The Research for Recommendation System Based on Improved KNN Algorithm

Bin Li*
School of Continuing Education
Anhui Open University
Hefei, China
e-mail: szbinlee@126.com

Sailuo Wan
School of Continuing Education
Anhui Open University
Hefei, China
e-mail: wsl@ahou.edu.cn

Hua Xia
School of Continuing Education
Anhui Open University
Hefei, China
e-mail: xiahua@ahou.edu.cn

Fengshou Qian
School of Continuing Education
Anhui Open University
Hefei, China
e-mail: qfs@ahou.edu.cn

***Abstract*—In this paper, we have researched two basic tasks of recommendation system score prediction and Top-N recommendation. We have improved K nearest neighbor (IKNN) algorithm with compression and global effect. In experiments, the methods of Top-10 recommended mainly refer to the score on the basis of prediction. We recommended the items whose scores are the highest. The experimental results show that using IKNN algorithm recommendation system score predicted mean square difference (RMSE) has reduced significantly. Meanwhileit has a well recommendation precision improvement.**

Keywords—recommendation algorithm, recommendation system, KNN, RMSE

I. INTRODUCTION

At present, publics are in an era with information explosion. Persons have to takea lot of time and energy searching for information they need. However, although the needed information is found, it is mixed with many “noises”, which hinders people to find truly useful information from mass information and therefore the utilization efficiency for information reduces on the contrary. The above-mentioned phenomenon is called information overload [1]. And we will take a lot of time to search the important information which we just need. At the same time, the efficiency of the use of information has also been reduced, which we call information overload. Search engines solve the problem of information filtering more or less, but they can’t resolve the different needs of individual users. Is there a system which can searchimportant information from a large number of items automatically and then recommend them?

Therefore, recommendation algorithm is the core part of the whole recommender system and the performance of the recommender system depends to some extent on the recommendation algorithm. In this paper, we have researched the KNN recommender algorithm deeply, and then proposethe improved algorithm of KNN with compression and global effect.

II. RELATED WORK

According to difference of the data and recommendation algorithm used, the recommender systemsare mainly divided into three types: recommender system based on content, collaborative filtering recommendation system and recommender system based on social network.

The model extracts keyword information of items at first and then calculates their weights by using different methods (including term frequency (TF), document frequency (DF), information grain (IG), and term frequency-inverse document frequency (TF-IDF)). TF-IDF is the most commonly used [2]. CataldoMusto et al. presented a preliminary investigation towards the adoption of Word Embedding techniques in a content-based recommendation scenario [3]. Protasiewicz J et al. proposed the architecture of a content-based recommender system aimed at the selection of reviewers (experts) to evaluate research proposals or articles [4]. Jian Wet al. proposed two recommendation modelsbased on a frameworkof deep learning neural network to solve cold start for new items [5]. Based on a collaborative filtering strategy, an electricity plan recommender system (EPRS) is developed by Zhang Yuan et al [6]. In real life, users’ interests and choices of items tend to be influenced by their friends. With the rise of social networking sites such as Fackbook and Twitter, how to design recommendation systems by using social relationships among users has become a research hotspot of recommendation in recent years [7]. These algorithms are also known as social recommendation. Social network is a kind of relationship network with a specific structure formed due to interaction between people, between people and organizations, and between organizationsfor certain specific or same purposes. The relationship network consists of different nodes which can be various entities (including people and organizations) and lines connecting nodes. Moreover, the lines connecting nodes represent the relationship between the above-mentioned entities. Quijano-Sanchez et al. includes an analysis of group personality composition and trust between each group member to improve the accuracy of group recommenders [8].

978-1-7281-6521-9/20/$31.00 ©2020 IEEE

III. K-NEAREST NEIGHBOR ALGORITHM AND ITS IMPROVEMENT

The K-nearest neighbor (KNN) algorithm is widely used among classification algorithms [9], whose concept is derived from nearest neighbor classification principle: when judging the category of an object to be classified, the most similar object is searched from the existing training set to thus further judge the classification of the object to be classified according to the categorical attribute of the most similar object. Unlike the methodusing only one nearest neighbor object, KNN algorithm can refer to categorical attributes of more than one approximate object when judging the category of an object. In classification, KNN algorithm shows different results with varying K value.

A. KNN Algorithm

The common methods in the recommender system for calculating similarity mainly involve the cosine similarity, adjusted cosine similarity and Pearson correlation similarity methods. We can use the similarity calculation formula to calculate the similarity of users or items.

