IETI Transactions on Engineering Research and Practice
http://ieti.net/TERP/

2022, Volume 6, Issue 1, 38-45, DOI 10.6723/TERP.202207_6(1).0004

PERSONALIZED PRODUCT RECOMMENDATION BASED ON BIG DATA MINING

Yigang Tang[1,2, a],Xiaolan Xie[1,2,b,*] ,Yunlong Cui[3,c,*],Xinfei Li[1,2,d]
[1]The College of Information Science and Engineering, Guilin University of Technology Guilin 541000, China
[2]The Guangxi Key Laboratory of Embedded Technology and Intelligent System, Guilin 541000, China
[3]Zhejiang Academy of Surveying and Mapping, Hangzhou 310000, China
[a]tyg1069300778@gmail.com, [b]xie_xiao_lan@foxmail.com,[c]605069578@qq.com, [d]2217842045@qq.com
* Correspondence: xie_xiao_lan@foxmail.com,605069578@qq.com

Abstract With the wide application of e-commerce in all walks of life and various fields, the personalized recommendation of e-commerce in the era of big data has also become a focus of attention in today's society. Compared with the previous ones, e-commerce websites have undergone qualitative changes. While providing users with the services and products they need, e-commerce websites make it more difficult for users to quickly and accurately find products that meet their needs in the massive information. Based on the research on big data technology, an extensive collection of users' massive data for analysis, and accurate addition of personalized information recommendation services for different users in the e-commerce service model, through personalized recommendation, e-commerce websites can take the initiative to Users recommend the products they need and the content they prefer. On the one hand, it helps users to quickly and accurately find products that meet their needs and improves the user experience. On the other hand, it also improves the quality of service and improves the competitiveness of enterprises in many markets.

Keywords: Big data; recommended products; data mining; feature selection.

1. INTRODUCTION

In recent years, online shopping on e-commerce websites has gradually become a new shopping habit for people, turning e-commerce website viewers into actual consumers, meeting consumer needs, and enhancing consumer loyalty through different forms are the challenges faced by various e-commerce websites. primary issue. In the process of online shopping, before people finally decide to buy a certain product, they usually leave a lot of behavior information of browsing products on the e-commerce platform, which usually reflects the behavior pattern of users' shopping. Using data mining methods to analyze the user's behavior pattern data is conducive to better understanding the user's shopping habits and tendencies, thus providing the possibility to predict the user's purchasing behavior [1]. Mining the potential value of users' basic consumption information, analyzing massive data, and forming specific personalized recommendation services for each user, brings new marketing models to e-commerce and new market opportunities. It has become the mainstream of the e-commerce industry to provide different personalized recommendations for individual users. Accurately predicting users' shopping behavior through algorithms is of great significance to e-commerce platforms. Through the prediction results, products can be recommended to users in a personalized manner, which can improve users' shopping efficiency, facilitate more transactions, and increase operating income. On the one hand, it helps users quickly and accurately look for products that meet their needs, improving the user experience, on the other hand, it also improves the quality of service and improves the competitiveness of enterprises in many markets. The main content of this paper is the research and application of personalized recommendation based on e-commerce in the

2022, Volume 6, Issue 1, 38-45, DOI 10.6723/TERP.202207_6(1).0004

era of big data, including the relevant basic theories, the service mode of personalized recommendation of e-commerce, and the personalized information service based on big data.

2. RELATED WORK

The Random Forest [2] algorithm proposed by Breiman in 2001 has been very successful as a general classification and regression method. The method combines several stochastic decision trees and aggregates their predictions by averaging, showing excellent performance on data-heavy tasks. Wang Y proposed a fusion model based on a logistic regression algorithm and GBDT algorithm [3], which was used to recommend vertical industry products in a mobile environment, and got good performance. eXtreme Gradient Boosting (XGBoost) [4] is an ensemble learning algorithm based on gradient Boosting, its principle is to achieve accurate classification effect through iterative calculation of weak classifiers. This paper introduces XGBoost into the product recommendation algorithm of e-commerce, mines the user's behavior data information on the e-commerce platform, and establishes a classification prediction model, to recommend products to users individually. Omar Hasan [5] and others mainly focus on the research on user portraits. They believe that by building a user's basic information database, the user's personalized characteristics can be accurately positioned from multiple perspectives. Unstructured processing can analyze the potential consumption behavior of users and provide accurate service recommendations. Stanescu [6] and others focus on better capturing key information through user tags. The main analysis data is based on the user's consumption data, and the user's consumption behavior is depicted in the user's evaluation information and commodity retrieval records. Pena [7] and others mainly studied automatic user analysis technology. The basic process is to use various articles that users have published, and use words that can express features such as semantic nature to mark the basic characteristics of users, to analyze the user's basic characteristics. User groups are classified, and then personalized recommendations are made for related products and services.

