Recommendation System of E-commerce Based on Improved Collaborative Filtering Algorithm

Xiaoying Wang
School of Software Engineering
Chongqing University
Chongqing,China
wxyRoy@126.com

Chengliang Wang
School of Software Engineering
Chongqing University
Chongqing,China
wcl@cqu.edu.cn

Abstract—**With the rapid development of information technology, the information overload problem in e-commerce site is becoming increasingly serious. It is difficult for people to obtain their own needs from the massive items information quickly. Recommendation systems contribute to alleviating the problem of information overload that exists on the e-commerce site. Collaborative filtering algorithm is most widely used in the recommendation algorithm, but there are still sparse data problems in collaborative filtering algorithm.In this paper, an e - commerce recommendation system based on improved user-based cooperative filtering algorithm is presented, which attempt to bridge the sparsity problem by combining the characteristics of user ratings with user reviews, and using the theme LDA model based on Spark framework to extract user preference.**

Keywords-e-commerce; sparsity problem; collaborative filtering; LDA; user preference

I. Introduction

A. Research Background

With the rapid development of information technology, e-commerce has become a part of people's daily life. However, the increasingly prominent problem of information overload in e-commerce is coming with information's development. E-commerce sites provide users with a wide range of products, so that it is difficult for users to get the information from the mass of commodity information they are really interested in quickly and accurately. So, the recommendation system, as a powerful tool to solve the problem of information overload, has a wide range of applications in e-commerce.

Recommendation systems are mainly divided into 3 categories: content based recommendation, collaborative filtering based recommendation and hybrid recommendation. The most widely-used recommendation among them is collaborative filtering recommendation. However, there are still some problems in practical applications such as low accuracy, cold start and sparse data. Because the traditional collaborative filtering algorithm relies too much on the rating data, once the rating matrix is too sparse, the available rating information will be too small, and the similarity of the users or items is difficult to guarantee the accuracy.

B. The Current Research Situation

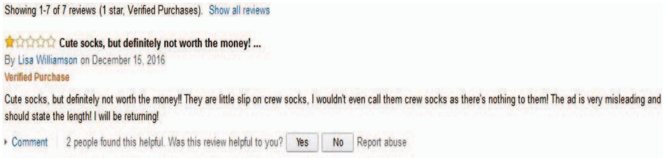
In order to produce accurate recommendations, researchers have made many improvements to the traditional recommendation algorithms. Breese and others have analyzed the various collaborative filtering algorithms and its improvements [1]. Goldberg proposed a new collaborative filtering algorithm Eigentaste, they applied PCA (principal component analysis) to a dense subset of rating matrix to reduce the dimensionality of the matrix, so as to ensure the accuracy of the algorithm at the same time and reduce the complexity of the algorithm [2].

C. Improved Thought of Recommender System in E-commerce

Most of the current e-commerce system support users to rate and review with brief text on what they buy, such as Dangdang and Taobao. Review texts often contain rich and valuable information resources, and they are also an effective way for businesses to obtain feedback from users [3]. Usually, the reviews given by users are mostly out of their own perception of items and their main impression. Most users will describe a few aspects which most impress them in the reviews, and that a few impressions, largely determines the user rating of items. Fig. 1 shows a product rating and review from the online store Amazon.

Traditional rating forecasting methods typically only consider user ratings, and do not deal with user's textual review data. In fact, user's review texts contain a lot of valuable information, such as the topic of the review, the level of attention, and the good or bad situation of these levels. Lu and others proposed a method for processing these comments to dig out the various levels and infer the scores at all levels [4]. This paper is inspired by its method. In the big data environment, this paper designed a e-commerce recommendation system which is based on improved user-based cooperative filtering algorithm in which combining the characteristics of user ratings with user reviews, and using the theme LDA model based on Spark framework to extract the theme.

978-1-5386-0497-7/17/$31.00 ©2017 IEEE



Showing 1-7 of 7 reviews (1 star, Verified Purchases). Show all reviews

★☆☆☆☆ **Cute socks, but definitely not worth the money! ...**

By Lisa Williamson on December 15, 2016

Verified Purchase

Cute socks, but definitely not worth the money!! They are little slip on crew socks, I wouldn't even call them crew socks as there's nothing to them! The ad is very misleading and should state the length! I will be returning!

2 people found this helpful. Was this review helpful to you?

Figure 1. Example of a Amazon user comments

II. RECOMMENDATION SYSTEMS AND RELATED KNOWLEDGE

A. *Recommendation System Survey*

Recommendation system is a kind of information service technology, which is mainly to solve the problem of difficult selection caused by massive information. It can help users find the information they want quickly. Table 1 lists some of the websites that use the recommendation technology at home and abroad.

