Survey of Recommendation Based on Collaborative Filtering

Wang Juan1,*, Lan Yue-xin1, Wu Chun-ying1
1 China People’s Police University, Langfang 065000, China
66267476@qq.com

Abstract. This paper introduces the domestic and international research of collaborative filtering, and discusses the main problems of collaborative filtering algorithm, including data sparsity, cold start and accuracy of similarity measure. Then, future research and development trends of integrating deep learning to recommender systems are pointed out. In order to solve the data sparsity and cold start problems in the personalized recommendation system, a hybrid collaborative filtering recommendation algorithm is proposed, which combines the KNN model and XGBoost model. When deep learning is applied to recommendation system by integrating massive multi-sources heterogeneous data, it could improve the performance of the recommendation system.

1. Introduction
In the era of artificial intelligence, it is more and more convenient for users to obtain resources from platforms or systems, but it takes a long time to select the resources of interest from a large number of data resources. In order to solve the problem caused by “information overload”, Baidu, Google and other search engines search according to the user's keywords, and return the top ranked results to the user. But some users don't know the keywords they need to find. In order to make up for this shortcoming, a personalized recommendation system has emerged, which can help users discover resources that are not realized but are very interesting, thus saving time in searching.

Collaborative filtering is a widely used algorithm in recommendation systems. The statistical technique is used to find neighbor users who have the same or similar interests as the target users, and predict the items that the target users may like according to the preferences of neighbor users. The personalized recommendation systems use the scoring matrix composed of users and projects to predict the user's interest. The greater the user's interest in the project, the higher the score. Since the K-nearest neighbor algorithm is easy to use and fast, most recommendation systems use collaborative filtering model based on K-nearest neighbor algorithm.

1.1. Research Status of Collaborative Filtering Algorithm
Since the first recommendation system Tapestry[1] was introduced in 1990s. In the early stage, due to hardware limitations, it is difficult for the system to collect and store a large amount of data. People's research on recommendation system only stays in theoretical stage. However, with the rapid development of science and technology, more and more scholars have begun to conduct in-depth research on the recommendation system. Traditional recommendation methods mainly include collaborative filtering, content-based recommendation methods, and hybrid recommendation methods.
Collaborative filtering is currently the most widely used recommendation algorithm. Common collaborative filtering algorithm models are: neighborhood-based collaborative filtering method, hidden factor model-based collaborative filtering method, graph-based collaborative filtering method, labeling and time factor. Collaborative filtering methods, fusion models, etc.

Collaborative filtering algorithm is mainly divided into memory-based CF and model-based CF. The memory-based collaborative filtering algorithm is divided into a user-based collaborative filtering algorithm and an item-based collaborative filtering algorithm. The user-based collaborative filtering algorithm is the earliest recommendation algorithm. The algorithm first calculates the users similar to the target user's interest, and then recommends the user's favorite items in the similar user set. The item-based collaborative filtering algorithm recommends users to items similar to those they liked before. The model-based collaborative filtering algorithm mainly uses machine learning and data mining models to extract the implicit patterns of users and articles by classification, regression, matrix decomposition and other algorithms. The content-based recommendation algorithm extracts the user's interest preference for the item by analyzing the item content information. Both the content-based recommendation algorithm and the collaborative filtering recommendation algorithm have shortcomings. The hybrid recommendation algorithm solves the shortcomings of the single recommendation model by combining multiple recommendation techniques.

Abroad, Resnick et al. [2] proposed the first automated collaborative filtering system names Group Lens. Group Lens collects user preferences by scoring. Then, the system calculates the similarity between users, and selects a group of similar users to predict the current user's preference for new articles. In recent years, some scholars have studied the combination of deep learning and recommendation algorithms[3-6]. Wang H, Wang N, et al.[7] proposed a hybrid collaborative filtering algorithm. By combining self-encoder and Beiyesitu, the combination of recommendation algorithm and deep learning is realized, and good results have been achieved.

