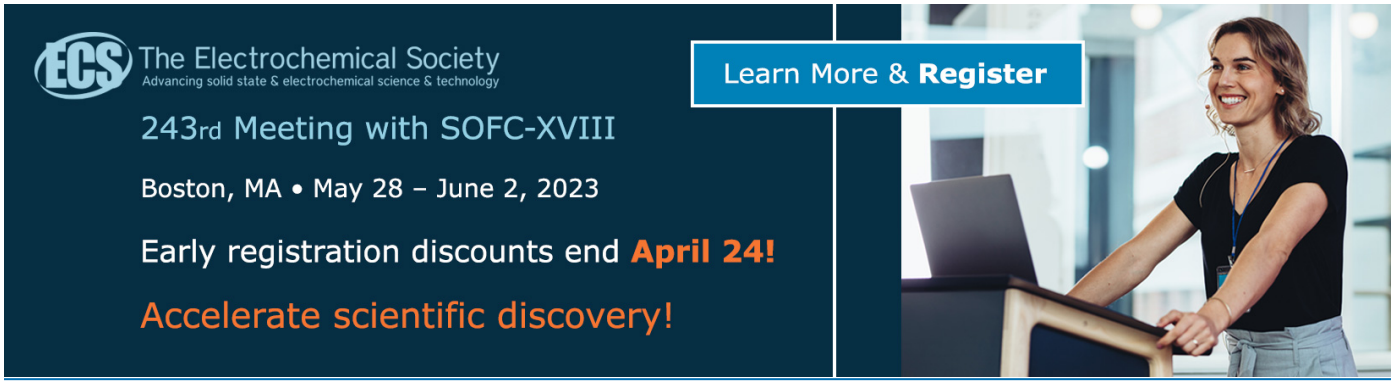
Research on Collaborative Filtering Recommendation Algorithm Based on Mahout and User Model



Research on Collaborative Filtering Recommendation Algorithm Based on Mahout and User Model

Bo Song [1], Yue Gao [2] and Xiao-Mei Li [3]

[1]The Software College, Shenyang Normal University, Shenyang, Liaoning, 110034, China

[2]The Software College, Shenyang Normal University, Shenyang, Liaoning, 110034, China

[3]The Liaoning Basic Education Research and Training Center, Shenyang, Liaoning, 110034, China

[*]Corresponding author's e-mail: songbo63@aliyun.com

Abstract. Based on the research status of recommendation system, the traditional collaborative filtering algorithm is explored. By establishing the concept of training sample and Mahout framework, a collaborative filtering algorithm based on Mahout and extracted user model can be proposed. Using this algorithm in the system of recommendation aims to solve the problem of transforming the traditional recommendation problem into the classification problem in machine learning. In this way, it can provide personalized recommendations for users. The analysis of the experimental results of the collected some data sets suggests that the algorithm based on user model and machine learning is effective and optimal, at the same time, the efficiency of the algorithm under Mahout framework is also verified.

1. Introduction

Collaborative filtering algorithm [1] is the most widely used personalized recommendation technology in business recommendation system, but when the amount of data is large, it will consume a lot of calculation time [2]. Machine-learning is the best way to solve the problem. Using this method, system administrators only need to implement the machine learning algorithm [3] based on the user model according to the rules of the training sample instead of doing a lot of repeated calculations. If you want to implement this algorithm, a software environment [4] that can run a large amount of data is also indispensable. In order to optimize the system and implement a collaborative filtering algorithm on the cloud computing platform, we chose Hadoop platform as the basis for implementation. Hadoop platform is an open source cloud computing platform, which uses the distributed-file-systems, that greatly improve the performance of parallel computing [5]. This paper mainly uses the collaborative filtering algorithm in the Hadoop-Mahout Machine Learning Framework to establish the user model and obtain the corresponding recommendation results. The collaborative filtering algorithm is a separate module in Mahout. The algorithm of this module can also be called the recommendation algorithm module [6]. This algorithm can greatly increase the transaction volume, so it has been widely used in e-commerce websites [7]. Similarly, the recommendation system developed based on the user model and the Mahout framework can make reasonable recommendations for online users [8], which greatly reduces the user's information query burden.

2. Research status of recommendation system

The recommendation system is an interdisciplinary discipline [9], which is developing for 20 years, that integrates information retrieval, prediction theory, function approximation theory and other fields. Nowadays, both the commercial field and the academic field are very concerned about the personalization of the recommendation system [10]. In the e-commerce field, Amazon, a famous e-commerce website, uses the recommendation system to its various applications. The website establishes a user preference model by collecting user information such as the behavior data, the browsed traces and the page stayed time [11]. It can find buyers interested in more products, which will bring more profits [12]. The world's some famous e-commerce websites are: GroupLens' News Recommendation System in 1994, Amazon in 1998(object-based collaborative filtering), Netflix Movie Recommendation System in 2006, and YouTube in 2016(applying deep neural networks to recommendation systems).