3. METHODOLOGY

3.1. Outlier Removal

The existence of outliers usually seriously affects the quality of modeling and prediction, so it is necessary to remove outliers existing in the data. There are shopping festivals in the acquired data. The total number of users browsing, favorites, adding to shopping carts, and purchasing on that day are 1.4, 1.5, and 2.4 times the average of the previous days, which are obvious outliers. Therefore, all the data of the day will be processed in the follow-up process and eliminated, and some data with null values also need to be eliminated. In addition, users who have no purchase records within one month may not have online shopping habits, and such users have no reference value for predictive modeling, so much data are also excluded.

3.2. Feature Selection

Raw data cannot be used directly for modeling, so it needs to be reduced to statistical features. Feature screening needs to be able to fully describe product information, user information, and user-product interaction. Therefore the features we use are listed in Table 1. In Table 1, the product

2022, Volume 6, Issue 1, 38-45, DOI 10.6723/TERP.202207_6(1).0004

features mainly reflect the popularity of the product. Generally, products with high viewing and purchase times have higher cost performance, so they can attract users to buy. User characteristics mainly reflect the user's shopping habits, such as their shopping frequency and whether the user prefers impulse shopping or repeated hesitation before purchasing. Interaction features take into account the interaction between users and products.

3.3. Data Set Partitioning

Usually, in the shopping process, users will choose whether to buy a product after comparing it with similar products, so the total amount of data generated is relatively large, and only some samples are used for modeling in the processing process. Divide the data set into a training set, validation set, and test set, which are 70%, 20%, and 10% of the original data respectively. Use the training set data to train the model, then use the test set data to test the fit of the model, and finally use the validation set. Set data to verify the fit of the model and prevent overfitting.

3.4. XGBoost Algorithm

XGBoost [4] is an implementation of Boosting Tree. Boosting is a very effective ensemble learning algorithm. The Boosting method can convert weak classifiers into strong classifiers, to achieve accurate classification results. XGBoost performs a second-order Taylor expansion on the loss function, using both first-order and second-order derivatives, and introduces a regular term suitable for tree models to control the complexity of the model. The regular term of XGBoost contains the number of leaf nodes of the tree and the L2 square sum of the output scores of each leaf node.

$$\tilde{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \tilde{y}_i^{(t-1)} + f_1(X_i)\right) + \Omega(f_t) + \Upsilon \tag{1}$$

Using Taylor expansion on the above equation can transform the objective function into:

$$\begin{aligned}\tilde{L}^{(t)} &\approx \sum_{i=1}^{n} l\left(y_i, \tilde{y}_i^{(t-1)} + f_1(X_i)\right) + \Omega(f_t) + \Upsilon \\ &= \sum_{i=1}^{n}\left[g_i w_q(x_i) + \frac{1}{2} h_i w_q^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right) w_j^2\right] + \lambda T\end{aligned} \tag{2}$$

Find the partial derivative of ω for the above formula, and then set the partial derivative equal to zero, it can be solved as:

2022, Volume 6, Issue 1, 38-45, DOI 10.6723/TERP.202207_6(1).0004

$$\omega^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{3}$$

Then substitute into the objective function to get:

$$\tilde{L}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{4}$$

Similarly, when selecting the split point, the goal is to minimize the objective function. Assuming that after a certain split point is selected for division, LI and RI are respectively the left and right nodes after division, and $I = I_L \cup I_R$, then the value of the loss function reduction after this division is:

$$L_s = \frac{1}{2}\left[\frac{\left(\sum_{i \in I_L} g_i\right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\sum_{i \in I_R} h_i + \lambda} + \frac{\left(\sum_{i \in I_J} g_i\right)^2}{\sum_{i \in I_J} h_i + \lambda}\right] - \gamma \tag{5}$$