TABLE I. EXAMPLES OF WEBSITES USING THE RECOMMENDATION TECHNOLOGY

Filed	Website
E-commerce	Amazon, eBay, Taobao
Music	Yaboo.com, Douban
Film	Netflix, MovieLens, Youku
News	GroupLens, Google News, PHOAKS

The working principle of the recommendation system is shown in Fig. 2 .

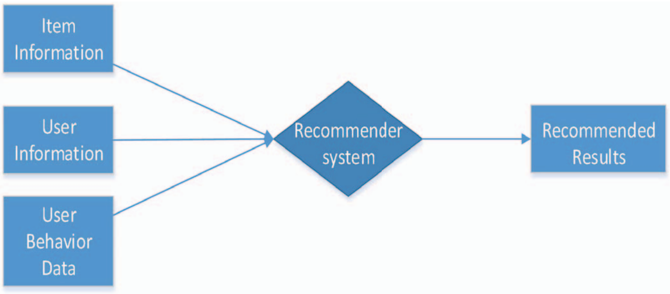


Figure 2. The working principle of the recommendation system

B. *Collaborative Filtering Algorithms Survey*

Traditional collaborative filtering algorithms predict user preferences and recommend items by taking advantage of the user's historical rating data [5,6]. In the user-based collaborative filtering algorithm, the predict rating that user u is possible to give to item i can be calculated by (1):

$$\text{pred}(u,\ i) = \bar{r}_u + \frac{\sum_{v\in N} \text{sim}(u,\ v)(r_{v.i} - \bar{r}_v)}{\sum_{v\in N} \text{sim}(u,\ v)} \qquad (1)$$

where N denotes the nearest neighbor set of user u, $r_{v,i}$ denotes the rating of item i given by user v, $\bar{r}$u rand $\bar{r}$v denote the average rating given by user u and user v respectively, sim(u,v) is the Person coefficient between user u and user v.

C. *Traditional LDA Model*

Latent Dirichlet Allocation (LDA) [7] is a generative probabilistic model, which aims to cluster words that co-occur in documents to form topics [8]. Each document d can be represented as a K-dimensional topic distribution θ_d, and each topic k is assigned a word distribution ϕ_k, which means the probability that a particular word is used for topic k. As shown in Fig. 3, a document can be considered as an ordered sequence of N words, a set of documents containing M document. α is a hyperparameter of θ and β is a hyperparameter of ϕ. We can use LDA to uncover hidden topics in reviews.

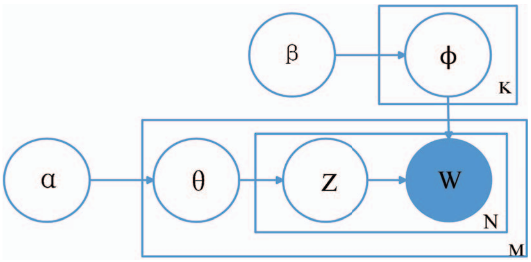


Figure 3. The LDA model diagram

D. *LDA Based on Spark*

For the purposes of document classification, the process of LDA is divided into two steps: the process of training and the process of reasoning . In big data environment, the amount of data is too large to have very good results after the training of traditional LDA model. Spark is a general parallel computing framework for the open class Hadoop MapReduce in Berkeley AMP labs. The distributed computing framework is implemented based on the MapReduce algorithm, and the job intermediate output and the final result can be stored in memory, thus reducing the I/O consumption of read-write HDFS to a certain extent. The greater the amount of data required to be read during processing, the greater the advantage of using Spark. Therefore, the system chooses to implement the extract of document topic features based on LDA model under the Spark framework.

III. CONSTRUCTION OF E-COMMERCE RECOMMENDATION SYSTEM BASED ON IMPROVED COLLABORATIVE FILTERING ALGORITHM

A. *System Function Module*

The system is divided into four functional modules, including user information management module, user purchase and review records module, commodity information management module, and recommendation module. Among them, the recommendation module is the core of the system which is divided into the following two parts:

- Hot items recommendation module: the system selects the top 10 items to show to the user according to the trade information. When the user enters the system, they can see the display of the hot items.
- Personalized recommendation module: According to the information of user's reviews and items, the system

infer what the items that the user may be interested in, then select Top-N items to recommend to the user.

B. *System Framework*

The overall framework of the system is shown in Fig. 4.