The recommended technology is not only popular in foreign countries, but it has also achieved great results in the researching of personalized recommendation in China. Some personalized websites, after the user logs in to the website, it may be directly recommending the products that the customer may purchase, or recommending some products that are of interest to the customer [13]. The recommended technology will be a wide range of application market in the virtual learning community, question and answer system and other fields.

3. Background and application of Mahout

The main purpose of Mahout is to implement adjustable machine learning algorithms [14]. Machine learning is a branch of artificial intelligence. "Machine learning is a science of artificial intelligence. The main research object in this field is artificial intelligence, especially how to improve the performance of specific algorithms in empirical learning." The Mahout algorithm runs on the Hadoop platform [15]. The Hadoop Cloud Platform is an open source framework for distributed applications that handle big data. Hadoop has the following advantages.

- Extremely scalable and reliable.
- Allow parallel work of big data.
- With software and hardware fault tolerance
- HDFS has high data throughput
- Provides both distributed storage and computing power.

4. Design Idea of Collaborative Filtering Algorithm

Personalized recommendations need to establish a user model [16], as shown in Table 1, where m represents the number of users and n represents the number of items, then the system's user model is an scoring matrix of $m \times n$. Use $U = \{U_1, U_2, U_3, \ldots, U_m\}$ to represent the user set. $I = \{I_1, I_2, I_3, \ldots, I_n\}$ to represent the Item set. And R represents the scoring matrix of $m \times n$. $r(i,j)$ is an element in R, and $i \in \{1, 2, 3, \ldots, m\}$, $j \in \{1, 2, 3, \ldots, m\}$. Definition $r(i, j)$ is a numerical type from 1 to 5. If the user scores 1 for the item, it means that the user does not like the item very much. If the user scores 5 points for the item, the user like the item very much.

Table 1. User_item score sheet

Customer	I_1	I_2	I_3	I_4	I_5	.....	I_{n-1}	I_n
U_1	4	2	?	3	4		$r_{(1,n-1)}$	$r_{(1,n)}$
U_2		4	5		3		$r_{(2,n-1)}$	$r_{(2,n)}$
U_3	5	4	3	4	4		$r_{(3,n-1)}$	$r_{(3,n)}$
.....								
U_{m-1}	$r_{(m-1,1)}$	$r_{(m-1,2)}$	$r_{(m-1,3)}$	$r_{(m-1,4)}$	$r_{(m-1,5)}$		$r_{(m-1,n-1)}$	$r_{(m-1,n)}$
U_m	$r_{(m,1)}$	$r_{(m,2)}$	$r_{(m,3)}$	$r_{(m,4)}$	$r_{(m,5)}$		$r_{(m,n-1)}$	$r_{(m,n)}$

5. Formatting the text

For recommended tasks, the scoring data only knows a part and needs to predict the unknown score [2]. According to the idea of User-based collaborative filtering algorithm, it is necessary to calculate the similarity between users. The following is an analysis of the specific steps. First, find a collection of users with similar interests to the target user. Secondly, find the items that the user in the collection likes and the target users have not heard of, and recommend to the target user. Finally, focus on the user and observe similar interests to the user. A group of users who recommend other items of interest to a group of similar interests to the user. Refer to Eq. (1), $w_{u,v}$ is to calculate the correlation between two users u and v, and $r_{u,i}$ indicates the rating of user u for item i, and $r_{v,i}$ indicates the rating of user v for item i, and $\bar{r}$ is the average score of the user rating. Refer to Eq. (2), $P_{a,i}$ is to calculate the final score of a user a for item i.

$$w_{u,v} = \frac{\sum_{i \in I}(r_{u,i} - \overline{r_u})(r_{v,i} - \overline{r_v})}{\sqrt{\sum_{i \in I}(r_{u,i} - \overline{r_u})^2}\sqrt{\sum_{i \in I}(r_{v,i} - \overline{r_v})^2}} \tag{1}$$

$$P_{a,i} = \overline{r_a} + \frac{\sum_{u \in U}(r_{u,i} - \overline{r_u}) \cdot w_{a,u}}{\sum_{u \in U}\left|w_{a,u}\right|} \tag{2}$$