Domestic, many researchers proposed improved algorithms for the problems of cold start, data sparsity and accuracy of similarity measure for traditional collaborative filtering recommendation systems. Cao Zhanwei [8] et al. proposed a matrix decomposition recommendation algorithm based on the LDA theme model, which introduced KL divergence in the similarity calculation, improved the similarity measure accuracy, and reduced the recommendation error. Zhang Haixia [9] et al. proposed an improved meta path-based collaborative filtering algorithm for weighted heterogeneous information networks, which calculated the final predicted scores by combining various methods using supervised machine learning algorithms. Mi Cui[10] et al. proposed a trust recommender model for mobile SNS based on contextual interest in cloud environment, which integrated the contextual interest of users, and combined the mobile SNS context similarity matrix with the trust matrix to alleviate the problem of low recommendation accuracy caused by data sparsity. Huang Xianying[11] et al. proposed self-adaptive K-means cluster algorithm, in which introduced the topological potential field theory in physics. The topological potential value was used to represent the user's importance and obtain the user's influence range.

The classic collaborative filtering method uses shallow models to learn the deep features of users. As more and more data on the Internet can be perceived, multi-source heterogeneous data, such as videos, images, text, contains rich user behavior information and personalized demand information. The hybrid recommendation method is more and more important because it can alleviate the problem of data sparsity and cold start in the traditional recommendation system. The hybrid recommendation method combining multi-source heterogeneous data still faces serious challenges, because of the multi-modality, data heterogeneity, large-scale, data sparsity and uneven distribution.

The earliest proposed application of deep learning to the recommendation algorithm was the Restricted Boltzmann Machine proposed in the second half of the NetFlix Prize contest. ACM Recommended System Annual Meeting pointed out that deep learning will be the next important direction of recommendation system. In recent years, deep learning and recommendation algorithms had good combination. How to make the perfect integration of deep learning, parallel computing and
recommendation system, improving the accuracy of recommendation system by combing deep learning, is the general trend and general direction of future recommended algorithms.

2. The basic principle of KNN algorithm
The K nearest neighbor (KNN) model has been used in recommendation system algorithms very early and is still popular. The algorithmic idea of the KNN model is: First, calculate the similarity of the user or the item, and then select the K users or items with the highest similarity to the target user or item. A predictive score is obtained by finding k-nearest neighbors of the item using a linear weighted combination of k-nearest neighbors.

The K nearest neighbor algorithm steps:
1) Obtain training data and extract features;
2) Part of the data is used as the training data set, and the remaining data is used as test data;
3) Assume the k value, which is generally an odd number;
4) calculating k instances that are similar to the item to be classified;
5) Judging the categories of classified samples in the test data through majority voting rule, and comparing with the real values to determine the classification accuracy;
6) Change the k value, go to step 3, and finally select the k value with the highest classification accuracy.

3. The basic principle of XGBoost algorithm
Extreme Gradient Boosting(XGBoost) was proposed by Tianqi Chen in 2015 and has received extensive attention from industry and academia, which was based on the traditional Boosting Tree and Gradient Boosting Decision Tree algorithms. XGBoost method transformed several weak classifiers into one strong classifier, which can accurately predict. XGBoost method is a very effective integrated learning algorithm. The algorithmic idea of XGBoost is to select some samples and features to generate a simple model as the basic classifier. When generating a new model, learn the residuals of the previous model, minimize the objective function and generate a new model. This process is repeated and ultimately combines into a comprehensive model with high accuracy. The new model is established in the direction of the corresponding loss function gradient to correct residual.

4. Hybrid method
In order to solve data sparsity and cold start problems in the personalized recommendation system, this paper proposes a hybrid collaborative filtering recommendation algorithm, which combines the KNN model and the XGBoost model, and uses the scores predicted by the model-based personalized recommendation algorithm as features. The value, the true score value is used as the target value, and the hybrid model is trained to complete the recommendation. Regarding the hybrid method, the literature gives several suggestions, such as weighting method (weighting the recommendation results by different algorithms), feature combination method, and cascading method (the recommendation result generated by the previous recommendation algorithm is used as the current recommendation algorithm), feature expansion method, etc. This paper suggests a way to weight-mix the KNN model and the XGBoost model.

