Computing, Performance and Communication Systems (2021) Vol. 5: 30-37
Clausius Scientific Press, Canada

DOI: 10.23977/cpcs.2021.51006
ISSN 2371-8889

Research on e-Commerce User Behavior Analysis Based on Big Data Collaborative Recommendation Algorithm

Xuecong Cao, Sisi Chen, Zhaoming Li*

State Grid Huitongjincai(Beijing) Information Technology Co., Ltd., Beijing, China

**Corresponding Author*

Keywords: Big data, Collaborative recommendation algorithm, Restricted boltzmann machine: e-commerce, User behavior

Abstract: With the rapid development of Internet, users tend to purchase their favorite products through Internet transactions and online payment. The general trend of e-commerce development in China is that physical trading places are gradually replaced by online trading platforms on the Internet. In this paper, the restricted Boltzmann machine based on category conditions is used to describe the user's own interest preference by using the objective label of the project itself. In this process, only the project information that the user has scored is used, which strengthens the user's personalized needs. The method fully mines user behavior information, replaces commodity content big data with user behavior information as a recommended data set, and can actively push commodity content that users may be interested in to users. Experimental results show that the accuracy of RBM (Restricted Boltzmann machine) model with nearest neighbor is higher than that of the original model, and the anti-over-fitting ability of the model is also improved.

1. Introduction

With the rapid expansion of the Internet, the utilization rate of information resources is getting lower and lower, and the problem of “information overload” is becoming more and more serious. Under this background, the recommendation system that mining the information that users are interested in from massive data to meet their personalized needs arises at the historic moment. For consumers of information, in the face of massive Internet information, how to quickly identify and find valuable information for themselves has become increasingly difficult; For information service providers, it is difficult to find potential target users of information, so that more consumers can pay attention to their own information and increase the interest conversion rate of information [1]. Therefore, recommendation algorithm came into being. Using recommendation algorithm to analyze users' historical behavior, alleviate the problem of “information overload” and make personalized recommendations for users is of great significance in the era of big data.

Based on various consumer behaviors on the e-commerce trading platform, this paper analyzes the big data collaborative recommendation algorithm, analyzes online products, consumers' interests, demands and evaluations from the comprehensive perspectives of probability analysis, attribute analysis and commonality analysis, and establishes a data model to conduct in-depth research on consumers' behaviors, so as to know consumers' shopping intentions well and promote the innovative development of the e-commerce industry.

2. Collaborative Filtering Recommendation and Analysis and Modeling of User Behavior Information

2.1 Collaborative Filtering Recommendation

CF(Collaborative filtering), also commonly known as collaborative filtering or social filtering, is based on the assumption that users are similar, and similar users have similar interests and hobbies, and the historical behavior of users represents their interests and hobbies, and judges whether users are similar according to their interests and hobbies, and measures the similarity. For example, in daily life, people often refer to the opinions or behaviors of friends around them to buy some goods or make some choices. In collaborative filtering technology, users are related, they can be friends with each other, and there are neighbor users. According to the principle of sharing interests, the preferences of neighbor users are consistent or similar, so for the current user, the neighbor's preference items are recommended. CF technology finds the neighbor set of a specific user through the preference and rating information of all users and the measurement of user similarity, and makes project recommendation for it according to the interest information of the nearest neighbor.

In 1998, Breese and other scholars divided collaborative filtering algorithms into memory-based algorithms and model-based algorithms [2]. Among them, memory-based collaborative filtering technology can be divided into user-based collaborative filtering technology, project-based collaborative filtering technology and model-based collaborative filtering technology. Commonly used models include Bayesian network, clustering, dimension reduction, linear regression and so on. Fig. 1 is a schematic diagram of cf technology classification.