6. Distributed Item-based collaborative filtering

User-based collaborative filtering requires frequent updating of the user similarity matrix, while the recommended number of items remains relatively stable and the commodity similarity matrix update frequency is low. Therefore, an Item-based collaborative filtering recommendation strategy is proposed [6]. The following is an analysis of the specific steps. First, calculate the similarity between items. Second, a recommendation list is generated for the user based on the similarity of the item and the historical behaviour of the user. Finally, with the item as the center, similar items are calculated by observing the user's preference behaviour for the item. These similar items can be considered to belong to a particular set of categories. Then calculate the category to which it belongs based on a user's historical interests. See if the category belongs to one of these group categories, and finally recommend the items belonging to the group category to the user. Refer to Eq. (3), $w_{i,j}$ is to calculate the correlation between two item i and j, and $r_{u,i}$ indicates the rating of user u for item i, and $r_{u,j}$ indicates the rating of user u for item j, $\bar{r}$ is the average score of the Item. Refer to Eq. (4), $P_{a,i}$ is to calculate the final score of a user a for item i, and guessing whether user a is interested in item i.

$$w_{i,j} = \frac{\sum_{u \in U}(r_{u,i} - \overline{r_i})(r_{u,j} - \overline{r_j})}{\sqrt{\sum_{u \in U}(r_{u,i} - \overline{r_i})^2}\sqrt{\sum_{u \in U}(r_{u,j} - \overline{r_j})^2}} \tag{3}$$

$$P_{a,i} = \frac{\sum_{j \in I} w_{i,j} \cdot r_{a,j}}{\sum_{j \in I}\left|w_{i,j}\right|} \tag{4}$$

7. Distributed Model-based collaborative filtering

7.1. Bayesian algorithm

Naive Bayes [17] classification is a classification algorithm. The idea is to solve the probability of occurrence of each category under the conditions of this occurrence for the given item to be classified. Which is the largest, it is considered which category the item to be classified belongs to.

The Bayesian algorithm flow is divided into three phases, as shown in Figure 1. Each data sample is represented by a feature vector A (a_1, a_2, a_3, ..., a_l). Each data sample A has a category B_i to which the sample belongs, $P(B_i)$ represents the probability of each category, and $P(a_i|B_i)$ represents each category of each attribute after the occurrence of B_i. Probability, $P(B_i|(a_1,a_2,a_3,...,a_i))$ denotes the probability that the new feature vector A is divided into B_i, and $max(P(B_i|(a_1,a_2,a_3,...,a_i)))$ denotes the

class *A* where the new feature vector *i* is most likely to be divided. The naive Bayesian algorithm is the most critical throughout the recommendation process.

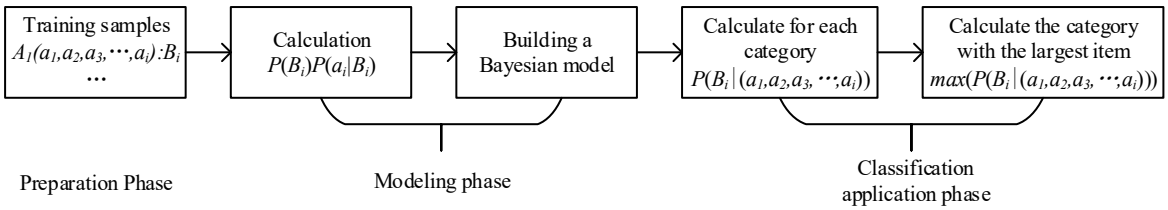


Figure 1. Bayesian algorithm flow chart

7.2. Model-based design

The core idea of collaborative filtering recommendation algorithm based on user model is data mining and machine learning. The user model is constructed using user historical data, and the predicted products are scored accordingly. First, the algorithm of Naive Bayes collaborative filtering is used for analysis. Then, based on the model-based collaborative filtering recommendation, this is a machine learning method, which is recommended by offline calculation. Usually, it first divides the data set into a training set and a test set based on historical data, and uses the training set to train. Generate a recommendation model, then apply the recommendation model to the test set to evaluate the pros and cons of the model. If the model reaches the actual required accuracy, the recommended model can be used to predict the recommendation. Since the score ranges from 1 to 5, it is divided into 5 categories. class indicates the category of the label. Eq. (5) is a variant of the Bayesian formula, and classj indicates the label j, and classSet = {1, 2, 3, 4, 5} indicates the label data set. It is noteworthy that the article uses five types of labels. Xo is user o. xo is the rating of user. P is the probability of calculating the condition. In this way, the more the number of labels, the more accurate the recommended results.