3.5. Other Classification Models

For the convenience of comparison, we also adopted two other general classification methods SVM and random forest comparison. SVM (Support Vector Machine) is a classification model. Its principle is to find a maximum hyperplane in the feature space so that the distance from all samples to the plane is the largest (to find the distance from the sample set to the plane, that is, Find the distance from the nearest sample point to the hyperplane), our learning goal is to maximize this distance. The kernel function can be used to map to a high-dimensional space, and the kernel function can be used to solve nonlinear classification. The classification idea is very simple, that is, maximizing the interval between the sample and the decision surface has a better classification effect, but it is difficult to train large-scale data. The principle of random forest is to randomly build a large number of classification trees, each tree classifies the samples individually, and the final classification result is determined by the respective classification results of each tree through voting. The random forest algorithm improves the accuracy of classification, and the results are robust and easy to tune parameters, but it runs slower.

3.5. Evaluation standard

For commodity recommendation algorithms, we are more concerned about the accuracy of the prediction of positive samples, so the precision, recall, and F1 value of positive sample predictions are used as evaluation indicators, which are defined as follows:

2022, Volume 6, Issue 1, 38-45, DOI 10.6723/TERP.202207_6(1).0004

$$precision = \frac{T_P}{T_P + N_P} \tag{6}$$

$$recall = \frac{T_P}{P} \tag{7}$$

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{8}$$

Among them, P is the total number of positive samples, TP is the number of correctly predicted positive samples, and NP is the number of incorrectly predicted positive samples. At the same time, for the data processing of a large number of samples, the operation speed is also an important evaluation index. This experiment was run on a personal computer (CPU: AMD 4800H 2.3 GHz; RAM: 16 G), and the running time was recorded with the python language time() function.

4. DATA SET

The data for our experiments were collected from Alibaba, which recorded user behavioral data such as clicks, browsing, adding to shopping carts, and payments on the website. Behavioral data contains about 5.8 billion behavioral records, describing user behavior data such as clicks and payments on the website over a period of time. What we need to do is to calculate the products that users may like based on these data, and recommend them to users to view. The descriptions of datasets M and N are shown in Table 1 and Table 2, respectively. As shown in Table 1, users' behaviors toward items can be divided into the following two types: "like" and "dislike". Since what we need to do is to recommend items to users that they might like, we transform this recommendation problem into a classification problem, where we set "dislike" as a negative sample, and add to cart and purchase behavior as a positive sample (like). We analyze and process it according to the data information described in the table, which is expanded in detail in Section 3. Finally, the dataset is divided into three subsets: training set, validation set, and test set to train and evaluate our proposed model.

Table 1. User behavior dataset M description.

Column	Description
User-id	distinguish users
Item-id	distinguish items
Behavior-type	Including click, collect, add-to-cart, and payment
User-geohash	user location when the behavior occurs
Item-category	The category id of the item
time	The time of the behavior

Table 2. Vertical Industry Commodity Dataset Description N.

Column	Description
User-id	distinguish users
Item-geohash	item location, it can be null
Item-category	The category id of the item

IETI Transactions on Engineering Research and Practice
http://ieti.net/TERP/

2022, Volume 6, Issue 1, 38-45, DOI 10.6723/TERP.202207_6(1).0004

5. PERFORMANCE

5.1. Feature Variable Importance Analysis

Through the modeling of XGBoost and random forest, the contribution of each feature variable to the model can be judged, to determine which feature variables have a more significant impact on the user's purchasing behavior. The analysis results are shown in figure 1 and figure 2.

Table 3. Statistics for modeling.