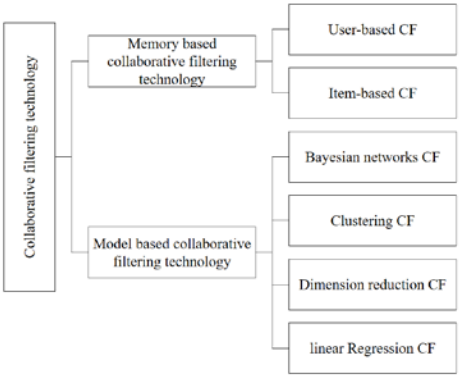
Collaborative filtering technology
Memory based collaborative filtering technology
Model based collaborative filtering technology
User-based CF
Item-based CF
Bayesian networks CF
Clustering CF
Dimension reduction CF
linear Regression CF

Fig.1 Cf Technical Classification

In addition, with the deepening of the research, CF technology which is integrated in various ways is becoming more and more mature, for example, adopting multiple models to float, based on the combination of memory and models, etc., but it still belongs to the above classification after tracing its roots.

2.2 Analysis and Modeling of User Behavior Information

(1) Analysis of user behavior information

The user's behavior information (including browsing record, browsing time, rating and comment, etc.) exists in the form of log on the computer. Literature [3] gives the definition of user behavior in commodity recommendation system: it refers to five aspects: the user who produces behavior, the object of behavior, the kind of behavior, the context of producing behavior and the content of behavior, as shown in Figure 2.

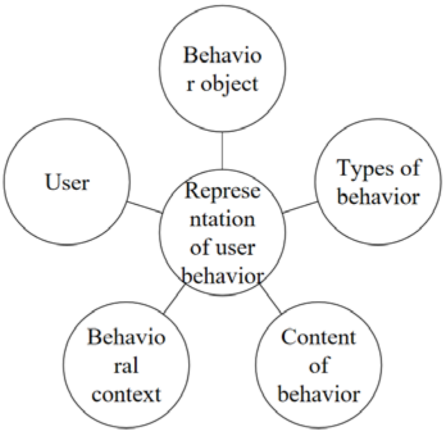

Fig.2 Representation Content of User Behavior

User behavior can be divided into explicit and implicit behavior [4], and explicit user behavior mainly refers to the behavior that users like or dislike when evaluating the quality of goods. For example, some e-commerce websites use the scoring system to evaluate the product content that users have seen, or comment messages in the comment area to express their views on the product content. Implicit user behavior refers to the user behavior that can't directly see whether the user likes or dislikes the commodity content, such as the length, frequency and time period of purchasing the commodity content. Mining hidden user behavior is a difficult point and trend in the analysis and research of user behavior in recent years, which has great potential and value.

(2) User behavior information modeling

The purpose of user behavior information modeling is to find the intrinsic relationship between users and programs, and to establish a model that can reflect users' interest in commodity content [5]. Therefore, a user-content binary association model matrix $M(a,b)$ can be designed. This is a two-dimensional vector representation. The horizontal axis represents users, the vertical axis represents commodity content, and 1 and 0 represent related and unrelated. The following simple algorithm flow can be used to explain whether users are associated with commodity content or not [6-7]. The algorithm flow is as follows.

1) Through the explicit user behavior scoring system, such as the 5-point scoring system, the user scores the product content $\geqslant$ 3, indicating that the user is related to the product, and the output result is 1; Otherwise, continue to execute.

2) To judge the time length of the user's purchase of goods, an intermediate value t can be assumed. When the user's viewing time is greater than or equal to t, there is relevance, and the output is 1; Otherwise, continue to execute the algorithm.

3) Judging the number of times the user purchases goods, assuming a suitable critical number n, when the number of times the user watches is greater than n, it shows that there is relevance, and the output is 1; Otherwise, there is no correlation between the user and the commodity content, and the output result is 0.

3. Collaborative Recommendation Algorithm Based on Big Data

Among the existing recommendation algorithms, one of the model-based collaborative filtering recommendation algorithms is based on the Restricted Boltzmann Machine (RBM). The recommended results of this model algorithm have high accuracy, and the RBM model can be used as the bottom layer of deep learning, which has attracted the attention and research of scholars in recent years. Therefore, this paper will also study how to improve the RBM model algorithm to improve the prediction accuracy of the model.