5. Conclusion
In the era of artificial intelligence, it is more and more convenient for users to obtain resources from platforms or systems, but it takes a long time to select the resources of interest from a large number of data resources. In order to solve the data sparsity and cold start problems in the personalized recommendation system, a hybrid collaborative filtering recommendation algorithm is proposed, which combines the KNN model and XGBoost model. When deep learning is applied to recommendation system by integrating massive multi-sources heterogeneous data, it could improve the performance of the recommendation system.
Acknowledgements

Supported by Hebei Statistical Scientific Research Program Project(Grant No.2018HZ03); Key R&D Projects in Hebei Province(Grant No.18210806); National Education Information Technology Research Project(Grant No.186140038); Research and Practice Project on Higher Education Reform in Hebei Province (Grant No. 2018GJJG450).

References

[1] Goldberg D, Nichols D, et al. Using Collaborative Filtering to Weave an Information Tapestry. Communications of the ACM, 1992, 35(12): 61-70.

[2] Resnick P, Iacovou N, Suchak M, et al. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. Proc of the ACM Conference on Computer Supported Cooperative Work. Chapel Hill, USA, 1994: 175-186.

[3] Konstan J A, Miller B N, Maltz D, et al. GroupLens: Applying Collaborative Filtering to Usenet News. Communications of the ACM, 1997, 40(3): 77-87.

[4] Gong S. A Collaborative Filtering Recommendation Algorithm Based on User Clustering and Item Clustering[J]. Journal of Software, 2010, 5(7):745-752.

[5] Li W, He W. An Improved Collaborative Filtering Approach Based on User Ranking and Item Clustering[M]// Internet and Distributed Computing Systems. 2013:134-144.

[6] Zhang J, Lin Y, Lin M, et al. An effective collaborative filtering algorithm based on user preference clustering[J]. Applied Intelligence, 2016, 45(2):230-240.

[7] Wang H, Shi X, Yeung D Y. Collaborative recurrent autoencoder: Recommend while learning to fill in the blanks// Proceedings of the Advances in Neural Information Processing Systems. Barcelona, Spain, 2016:415-423.

[8] Cao Zhanwei, Hu Xiaopeng. Recommendation algorithm combining theme model[J/OL]. Application Research of Computers, 2019(06):1-8.

[9] Zhang Hai-xia, Lv Zhen, Zhang Chuan-ting, Yuan Dong-feng. An Improved Collaborative Filtering Recommendation Algorithm with Weighted Heterogeneous Information[J]. Journal of University of Electronic Science and Technology of China, 2018, 47(01):112-116+152.

[10] Mi Cui, Chen Jing, Su Yan-yuan, Zhang Ya-ming, Liu Hai-ou. Trust Recommender Model for Mobile Social Network Integrating Cloud Environment User's Situational Interest[J]. Journal of Chinese Computer Systems, 2018, 39(03):484-489.

[11] Huang Xian-ying, Li Qin-dong, Xiong Li-yuan. Collaborative filtering recommendation algorithm with topology potential combined user clustering [J]. Computer Engineering and Design, 2018(1):90-95.

[12] Wang Junshu, Zhang Guoming, Hu Bin. A Survey of Deep Learning Based Recommendation Algorithms[J]. Journal of Nanjing Normal University(Engineering and Technology Edition), 2018, (4):33-43.

[13] Sedhain S, Menon A K, Sanner S, et al. AutoRec: Autoencoders meet collaborative filtering // Proceedings of the 24th International Conference on World Wide Web. Florence, Italy, 2015:111-112.

[14] Wang H, Shi X, Yeung D Y. Collaborative recurrent autoencoder: Recommend while learning to fill in the blanks// Proceedings of the Advances in Neural Information Processing Systems. Barcelona, Spain, 2016:415-423.

[15] Leng Ya-Jun, Lu Qing, Liang Chang-Yong. Survey of Recommendation Based on Collaborative Filtering[J]. Pattern Recognition and Artificial Intelligence, 2014, 27(8):720-734.

[16] Gong S. A Collaborative Filtering Recommendation Algorithm Based on User Clustering and Item Clustering[J]. Journal of Software, 2010, 5(7):745-752.

[17] Li W, He W. An Improved Collaborative Filtering Approach Based on User Ranking and Item Clustering[M]// Internet and Distributed Computing Systems. 2013:134-144.

[18] Zhang J, Lin Y, Lin M, et al. An effective collaborative filtering algorithm based on user preference clustering[J]. Applied Intelligence, 2016, 45(2):230-240.