Overview of Collaborative Filtering Recommendation Algorithms

Zhen Zhang¹,², *, Taile Peng², a and Ke Shen², b

¹College of Telecommunications, China University of Mining and Technology, Xuzhou, China
²School of Computer Science and Technology, Huaibei Normal University, Huaibei, China

*Corresponding author e-mail: 380869815@qq.com,ATAILEP@163.com, hbshenke@163.com

Abstract. In this paper, the key technologies of collaborative filtering recommendation algorithm and the bottleneck in its development are summarized, the problems of different technologies are analyzed. Besides, the application prospect of collaborative filtering technology is prospected.

1. Introduction

With the rapid popularization of e-commerce, social network, MOOC, P2P and cloud computing, the scale of Internet information resources is rapidly expanding [1]. Faced with the mass data, how to use intelligent technology to discover products and services that meet individual interests and preferences has become an important research topic [2]. Collaborative filtering recommendation is one of the most popular technology in current recommendation systems [3]. Collaborative filtering algorithm analyses user preferences and finds user sets with similar interests to designated users. According to the viewpoint of similar users in user set, the recommendation list of target users is deduced, and then user preferences are predicted. Collaborative filtering recommendation is based on the assumption that if users have similar ratings on some items, indicating that they have similar preferences, they also have similar ratings on other items.

In the industry, recommendation systems play an important role in various highly rated website platforms and e-commerce systems, such as Google News, Amazon.com, Netflix, Last.fm, YouTube, Spotify, Twitter, LinkedIn, and Facebook.

In academia, many domestic and foreign researchers and research institutions have carried out on innovative development and in-depth research in the field of recommendation systems. At present, recommendation system has developed into an independent and popular research discipline. In 2007, the Computer Association (acm) began to host the International Conference on Recommendation Systems. In many international conferences in the fields of human-computer interaction, information retrieval, data mining and machine learning, (such as ACM SIGIR, ACM SIGMOD, ACM SIGKDD, UMAP, IUI, WWW, ICML, ICDM, etc.), the number of academic papers on recommendation system is also increasing. Some high-level academic journals have also set up special journals on recommendation systems to collect and publish the latest research results in this field, such as: in 2007, IEEE Intelligent Systems; in 2008, AI Communications, in 2012 and 2014, User Modeling and User-
Adapted Interaction, in 2013, ACM Transactions on Interactive Intelligent Systems, in 2015, ACM Transactions on Intelligent Systems and Technology. In 2015, Springer published the book of Recommender Systems Handbook (second edition). It introduced in detail and systematically the main concepts, theoretical basis, technical methods, development trends, future challenges and practical application of the recommendation system.

2. Main Challenges of Collaborative Filtering Recommendation System

Recommendation system has prominent performance in stimulating users' purchase enthusiasm and promoting enterprise revenue, which has attracted more and more attention of enterprises. However, with the rapid growth of Internet data and the changing application scenarios, the demand for applications in various fields is also increasing. There are still many new problems and challenges to be solved urgently in recommendation system, such as Data Sparsity [4], Cold Start Problem, scalability [5], accuracy and diversity [6-7], Top attack problem [8-9], context awareness problem [10], privacy protection problem [11], etc.

2.1. Data Sparsity

The problem of data sparsity is caused by insufficient or even missing useful information. With the rapid development of various application platforms and systems in the Internet, the number of users, especially the number and types of projects and services, has increased explosively. At the same time, in the face of massive services and projects, the rating information or useful historical information is relatively limited, and the useful historical information matrix between users and projects is often very sparse, even with the increase of users and projects, it will show super high-dimensional and uneven distribution.

The cold start problem in recommendation system is equivalent to the extreme case of data sparsity problem. Because of the lack of user feedback data, the system cannot predict based on historical information or give a valid recommendation [12]. Therefore, In this case of high sparsity, it can neither accurately describe users' interest preferences nor accurately calculate the similarity of their interests, which will lead to the decline of recommendation quality and accuracy.

In order to alleviate this problem, the main solutions are: converting other auxiliary information into scoring information, so that relevant recommendations can be given to each user based on available data, such as recommendation algorithm based on demographic information, recommendation algorithm based on social relations and hybrid recommendation algorithm [13-16]; fill in the vacancy value of the original sparse matrix [17]; use dimensionality reduction technology to extract features to make the data mapped to low-dimensional space relatively dense [18]; consider the relevance of second-order or higher-order global information [19-20]; combine with content-based recommendation methods [21]; improve the accuracy of similarity measurement methods [22-23].

2.2. Cold Start Problem

New users and new objects are cold start problems. In recommendation system, especially in collaborative filtering system, new objects must wait for a period of time before users can view and evaluate clicks, scores, comments and so on. It is impossible to analyze and recommend this object before the evaluation reaches a certain number. Different from the new user problem, this kind of problem is generally considered to use the combination recommendation method to deal with. Paper [24-25]: Using object entropy, popularity and user personality attributes to improve the effect.