3.1 Collaborative Filtering Framework Based on Rbm Model

The main problem of applying RBM model to collaborative filtering algorithm is how to effectively deal with unrated items. Literature [8] firstly improves the visible unit of the traditional RBM model, and adopts Softmax unit as the visible layer unit; Then, a special visible unit “Missing” is introduced to indicate that the user is not connected with any hidden unit for items without scoring. The model is shown in Figure 3.

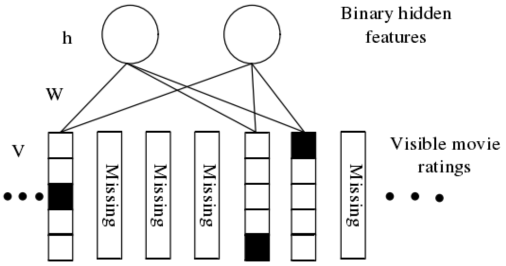

Fig.3 Constrained Boltzmann Machine Applied to Collaborative Filtering

Due to the special structure of the restricted Boltzmann machine, the following conditional independence properties can be obtained: (1) given the visible layer, the hidden layer units are independent; Given the hidden layer, the visible layer units are independent:

$$p(h|v)=\prod_i (h_i|v)$$
$$p(v|h)=\prod_j (h_j|h) \qquad (1)$$

If both hidden layer and visible layer are Bernoulli distribution, the conditional distribution formula can be obtained:

$$p(h_i=1|v)=\sigma(c_i+W_i'v)$$
$$p(v_i=1|h)=\sigma(b_j+W_j'h) \qquad (2)$$

In which $\sigma(r)$ is Sigmoid function.

Therefore, Gibbs sampling can be used to estimate the distribution of restricted Boltzmann machines. In practice, the sampling algorithm is very complex. Usually, it is approximated by k-step sampling without waiting for the sampling convergence. Generally, good results can be achieved by taking $k=1$. This is called the contrast divergence algorithm [9].

The key of applying RBM model to collaborative filtering is how to predict the scores of Missing items. To solve this problem, the above model adopts the method that each user has a separate RBM, all RBM corresponds to a common hidden layer, and the weights between the visible layer and the hidden layer of all RBM and their respective offset items are shared. The method of sharing weights and offsets is to consider that the number of movies scored by each user is far less than the number of all movies, so the number of the same movies scored by different users is less, and the weights of RBM models of different users only overlap in a small part.

3.2 Improved Design of Restricted Boltzmann Machine Based on Class Conditions

3.2.1 Model Structure

RBM training is realized by maximizing the likelihood probability of data, which belongs to unsupervised training. Since it is difficult to calculate the average expectation of model data during RBM training, the literature [10] proposes a Contrastive Divergence (CD) algorithm, which uses Gibbs sampling value as the expected value of the model. Firstly, the CD algorithm performs k state transitions by performing block Gibbs Sampling with each training data as the initial state. Then, the transferred data is used as the estimated mean value of Negative Phase in RBM training to update the parameters.

Experiments show that only one state iteration is needed to ensure the good learning effect of the model. Given the training data $v^{(n)}$, the update of the connection weight W_j is shown in formula (3).

$$\Delta W_j = P\left(h_j = 1 | v^{(n)}\right)\cdot v^{(n)} - P\left(h_j = 1 | v^{(n)-1}\right)\cdot |v^{(n)-} \qquad (3)$$

In which: $\hat{h} - P\left(h|v^{(n)}\right)$ is the posterior activation probability of hidden cell layer h given $v^{(n)}$, and $v^{(n)}$ - $P\left(v^{(n)}|\hat{h}\right)$ is the posterior activation probability of visible cell layer $|v^{(n)-}$ given $\hat{h}$.

Unlike literature [10], which takes category information as an additional visible unit to participate in the model, lCRBM(label Condition RBM) designed in this paper takes label unit as the condition unit of RBM. Therefore, the structure of lCRBM model is shown in fig. 4.