1) Similarity calculation

Assuming that vectors i and j separately represent the scores of items i and j in m-dimensional space, the formula for calculating the similarity sim(i, j) between the two items is shown in Formula (1).

$$sim(i,j) = \cos(i,j) = \frac{i \cdot j}{\|i\| * \|j\|} \qquad (1)$$

Actually, users likely have different scoring scales for different items, while the scaling problem is not taken into account in Formula (1). In Formula (2), the adjusted cosine similarity method is used to solve the problem by subtracting the average score of the item from the score. When a user scoresthe item i and item j, and the Ui, jdenotes the set of the users. Moreover, Ui denotes the set of all users who score item i while Uj represents the set of all users who rate the item j. Therefore, the similarity between itemi and item j as Formula (2).

$$sim(i,j) = \frac{\sum_{c \in U_{i,j}} (R_{c,i} - \overline{R}_i)(R_{c,j} - \overline{R}_j)}{\sqrt{\sum_{c \in U_i} (R_{c,i} - \overline{R}_i)^2} \sqrt{\sum_{c \in U_j} (R_{c,j} - \overline{R}_j)^2}} \qquad (2)$$

Where, Rc, irefers to the score of the user c on the item i, $\overline{R}_i$ represent the average score of items i and $\overline{R}_j$ represent the average score of itemj.

2) Select K nearest neighbors and calculate predicted values

While the calculation of the similarity of all the items evaluated by the user u, K items with the highest similarity with the item j are got, and the set of the K items is represented by N(u, j).

Through the calculation of various similarity calculation methods, the similarity between the items can be obtained. By selecting the K neighbors, we obtain K neighbors of the item, and the corresponding recommended objects need to be generated in the next step. Before generating the recommended object, we need to calculate the user's score for the proposed item. The user u's score for item j is calculated as shown in Formula (3):

$$\hat{R}_{u,j} = \frac{\sum_{n \in N(u,j)} sim(n,j) R_{u,n}}{\sum_{n \in N(u,j)} sim(n,j)} \qquad (3)$$

Here, $sim(n, j)$ indicates the similarity between the item n and the item j, and $R_{u,n}$ indicates that the rating value of the item n is given by the user u.

And then the user can predict the score of the unrated items in the system, and then recommend to the user by selecting the N items with the highest predicted scores.

B. Improve the KNN Algorithm

1) Data sparsity improvement

As the number of users and items in the recommendation system increases, the amount of data in the entire system becomes very big, and the number of items used by two users becomes very small, which results in sparse data generation. We can see from the similarity formula that if the intersection of two items is much smaller than the intersection of other items, the reliability of the similarity of those two items will be low. If we want to raise the reliability of the prediction result, we need to compress the similarity in accordance with the size of the intersection as Formula (4):

$$\overline{sim(i,j)} = \frac{|U_{i,j}|}{|U_{i,j}| + \gamma} * sim(i,j) \qquad (4)$$

Here, $\overline{sim(i,j)}$ denotes the similarity between compressed itemi and item j, and γ meansthe compression factor which specified by the user.

2) Global effect improvement

Since the user or the item itself has many self-characteristics, there is a subtle tendency to score while the user scores the item. For instance, some users are strict raters and they want togive items a low score, while some other not-strict raters will give a high value on items. For the same reason, there is a similar situation for items [10]. The factors in recommendation system are called global effect (GE)which include average of all items, the time interval of scoring, and the scoring times of items and so on [11]. The KNN algorithm is improved by the factors of GE, and Formula as follow.

$$\hat{R}_{u,j} = om + \sum_{i=1}^{t} \theta_i x_{u,j} + \frac{\sum_{n \in N(u,j)} \overline{sim(j,n)} (R_{u,n} - om - \sum_{i=1}^{t} \theta_i x_{u,n})}{\sum_{n \in N(u,j)} \overline{sim(j,n)}} \qquad (5)$$

Where, omrepresents the average score of all of the items. And θ_i refers to the ith of GE,t denotesthe number of GEand $x_{u,j}$ refers tothe explanatory variables of the user u and the jth item. $N(u, j)$ denotesthe set of the most similar k items to item j for user u, and $\overline{sim(j,n)}$ refers to the compressed similarity of item j and item n.

IV. Experimental Data and Analysis of Experimental Results

A. Experimental Data

TheMovieLens datasethas beenobtained from a very famous movie scoring website www.movielens.org. Itisconsistedby three different sized datasets: the ML-100K, theML-1M and ML-10M datasets which are composed of 100,000, 1 million, and 10 million scoring records, respectively.In the study, 80% of the records from each dataset as the training set and the remainingdata as the test dataset.

B. The Result based on Improved Algorithm

In the prediction of the item, we used the Root Mean Squared Error (RMSE), and the recommended precisionwas used to evaluate the recommended effect. The Formula as follow.

$$RMSE=\sqrt{\frac{1}{n_u}\sum\left|R_{u,i}-R_{u,i}^{(real)}\right|^2} \quad (6)$$

Where, n is the number of items that the user u scores in the system, $R_{u,i}$ is the predicted score value of the user u for the unrated item i at the time of the score prediction test. And $R_{u,i}^{(real)}$ is the real prediction score of the user u for the item iusing test data set, nudenotes the number of user-item pairs in the recommendation system.