2.3. Scalability

For large-scale Internet sites, recommendation systems usually need to face more than one million users and commodities, and at the same time need to search the similarity in the whole user space in real time, so as to complete the recommendation for target users. Therefore, when the number of users and products in recommendation system increases rapidly, the traditional recommendation system faces severe challenges in real time and scalability of the algorithm. Many scholars' research results show that the recommendation accuracy and real-time performance of the recommendation system are usually a
contradiction, it is difficult for the recommendation system to meet the real-time requirements while ensuring high-quality recommendation accuracy. Therefore, on the premise of guaranteeing the best recommendation quality, how to ensure the real-time and scalability of the recommendation system is one of the challenges faced by the recommendation system.

In a word, the problems mentioned above restrict the further development and application of recommendation system, which is a difficult problem and challenge that needs to be paid close attention to.

3. Comparison of Nearest Neighbor Selection Methods
Computing the similarity of users or items is an important part of collaborative filtering recommendation algorithm. In collaborative filtering recommendation technology, there are several commonly used methods to calculate the similarity:

3.1. Pearson Correlation Similarity
Assuming that the item set scored by user(i) and user(j) is \( I_{ij} \), the similarity \( \text{sim}(i,j) \) can be obtained by Pearson correlation:

\[
\text{sim}(i,j) = \frac{\sum_{c \in C} (R_{ic} - \bar{R}_i)(R_{jc} - \bar{R}_j)}{\sqrt{\sum_{c \in C} (R_{ic} - \bar{R}_i)^2 \sum_{c \in C} (R_{jc} - \bar{R}_j)^2}}
\]

(1)

\( R_{ic} \) and \( R_{jc} \) represent the score of user(i) and user(j) on item c, \( \bar{R}_i \) and \( \bar{R}_j \) represent the average score of user(i) and user(j) on all items.

Because Pearson correlation coefficient is obtained by linear regression formula, it is necessary to satisfy the assumptions of linear relationship between the data and that the residual is independent of each other and the mean value is zero. When these conditions are not satisfied, the accuracy of calculation will be reduced.

3.2. Jaccard coefficient
Jaccard coefficient is used to measure the degree of overlap of binary data. It is defined as follows:

\[
\text{sim}(i,j) = \frac{|R_i \cap R_j|}{|R_i \cup R_j|}
\]

(2)

In e-commerce, Jaccard coefficients are usually used to compare shopping cart data of different users, and this similarity measure, which is only applicable to binary type, also limits its further application in recommendation system.

3.3. Cosine Similarity
Cosine similarity is often used in the field of information retrieval. The similarity is measured by calculating the angle between document vectors.

\[
\text{sim}(i,j) = \cos(I,J) = \frac{\sum_{c=1}^{n} R_{ic}R_{jc}}{\sqrt{\sum_{c=1}^{n} R_{ic}^2 \sum_{c=1}^{n} R_{jc}^2}}
\]

(3)

In practice, different scales of user ratings are inconsistent, some users tend to score high, while some users tend to score low, then cosine similarity can not accurately measure the similarity between users.

3.4. Modified Cosine Similarity
The common cosine similarity measurement method does not consider that different users may have different scoring scales. Therefore, the modified cosine similarity improves the defect and eliminates the difference of rating scales by considering the average of user ratings.
sim(i, j) = \frac{\sum_{c=1}^{n}(R_{i,c} - \bar{R}_i)(R_{j,c} - \bar{R}_j)}{\sqrt{\sum_{c=1}^{n}(R_{i,c} - \bar{R}_i)^2 \sum_{c=1}^{n}(R_{j,c} - \bar{R}_j)^2}} \tag{4}

4. Comparison of Recommendation Methods

4.1. Average Weighting Strategy

A basic assumption of collaborative filtering algorithm is that users with similar preferences will give similar ratings for the same project. Therefore, after the target user's nearest neighbor set is generated, it can predict the target user's score for the non-scored items according to the user's score in the nearest neighbor set.

\[ P_{u,i} = \frac{\sum_{v \in \text{KNB}} \text{Sim}(u,v) \times R_{v,i}}{\sum_{v \in \text{KNB}} \text{Sim}(u,v)} \tag{5} \]

At present, most collaborative filtering recommendation systems use average weighting strategy to generate recommendation.

\[ P_{u,i} = \bar{R}_u + \frac{\sum_{v \in \text{KNB}} \text{Sim}(u,v) \times (R_{v,i} - \bar{R}_v)}{\sum_{v \in \text{KNB}} \text{Sim}(u,v)} \tag{6} \]

\( P_{u,i} \) represents the predictive score of user(u) for item(i), and KNB represents the nearest neighbor set of target user(u).