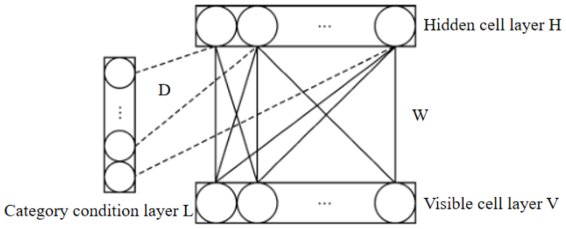

Fig.4 Lcrbm Unit Connection Diagram

It can be seen from fig. 3 that lCRBM consists of three kinds of units, namely visible unit V, hidden unit h and category condition unit L; At the same time, the hidden cell layer, visible cell layer and class condition layer are interconnected, and the cells in the same layer are not connected; W represents the connection weight between the hidden cell layer and the visible cell layer, and D represents the connection weight between the hidden cell layer and the class condition layer.

Given the state (v,h,l) of each unit, the energy definition of lCRBM model is shown in formula (4).

$$E(v,h,l) = -\sum_{i=1}^{N} v_i b_i - \sum_{j=1}^{M} h_j c_j - \sum_{i=1}^{N}\sum_{j=1}^{M} W_{ij} v_i h_j - \sum_{i=1}^{L}\sum_{j=1}^{M} D_{ij} l_i h_j \qquad (4)$$

In which: l is the class condition unit, D_{ij} is the connection weight between hidden unit h_j and class condition unit l_i, and L is the total number of class condition units.

According to the energy function of lCRBM model, the posterior activation probability of visible cells in lCRBM can be deduced as shown in formula (5).

$$p(v_i|h) = \sigma\left(\sum_{j=1}^{M} h_j W_{ij} + b_i\right) \qquad (5)$$

In the same way, the posterior activation probability of hidden cells in lCRBM is shown in formula (6).

$$p\left(h_j \middle| v,l\right)=\sigma\left(\sum_{i=1}^{N} v_i W_{ij} + c_i + \sum_{i=1}^{L} l_i D_{ij}\right) \quad (6)$$

It can be seen from formula (6) that the difference between hidden units in lCRBM and formula (2) is that class units participate in the calculation of posterior activation probability of hidden units, which is also a method for lCRBM to introduce class information into model training, so as to make the training category-specific, weaken the problem of feature homogenization which is easy to occur when RBM unsupervised training, and further improve the data fitting degree.

4. Experimental Analysis

In this experiment, Matlab2015b software was used, and the data set was MovieLens100K data set. The data set is randomly selected and divided. 85% of the data set is divided into experimental training set, and the remaining 15% is divided into experimental test set. Under the condition that the training set and the test set are completely identical, the average root mean square error (lCRBM) is taken as the evaluation index, the standard RBM collaborative filtering algorithm is taken as the comparative reference algorithm of this experiment, and the average of 10 experimental results is taken as the final prediction result.

When training RBM model and the algorithm in this paper, we must first determine the parameters of the model. Literature [8] and Literature [9] introduce the selection of model parameters in detail. In order to ensure the accuracy and comparability of the experimental results, the same experimental parameters are adopted for the algorithm and RBM algorithm in this paper. The determination of parameters in this paper is based on literature [10], and the relative optimal values are verified by practical experiments. The main parameters are shown in Table 1.

Table 1 Setting of Main Parameters of Model

Parameter	Parameter value
Hidden layer node	70
Weight attenuation coefficient	0.0006
Weight learning efficiency	0.001
Visible layer offset learning efficiency	0.001
Implicit layer bias learning efficiency	0.01
Iterations	100
Iteration times of CD algorithm	4

After determining the model parameter values, because the algorithm in this paper contains class conditions, it is necessary to consider the influence of different number of class conditions on the experimental results and find out the optimal number of class conditions. The influence of different nearest neighbors on the algorithm is calculated through experiments, and the calculation results are shown in Figure 5.