We used KNN algorithm and improved KNN (IKNN) algorithm to experiment with three different data sets of MovieLens. In the experiment, the algorithm would get the best result while the value of K=10.We use various algorithms to select the scores of the top 10 movies based on the scores of different users' prediction scores for movies. And the experimental results are shown in the table I.

Table I. Performances of KNN and IKNN Algorithms in Different Datasets

Dataset	Algorithm	RMSE	Precision
ML-100*K*	*KNN*	1.0758	85.16%
	IKNN	0.9462	88.57%
ML-1*M*	*KNN*	1.0155	87.33%
	IKNN	0.8827	**89.22%**
ML-10*M*	*KNN*	1.0169	86.74%
	IKNN	**0.8776**	88.30%

It can be seen from Table I that the experiments on three different datasets, the performance of the IKNN algorithm was much better than the KNN algorithm. When the size of the data set is 10M, the minimum RMSE value was 0.8776, and the RMSE value obtained by the IKNN algorithm was 0.13 lower than the KNN algorithm on average. The recommendation precision rate of both algorithms was above 85%, and the recommendation effect of IKNN algorithm was better than KNN algorithm, and the average accuracy rate was improved by nearly 2.3%. Among them, when the dataset size is 1M, the best recommendation precision rate is 89.22%.

V. Conclusion

In the study, we introduced the background and significance of the recommendation system, and summarized and reviewed the research and development of the recommendation system and the results obtained. And then different recommendation systems were described in detail. The KNN algorithm was improved by two aspects of data sparsity and global effect factors.

Although some progress had been made in this study, issues such as the diversity of recommended items and the sparseness of data need to be further addressed. At the same time, in this study, only the user's scoring information was considered, and other user's own information was not used, such as the user's own characteristics (e.g. gender, age, etc.) and the time of scoring the item. In future research, more factors could be considered to obtain better results.

Acknowledgement

This work was supported by the Key Program of the Foundation for Young Talents in the Colleges of Anhui Province under Grant 2013SQL097ZD; and the Program of Quality Engineering in the Colleges of Anhui Province under Grant 2017SJJD101, 2018XFSYXM010, 2019ZDJG08, 2019ZDJG09, 2019ZDJG10, 2019JXJJ88, the Key Program of the Natural Science Foundation of Anhui Education Committee under Grant KJ2019A0969.

References

[1] R.T. Arnold, “Introduction to Modern Information Retrieval,” Third Edition. Library Resources & Technical Services, vol. 44, pp. 239-240, 2011.

[2] G. Salton. “A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART).” Journal of the American Society for Information Science & Technology, vol. 23, pp. 75-84, 2010.

[3] C. Musto, G. Semeraro, M. D. Gemmis, et al. “Learning Word Embeddings from Wikipedia for Content-Based Recommender Systems”. European Conference on Information Retrieval. Springer, Cham, 2016.

[4] J. Protasiewicz, et al. “A recommender system of reviewers and experts in reviewing problems.” Knowledge-Based Systems, 2016:S0950705116301381.

[5] W. Jian, J. He, C. Kai, et al. “Collaborative filtering and deep learning based recommendation system for cold start items”. Expert Systems with Applications, vol. 69, pp. 29-39. 2017.

[6] Z. Yuan, M. Ke, K. Weicong, et al. “Collaborative Filtering-based Electricity Plan Recommender System.” IEEE Transactions on Industrial Informatics, vol. 1, 2018.

[7] B. Yang, Y. Lei, J. Liu, et al. “Social Collaborative Filtering by Trust.” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, pp. 1633-1647, 2017.

[8] L. Quijano-Sanchez, J. A. Recio-Garcia, Diaz-Agudo B , et al. “Social factors in group recommender systems.” ACM Transactions on Intelligent Systems and Technology, vol. 4, pp. 1-30. 2013.

[9] Bin L, Jun L, J.M. Y, et al. “Automated Essay Scoring Using the KNN Algorithm”, International Conference on Computer Science and Software Engineering, CSSE 2008, Volume 1: Artificial Intelligence, December 12-14, 2008, Wuhan, China. IEEE, 2008.

[10] R. M. Bell and Y. Koren. “Seventh IEEE International Conference on Data Mining (ICDM 2007) - Scalable Collaborative Filtering with Jointly Derived Neighborhood Interpolation Weights”. pp. 43-52. 2007.

[11] L. Bin, M. Ning, L. Ninghui and G. Yuliang. “A recommendation algorithm based on the hybrid model.” Scientific Bulletin Series: Electrical Engineering and Computer Science, vol. 81, pp. 189-204, 2019.

[12] V. R. Prakash, Saran (2019). An Enhanced Coding Algorithm for Efficient Video Coding. Journal of the Institute of Electronics and Computer, 1, 28-38.