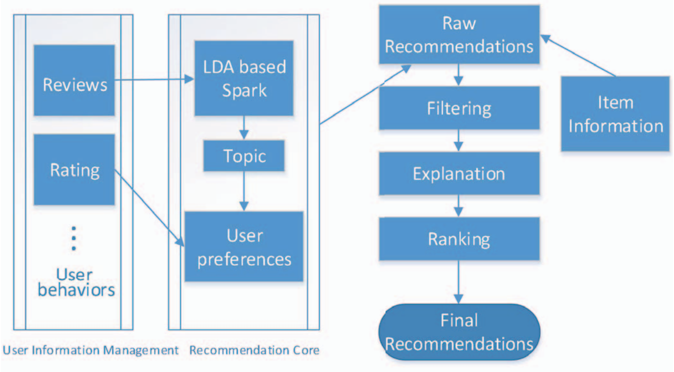


Figure 4. The system framework diagram

IV. Implementation of System Algorithm

The algorithm of this recommendation system using LDA topic model based on Spark to analyze user reviews, and generate the user preferences. Next, use the generated user preferences to calculate user preferences similarity, and combine user preferences similarity with the traditional user rating similarity to produce the final user similarity. Finally, use the user-based collaborative filtering thought to predict user ratings and generate the items recommended. The algorithm flowchart of this recommendation system is as follows:

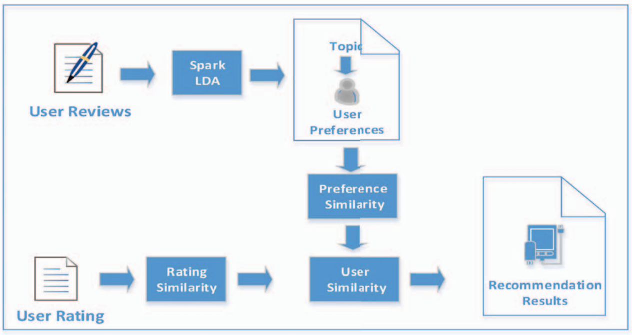


Figure 5. The algorithm framework diagram

A. *The Generation of User's Preference*

1) Pretreatment of User's Reviews: The original comment data tends to carry a lot of user's subjective ideas, many of which are irrelevant to the item, therefore they need to be pre-processed for comments.

2) Extraction of Item's Features: Item's features include intuitive features and hidden features. For example, "The size of the phone is right, but the standby time is a little short", in which the "size" is an intuitive feature, while the "power capacity" is a hidden feature. The technology of extracting hidden features is not mature at present, so the hidden features have not been considered in this system, which will be taken as the next research direction.

After the preprocessing of the comments, we extract nouns, verbs and adjectives as feature words from the generated data sets, and expressed in vector form. These feature words indicate the user's preference characteristics.

3) Pruning the Item's Feature: Some of the features that were extracted from items appear frequently in the reviews, but have nothing to do with the features of the items. Therefore, it is necessary to prune the extracted features.

The pruning process is mainly uses the HowNet to query the semantics of the feature words to find the semantic similarity of the feature words, and then realize the de-duplication and merge of the text, providing the basis for the topic discovery. The concrete process is:

Using HowNet to calculate the semantic similarity between each feature word. If the similarity is 1, then deleting the repeated feature words according to the probability of the occurrence of feature term in this semantic, and only the words with higher semantic probability are retained. Delete the feature that is not included in HowNet and merge the semantic similarity terms. When the similarity is greater than the threshold value, the semantic similarity terms are merged according to the probability of the feature appearing in the semantic, and the words with higher probability are retained.

4) Assessing Emotional Tendencies: User's reviews contain many words that represent their emotions, mostly adjectives and adverbs. We extract the emotional words associated with the item's features in the reviews, and use the HowNet to establish the emotional dictionary which is used to calculate the emotional tendency of emotional words.

5) Discover the Review's Topic: We utilize Spark-based LDA model to generate its K-dimensional topic distribution $\theta_{u,i}$. The $\theta_{u,i}$ is a probability distribution, which indicates the topics allocations of that review expressed by user u. These topics represent the user's personal preferences.

6) Evaluate the Emotional Intensity of Preference Topics: Each review of the user often corresponds to a specific rating $r_{u,i}$. We consider the high rating of reviews can provide more valuable topic distribution, and take the advantage of review attitude that extracted from rating information. The rating represents the attitude of the user, and the higher the score, it represents the characteristics of the product in line with the user's personal preferences, and the corresponding distribution of the theme can reflect the characteristics of user preferences [9]. We define the review attitude $t_{u,i}$ as follows:

$$t_{u,\,i} = \frac{2}{1+e^{-(r_{u,i}-\bar{r}_u)}} - 1 \qquad (2)$$

where $r_{u,i}$ is the rating of item i given by user u, $\bar{r}_u$ denotes the average rating of user u, the $t_{u,i}$ is in the range (0,1). In this way, $t_{u,i}$ indicates the attitude of review by utilizing the user rating information. When $t_{u,i} > 0$, it indicates that the user u has preferences for the items i, and conversely, the user u have a tendency to hate on item i [9]. The user preference with review attitude is defined as follows:

$$p_u = \frac{\sum_{i\in I_u} \theta_{u,i} t_{u,i}}{|D_u|} \quad (3)$$

where D_u is the set of all reviews given by user u, I_u represent the set of all items reviewed by user u, and $|D_u| = |I_u|$.