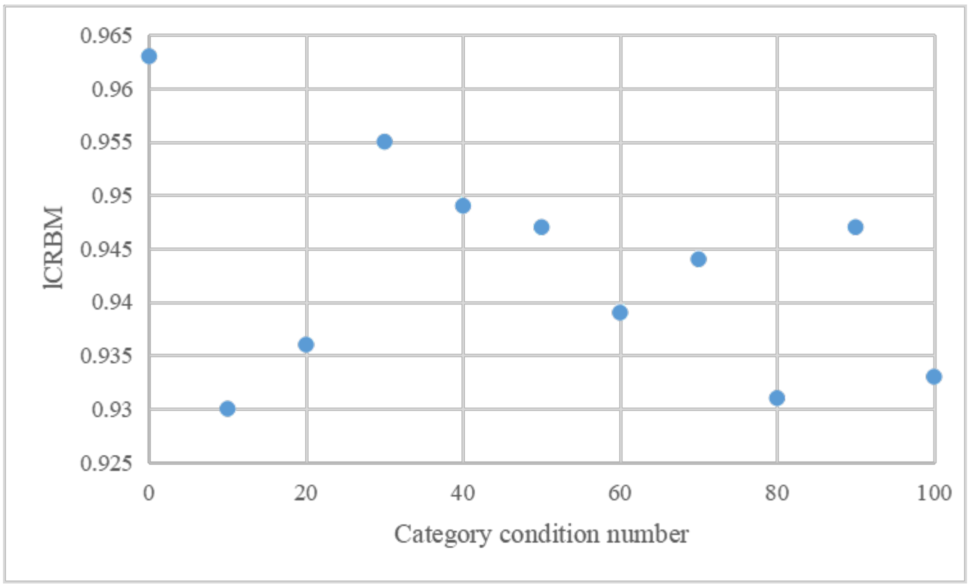

Fig.5 The Influence of Category Condition Number on Lcrbm

It can be seen from fig. 6 that after the number of users' nearest neighbors reaches 10, the influence of the number of nearest neighbors on lCRBM value fluctuates in a small range and tends to be stable basically. Considering the computational complexity and reducing unnecessary errors, the number of category conditions is set to 20 in subsequent experiments.