$$class = \arg\max_{class_j \in classSet} P(class_j) \prod_o P(X_o = x_o | class_j) \tag{5}$$

$$P(A, B|C) = P(A|C)P(B|C) \tag{6}$$

According to the relationship matrix of Table 1, to predict how many points U_1 evaluates to I_3. Refer to (4), using the model-based collaborative filtering algorithm and the Naive Bayes classification algorithm, a machine learning algorithm that can obtain regular data from the automatic analysis of data is designed. The output value is obtained according to the training samples.

$$\begin{aligned} class &= \arg\max_{class_j \in \{1,2,3,4,5\}} P(class_j)P(U_2 = 5|class_j)P(U_3 = 3|class_j) \\ &= \arg\max_{class_j \in \{1,2,3,4,5\}} \{0, 0.0069, 0.0139, 0.0102, 0\} = 3 \end{aligned} \tag{7}$$

From the above calculation, it can be concluded that User 1 gives Item 3 a score of 3, indicating that the user is interested in Item 3 and can make recommendations.

8. Experimental results and analysis

The computer memory is 16G, 1T of the hard disk, 2.4GHz of the processor and the operating system Ubuntu 12.04 desktop 64bit. The experimental analysis comes from data provided by the MovieLens website [18]. The data set includes user information, movie related data, and user ratings. Based on these data, a correlation matrix is formed. For the user-based collaborative filtering recommendation strategy and the item-based collaborative filtering recommendation strategy, in principle, it is not established the user module, and directly the calculation is performed. The Naive Bayes collaborative filtering algorithm based on the user model uses the machine learning knowledge content when building the model. The principle is to convert the recommendation problem into a classification problem and derive the output value based on the training sample. As shown in Fig. 2, it can be seen from the analysis of the three algorithms that the collaborative filtering algorithm based on the user model has the smallest error with the average number regardless of the number of user ratings. In most

cases, its value is between the other two algorithms. It can be used to analysis the user's information needs. The corresponding resources are accurately and precisely pushed forward through the establishment of the user model.

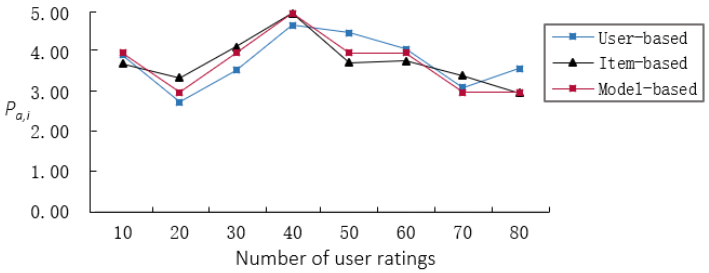


Figure 2. Test score chart for three algorithms

Figure 3 is a recommended process analysis of the Naive Bayes collaborative filtering algorithm based on the user model. During the registration process, the user will fill in some indicators and then analyze them. First, the user's basic information is taken as the user feature vector, and the training samples are extracted from them. According to the training samples, preliminary recommendations are obtained. Continuously filter the sample and get the final recommendation.