B. Calculation of User Preference Similarity

There are a lot of similarity calculation formula, such as the Pearson correlation coefficient formula, generalized Dice coefficient method and so on. We use the cosine similarity formula in algorithm, and the similarity degree of the user preference is calculated according to the attitude of the different users.

The user preference similarity's calculation is shown as follows:

$$\text{Psim}(u, \quad v) = \frac{\sum_{j=1}^{k} p_{uj} p_{vj}}{\sqrt{\sum_{j=1}^{k} p_{uj}^2} \sqrt{\sum_{j=1}^{k} p_{vj}^2}} \quad (4)$$

In which, p_{uj} and p_{vj} represent the preferences of users and users on the topic of j respectively.

C. Calculation of User Rating Similarity

Firstly, we utilize the improved cosine similarity method calculate the rating similarity of users. As follows:

$$\text{Usim}(u, \quad v) = \frac{\sum_{i\in I} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i\in I} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i\in I} (r_{v,i} - \bar{r}_v)^2}} \quad (5)$$

where I is the set of items, $\bar{r}_u$ and $\bar{r}_v$ denotes the average rating of user u and v, $r_{u,i}$ is the rating of item i given by user u, and $r_{v,i}$ is the rating of item i given by user v.

Finally, using parameters α to balance the importance of user preference similarity and rating similarity, and calculate the user's final similarity sim(u,v). For any two users u and v, user similarity sim(u,v) is calculated as follows:

$$\text{sim}(u, \ v) = \alpha\text{Psim}(u, \ v) + (1-\alpha)\text{Usim}(u, \ v) \quad (6)$$

where α values range from 0 to 1.

D. Scoring Forecasts and Produce TOP-N Recommendations

After calculation of user similarity, we put sim(u,v) into the (1) to calculate the prediction rating, and then select Top-N items which have higher rating from the large predicted items to recommend to the target user.

V. Conclusion

This paper constructs an e-commerce recommendation system which is based on improved collaborative filtering algorithm. In the system algorithm, we incorporate the review topics into user-based collaborative filtering algorithm, utilize LDA model based on Spark to generate review topics distribution, and then establish user preferences similarity according to the review topics. In the end, we establish final user similarity by combing user preferences similarity and user rating similarity calculated with rating information. Practical application shows that our system can improve the quality of recommendation, and effectively alleviate the sparsity problem of collaborative filtering.The future research will focus on establishing the user preferences and item features more accurately, and constructs a recommendation system with better recommended results.

[1] Breese J S, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering[C]// Fourteenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc. 2013:43-52.

[2] Goldberg K, Roeder T, Gupta D, et al. Eigentaste: A Constant Time Collaborative Filtering Algorithm[J]. Information Retrieval Journal, 2001, 4(2):133-151.

[3] Resnick P, Iacovou N, Suchak M, et al. GroupLens: an open architecture for collaborative filtering of netnews//Proceedings of the 1994 ACM conference on Computer supported cooperative work. ACM, 1994: 175-186.

[4] Breunig M M, Kriegel H P, Ng R T, et al. LOF: identifying density-based local outliers[C]// ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, Usa. CiteSeer, 2000:93-104.

[5] Resnick P, Iacovou N, Suchak M, et al. GroupLens: an open architecture for collaborative filtering of netnews[C]//Proceedings of the 1994 ACM conference on Computer supported cooperative work. ACM, 1994: 175-186.

[6] Linden G, Smith B, York J. Amazon.com recommendations: item-to-item collaborative filtering[J]. IEEE Internet Computing, 2003, 7(1):76-80.

[7] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3:993-1022.

[8] Chen L, Chen G, Wang F. Recommender systems based on user reviews: the state of the art[J]. User Modeling and User-Adapted Interaction, 2015, 25(2):99-154.

[9] LI Wei-lin, WANG Cheng-liang, WEN Jun-hao.Collaborative Filtering Recommendation Algorithm Based on Reviews and Ratings[J]. Journal of Computer Applications, 2017(2):361-364.