Feature description	Types of features	Feature number
The number of interactions with the product by the user within 1 day	Interactive Features	1-4
The number of interactions of users with similar products in 1 day	Interactive Features	5-8
The number of interactions with the product by the user within 3 day	Interactive Features	9-12
The number of interactions of users with similar products in 3 day	Interactive Features	13-16
Total product views (favorites, add to cart, purchase)	Product features	17-20
The number of browsing (favorites, adding to shopping carts, purchasing) of products in the last 3 days	Product features	21-24
User total browsing (favorite, add to cart, purchase) volume	User Features	25-28
User browsing (favorites, adding to the shopping cart) purchase volume ratio	User Features	29-31

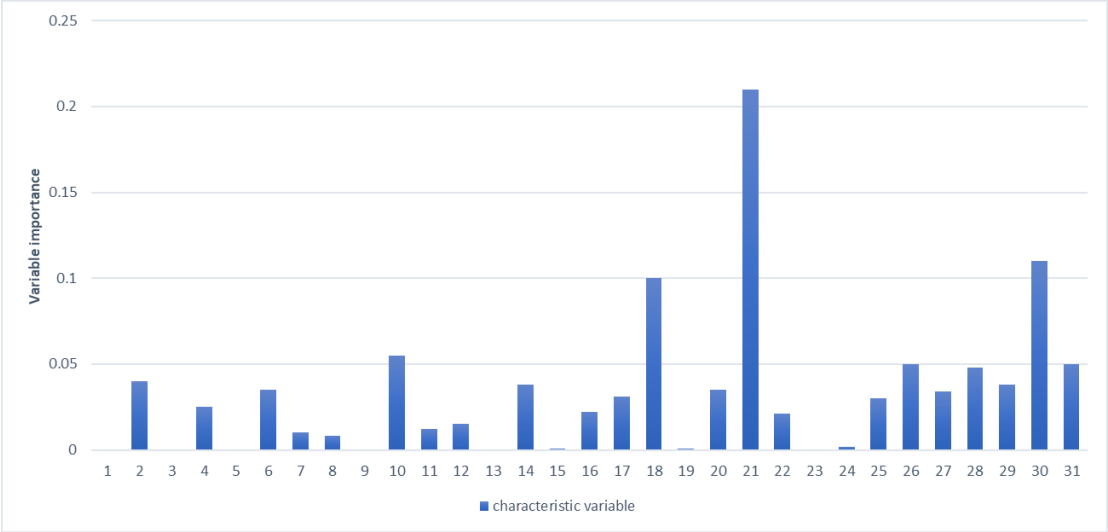

Figure 1. Feature variable importance.

2022, Volume 6, Issue 1, 38-45, DOI 10.6723/TERP.202207_6(1).0004

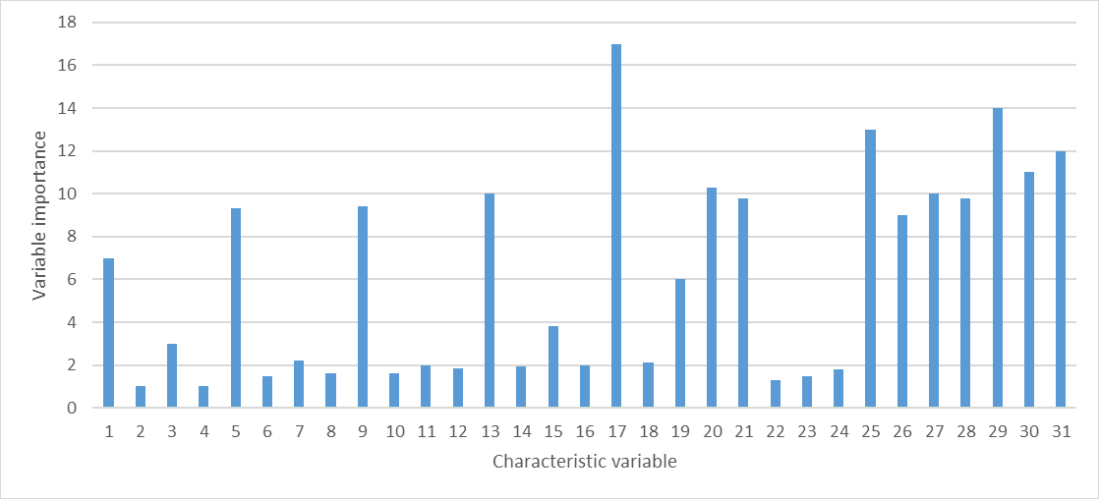

Figure 2. Feature variable importance.