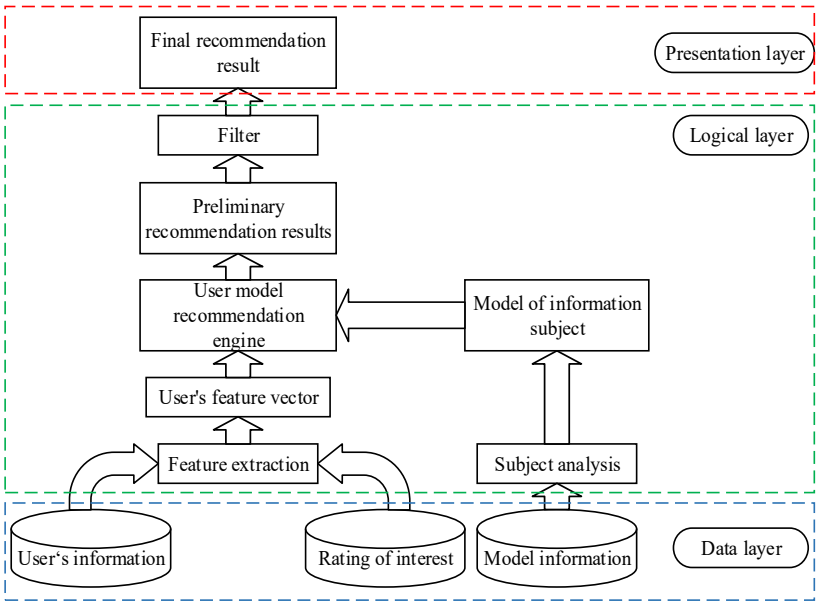


Figure 3. Recommended process analysis chart

9. Conclusion

The user-based collaborative filtering recommendation algorithm proposed in this paper is based on the Mahout framework, combined with the Nave Bayes classifier [19], to establish a user model and form a personalized recommendation. From the analysis of the three algorithms, it can be concluded that the collaborative filtering recommendation algorithm based on the user model is an algorithm of machine learning. In today's era of machine learning, we need to find regular models from the issue and machine the complex problems. The recommendation algorithm running on the Hadoop platform not only solves practical problems but also implements more an effective and more personalized recommendation for users. At the same time, in a network environment with large data, the running speed is improved and the workload of the server is reduced. In the construction of the online recommendation field, model-based algorithm has essential reference values to improve the quality and speed of user search information.

Acknowledgments
The authors acknowledge—(1) The Basic Research Foundation of Universities in Liaoning (Grant: 2017L317); (2) The 13th Five-Year Plan of Education Science in Liaoning Province. (Grant: JG18DB451).

References
[1] Jiao, D. J., Meng, X. w. Improvement of Apache Mahout collaborative filtering algorithm evaluation method[J]. Journal of Jinan University (Natural Science Edition),2016,30(01):47-50.
[2] Huang, X. Y., Long, Y. Y., Xie, J. Collaborative Filtering Recommendation Algorithm Based on User Interest Clustering [J/OL]. Computer Application Research, 2019(09): 1-7.
[3] Tu, E. M. Design of machine learning algorithm based on graph theory and its application in neural network [D]. Shanghai Jiaotong University, 2014.
[4] Xu, W. J., Liu, Q. K., Zheng, X. W., etal, Distributed Course Recommendation Algorithm Based on Hadoop-Mahout[J]. Computer Applications and Software, 2018, 35(03): 236-240.
[5] Fu, W., Du, L., Zhang, K. B., etal, Service recommendation algorithm based on trusted alliance for smart communities[J]. Computer Engineering, 2019, 45(02): 310-314.
[6] Sarwar, B., Karypis, G., Konstan, J., etal, Item-Based Collaborative Filtering Recommendation Algorithms[C]//Proceedings of the 10th international Conference on World Wide Web. Hong Kong: ACM Press, 2001: 285-295.
[7] Zhang, Y. Z. Research on recommendation algorithm for single-class collaborative filtering [D]. University of Science and Technology of China, 2018.
[8] Zhang, Y. H., Zhu, X. F., Xu, C. Y. A hybrid recommendation method based on deep emotionanalysis and multi-view collaborative fusion based on user comments [J/OL]. Chinese Journal of Computers, 2019:1-19.
[9] Shinde, R. U., Raut, S. D. Typicality-based collaborative filtering recommendation system. Int. J. Innov. Res. Com-put. Commun. Eng. (IJIRCCE) 4(6), 12498–12504 (2016).
[10] Wu, Y. W., Qi, W., Yang, R. A News Recommendation System Based on Improved Collaborative Filtering Algorithm[J]. Computer Engineering and Science, 2017, 39(06): 1179-1185.
[11] Liang, L. J. Li, Y.G. Zhang, N. N., etal, Collaborative filtering algorithm based on user feature optimization clustering[J/OL]. Journal of Intelligent Systems, 2019(03):1-7.
[12] Xiao, C. L., Wang, N. Y. Wang, G. Parallel collaborative filtering recommendation algorithm based on social network and key users [J/OL]. Computer Application Research, 2019(10): 1-8.
[13] Gao, X. W., Shi, Z. B. Design and Implementation of New User Recommendation Algorithm Based on Mahout[J]. Computer Engineering and Science, 2015, 37(08): 1444-1449.
[14] Fan, Z. Mahout algorithm analysis and case combat. Beijing: Mechanical Industry Press, 2014, pp. 84-94.
[15] Yin, X. Personalized job recommendation system based on Mahout framework [D]. Tianjin University of Technology, 2018.
[16] Song, B., Gao, J., Wu, P. Improvement and Application of Collaborative Filtering Recommendation Algorithm in e-Learning System[J]. Journal of Shanxi University (Natural Science),2015,38(04):573-580.
[17] Feng, X. K., Liu, Y., Jiang, X. F. Application of Naive Bayesian Classification Algorithm in Data Prediction[J]. Software Guide,2011,10(05):65-66.
[18] GroupLens Research. MovieLens [EB/OL]. http://files.grouplens.org/datasets/movielens/,2019-01-10.
[19] Di, P., Duan, L. G. A new type of Naive Bayesian text classification algorithm[J]. Journal of Data Acquisition & Processing,2014,29(01):71-75.