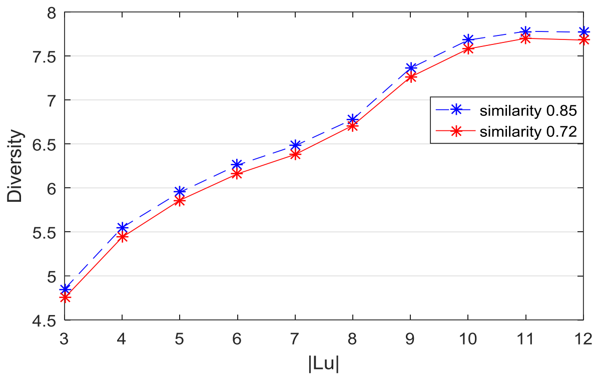


Figure 3 Relationship between recommended table length and diversity

As can be seen from Figure 3, as the length of the recommended table increases, the value of the diversity index becomes larger. At higher levels of similarity and when the recommended list length is fixed, small-scale floating similarity has little effect on the diversity index. When the list of recommendations increased to about eight, the increase in diversity tends to be flat. This shows that it is not possible to obtain continuous improvement of the diversity index by increasing the list of recommendations indefinitely. It is generally recommended that the length of the recommended list be selected to be 8 to 12 or so (short range diversity index is not good, high variety is not obvious, and the accuracy of the recommendation is affected).

## 6. Conclusion

This article focuses on the recommended diversity (individualization) problem. In order to avoid excessive "precise" homogeneity, thus affecting the recommended user experience and satisfaction, the recommendation diversity can be appropriately increased to improve. After analyzing the existing related researches, a personalized recommendation algorithm based on user behavior is proposed. The experimental results show that the target user clusters have higher similarity levels, and the algorithm can obtain better accuracy and diversity. It is also superior to the traditional Apriori algorithm.

For the next step, it is planned to further study the quantitative relationship between the influence of multi-feature intersection on the user's behavior contribution, so as to establish a model with hierarchical relationships to describe the user more stereoscopically and accurately, so that the algorithm obtains better performance. On the other hand, plans to acquire more data sets to verify the algorithm and further explore the robustness of the algorithm.

## Acknowledgements

Thanks to Taobao's support for the study of this article, which provides the data needed for the experimental study. At the same time, I would like to thank you for the valuable suggestions I gave during the research process.

## References

[1] Lu Zhou, Cheng Jing, Zhang Wei. Collaborative Filtering Recommendation Algorithm Based on User Interest Model Clustering[J]. Microcomputer Information, 2010, 26(33): 235-236.

[2] Zhang Fuguo, Zhang Fuguo. A Survey of Personalized Project Recommendation System Based on Tags[J]. Journal of the China Society for Scientific and Technical Information, 2012, 31(9): 963-972.

[3] Zhang F. Improving recommendation lists through neighbor diversification[C]// IEEE International Conference on Intelligent Computing and Intelligent Systems. 2009:222-225.

[4] Chakraborty J, Verma V. A survey of diversification techniques in Recommendation Systems[C]//Data Mining and Advanced Computing (SAPIENCE), International Conference on. IEEE, 2016: 35-40.

[5] Hu Rong,Pu P. Helping users perceive recommendation diversity[C]//Proceeding of the Workshop on Novelty and Diversity in Recommender Systems.New York: ACM,2011:43-50.

[6] An Wei, Liu Qihua, Zhang Liyi. Research Progress of the Diversity of Personalized Recommendation System [J]. Library and Information Service, 2013, 57(20): 127-135.

[7] Cai Z. Personalized Recommendation Systems Based on Users-Products-Tag Ternary Relation[J]. Value Engineering, 2012.

[8] Liu J G, Zhou T, Guo Q. Information filtering via biased heat conduction[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2011, 84(3 Pt 2):037101.

[9] Jambor T, Wang J. Optimizing multiple objectives in collaborative filtering[C]//Proceedings of the fourth ACM conference on Recommender systems. ACM, 2010: 55-62.

[10] Gan M, Jiang R. Improving accuracy and diversity of personalized recommendation through power law adjustments of user similarities[J]. Decision Support Systems, 2013, 55(3):811-821.

[11] Adomavicius G, Kwon Y O. Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques[J]. IEEE Transactions on Knowledge & Data Engineering, 2012, 24(5):896-911.

[12] Zhang M, Hurley N. Statistical Modeling of Diversity in Top-N Recommender Systems[C]// Ieee/wic/acm International Joint Conferences on Web Intelligence and Intelligent Agent Technologies, 2009. Wi-Iat. IEEE, 2009:490-497.

[13] Zhang Z K, Zhou T, Zhang Y C. Personalized recommendation via integrated diffusion on user–item–tag tripartite graphs[J]. Physica A: Statistical Mechanics and its Applications, 2010, 389:179–186.

[14] Cantador I, Castells P. Extracting multilayered Communities of Interest from semantic user profiles: Application to group modeling and hybrid recommendations[J]. Computers in Human Behavior, 2011, 27(4):1321-1336.

[15] Campos P G. Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols[M]. Kluwer Academic Publishers, 2014.