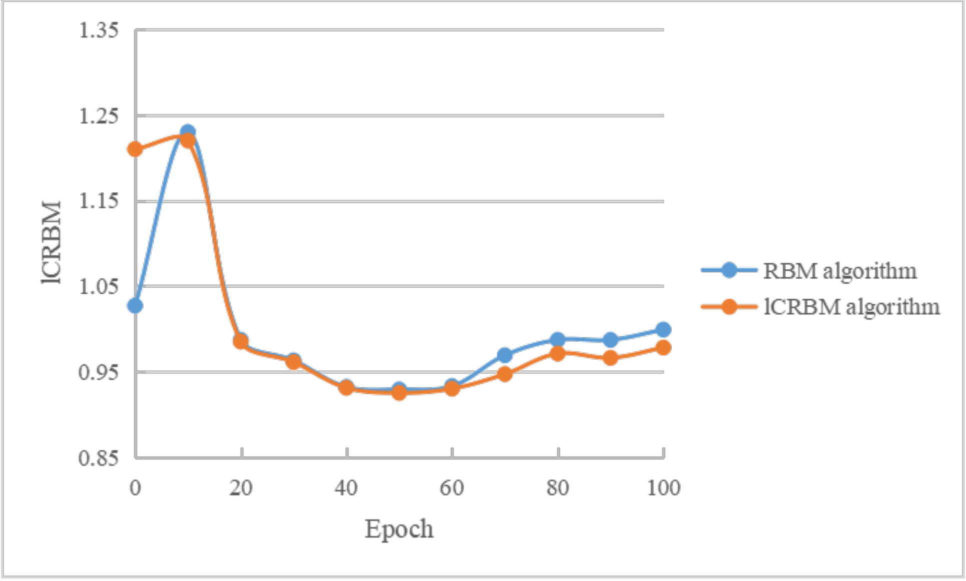

Fig.6 Comparative Experiment between This Algorithm and Rbm Algorithm

It can be seen from fig. 6 that the lCRBM value of this algorithm is always smaller than that of RBM algorithm, indicating that the accuracy of this algorithm is higher than that of RBM algorithm; It can be seen from Figure 3.3 that the improvement effect is obvious when the number of iterations is small and the number of iterations to achieve the optimal effect is small. When the number of iterations (epoch) reaches 50, subsequent iterations will cause over-fitting problems, resulting in the increase of lCRBM, and the recommendation accuracy of RBM algorithm will decrease rapidly. However, the recommendation accuracy of this algorithm is obviously lower than that of RBM algorithm, which shows that the anti-over-fitting ability of this algorithm is better than that of RBM algorithm.

5. Conclusion

Traditional search services can no longer meet the needs of Internet users. The rapid development of the Internet has brought a lot of convenience as well as some inconveniences. In this paper, the nearest neighbor is integrated when the RBM model is applied to collaborative filtering, which improves the discrimination ability of the RBM model, and further improves the accuracy of the prediction results of the model. The user's preference information is extracted by transforming the user's behavior into the user's score, which is used as the data input of the recommendation algorithm. At the same time, the clustering algorithm based on commodity label and user interest model is used to cluster users according to their interests, which is combined with the collaborative filtering recommendation algorithm based on users. Experimental results show that the accuracy of RBM model with nearest neighbor is higher than that of the original model, and the anti-over-fitting ability of the model is also improved.

References

[1] Shuang F, Chen C L P. A Fuzzy Restricted Boltzmann Machine: Novel Learning Algorithms Based on the Crisp Possibilistic Mean Value of Fuzzy Numbers[J]. IEEE Transactions on Fuzzy Systems, 2018, 26no. 1, pp. 117-130.

[2] Chang Hao, Yang Shengquan. (2020). Research on commodity recommendation algorithm based on collaborative filtering decision tree. Value Engineering, vol. 039, no. 009, pp. 127-129.

[3] Shen Weijie, Bian Longjiang, Zhang Xingjian, et al. (2019). Analysis and evaluation of quality information based on big data technology and application research of e-commerce procurement quality control strategy. Modern Management, vol. 9, no. 5, pp. 6.

[4] Wang Ye, Guo Lingli, Song Wenchao, et al. (2018). Research on equipment portrait recommendation algorithm in expert knowledge base based on big data technology. Computer Measurement and Control, vol. 26, no. 12, pp. 225-229.

[5] Li Xiaoying, Zhao Anna, Zhou Xiaojing, et al. (2019). Research and application of personalized product recommendation based on big data analysis and mining platform. Electronic Testing, no. 12, pp. 65-66.

[6] Bian Yuning, Li Yeli, Zeng Qingtao, Sun Yanxiong. (2020). Research on the Application of Improved Collaborative Filtering Recommendation Algorithm in Precision Marketing. Journal of Beijing Institute of Printing, vol. 28, no. 10, pp. 140-145.

[7] A Z J, B N C A, B D B P A, et al. A highly parameterizable framework for Conditional Restricted Boltzmann Machine based workloads accelerated with FPGAs and OpenCL[J]. Future Generation Computer Systems, 2020, 104:201-211.

[8] Liu Yan, Lei Jue. (2020). Research on ship information recommendation model based on collaborative filtering algorithm in big data environment. Ship Science and Technology, vol. 42, no. 04, pp. 155-157.

[9] Yu Tao. (2018). Research on Key Technologies of E-commerce Based on Big Data. Science and Technology Information, vol. 016, no. 036, pp. 35-36.

[10] Chopra P, Yadav S K. (2018). Restricted Boltzmann machine and softmax regression for fault detection and classification[J]. Complex & Intelligent Systems, vol. 4, no. 1, pp. 67-77.