Figure 1 shows the variable importance results of the XGBoost model, and Figure 2 shows the variable importance results of the random forest, where the 17th, 21st, 25th, and 29th variables are important in the two models The rankings are all in the top four, and their corresponding features are the total page views of the product, the page views of the product in the last three days, the total page views of users, and the user browsing/purchase ratio. The variables related to browsing actions have the greatest contribution to the model, because browsing is the most important way for users to interact with products, and its information richness is much higher than other features. Aside from browsing, the features associated with the user's "add to cart" action are the highest, as a user is likely to purchase within the next few days after adding an item to the cart. In the user-product interaction characteristics, we found that the feature variables related to the user's product interaction within three days are not less important than the user's product interaction on the previous day, which indicates that the user has a certain hesitation time before purchasing a product. The importance of the feature variables related to the user's interaction with similar products is slightly higher than the user's interaction with a certain product, which means that when most users buy a certain product, they will fully compare it with similar products, and finally select those browsed and popular items that are purchased more frequently.

In addition, the user features have high positions in the importance ranking, which means that different users have different shopping habits. Therefore, it is very necessary to make personalized recommendations for different users.

5.2. Results Comparison

After optimizing the parameters of the three algorithms through the interactive test, the modeling results were predicted using the test set samples. The results of the three algorithms are listed in Table 4. The results show that among the three classification algorithms, the prediction results of XGBoost are slightly higher than that of random forest, but the running speed is significantly faster than that of random forest. The principle of the classification tree is simple, and the operation speed is close to that of XGBoost, but the operation accuracy is poor. Therefore, XGBoost has the advantages of high accuracy and fast operation speed compared with the other two algorithms.

2022, Volume 6, Issue 1, 38-45, DOI 10.6723/TERP.202207_6(1).0004

Table 4. Statistics for modeling.

Model name	Accuracy	Recall	F1	Operation time
Random Forest	0.97	0.79	0.88	6.16
SVM	0.98	0.84	0.61	2.1
XGBoost	1.0	0.88	0.93	1.11

6. CONCLUSION

In this paper, the XGBoost classification algorithm is used, and feature extraction and classification modeling are carried out based on Alibaba's real user data and compared with random forest and decision tree calculation, more accurate prediction results are obtained. By performing an analysis of variable importance, we identified variables with high contributions to the model. This research allows us to mine the correlation between data and user behavior, which can improve the exposure rate of products. Moreover, the personalized recommendation algorithm can not only increase the user experience, but also bring greater development prospects and huge economic benefits to e-commerce enterprises.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No.61762031), The Science and Technology Major Project of Guangxi Province (NO.AA19046004), and The Natural Science Foundation of Guangxi (No.2021JJA170130).

References

[1]. Curme C., Preis T., Stanley HE., Moat HS., 2014, Quantifying the semantics of search behavior before stock market moves, *Proceedings of the National Academy of Sciences of the United States of America,* 111(32), pp.11600-11605.

[2]. Biau G., Scornet E., 2016, A random forest-guided tour, *Test*, 25(2), pp. 197-227.

[3]. Wang Y., Feng D., Li D., et al., 2016, A mobile recommendation system based on logistic regression and gradient boosting decision trees, *International joint conference on neural networks (IJCNN). IEEE 2016*, pp. 1896-1902.

[4]. Friedman J H., 2001, Greedy Function Approximation: A Gradient Boosting Machine, *The Annals of Statistics*, 29 (5), pp. 1189-1232.

[5]. Hasan O., Habegger B.A., Brunie L. et al., 2013, A Discussion of Privacy Challenges in User Profiling with Big Data Techniques: The EEXCESS Use Case, *IEEE International Congress on Big Data*, pp. 25-30.

[6]. Stanescu A， Nagar S， Caragea D., 2013, A Hybrid Recommender System：User Profiling from Keywords and Ratings, *IEEE/WIC/ACM International Joint Conferences on Web Intelligence（WI）and Intelligent Agent Technologies（IAT)*, pp. 73-80.

[7]. Peñas P., del Hoyo R., Vea-Murguía J., et al., 2013, Collective Knowledge Ontology User Profiling for Twitter - Automatic User Profiling, *IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies（IAT)*, pp. 439-444.