4.2. Top-N Recommendation Strategy

Top-N recommendation strategy is to calculate the weighted average of user(i)'s interest in different items in the "nearest neighbor" set, and take N items which rank first and do not belong to \( I_i \) (\( I_i \) represents the item set of user(i)'s score) as the Top-N recommendation set.

5. Comparisons of recommended quality assessment methods

The advantages and disadvantages of a recommendation system are measured by its prediction results. At present, the evaluation strategies commonly used in collaborative filtering recommendation algorithms are as follows.

5.1. Average Absolute Error (MAE)

Average absolute error is the most widely used evaluation method in the recommendation system. It is obtained by calculating the absolute error between the predicted value and the actual value.

\[ \text{MAE} = \frac{\sum_{i,j} |P_{ij} - r_{ij}|}{n} \tag{7} \]

Variable \( n \) is the total score, \( P_{ij} \) represents user(i)'s predictive score for item j, \( r_{ij} \) represents user(i)'s actual score for item(j). The smaller the MAE value, the higher the recommendation accuracy.

5.2. Root Mean Square Error (RMSE)

RMSE, also known as standard square deviation, reflects the discreteness of scoring data.

\[ \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i,j} (P_{ij} - r_{ij})^2} \tag{8} \]

The smaller the RMSE value, the higher the recommendation accuracy.
5.3. Recall  

The recall rate is used to reflect the recommendation rate of the project to be recommended.

\[
\text{Recall} = \frac{|\text{test}\cap\text{top-N}|}{|\text{test}|}
\]  

(9)

The test represents the number of items in the test data set, the top-N represents the N items recommended by the system to users. The greater the Recall value, the greater the probability of recommendation.

6. Conclusion  

This paper mainly introduces the commonly used nearest neighbor selection algorithm, recommendation algorithm and evaluation strategy of recommendation algorithm when using collaborative filtering algorithm for recommendation. Collaborative filtering recommendation algorithm mainly has the problems of data sparsity, cold start and robustness, as well as the recommendation efficiency in large data environment. Researchers have proposed a variety of solutions, the most common one is to introduce methods from other fields, and the interdisciplinary research of collaborative filtering has been further developed. With the rapid growth of information on the Internet, improving collaborative filtering recommendation based on large data is a formidable challenge.

Acknowledgments  

This paper is jointly supported by Key Natural Science Research Projects in in Colleges of Anhui China (grant numbers KJ2016A630 and KJ2017A843).

Zhang Zhen (Correspondent Author) received her B.Sc. degree from Huaibei Coal Normal University in 1999 and M.Sc. degree from Hefei University of Technology in 2007, at present, she is studying for her doctorate in China University of Mining and Technology. She is now an associate professor of Huaibei Normal University. Her main research interests include recommender system and machine learning.

Peng Taile (Corresponding author) received his B.Sc. degree in 1998 from Jilin University of Technology and M.Sc. degree in 2008 from Hefei University of Technology and Ph.D. degree in 2016 from Shanghai University; Now he is a professor in Huaibei Normal University. His main research interests include video understanding and machine learning, and etc.

References  

[1] Eppler M J, Mengis J. The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines [J]. The Information Society, 2004, 20 (5): 325-344.

[2] Hors-Fraile S, Rivera-Romero O, Schneider F, et al. Analyzing recommender systems for health promotion using a multidisciplinary taxonomy: A scoping review [J]. International Journal of Medical Informatics, 2017, 114: 143-155.

[3] Ricci F, Rokach L, Shapira B. Recommender Systems: Introduction and Challenges [M]. In: Ricci F, Rokach L, Shapira B. (Eds) Recommender Systems Handbook (Second Edition), Springer US, 2015: 1-36.

[4] Adomavicius G, Tuzhilin A. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions [J]. IEEE Transactions on Knowledge and Data Engineering (TKDE), 2005, 17 (6): 734-749.

[5] Luo X, Xia Y N, Zhu Q S. Incremental Collaborative Filtering Recommender Based on Regularized Matrix Factorization [J]. Knowledge-Based Systems, 2012, 27: 271-280.

[6] Karakaya M Ö, Aytekin T. Effective methods for increasing aggregate diversity in recommender systems [J]. Knowledge & Information Systems, 2017 (5): 1-18.

[7] Sánchez P, Bellogín A. Time-Aware Novelty Metrics for Recommender Systems [C]. Proceedings of the European Conference on Information Retrieval, Springer, Cham, 2018,
357-370.

[8] Yang Z H, Sun Q D, Zhang Y L, et al. Uncovering Anomalous Rating Behaviors for Rating Systems [J]. Neurocomputing, 2018, 308: 205-226.

[9] Zhang F, Lu Y, Chen J, et al. Robust collaborative filtering based on non-negative matrix factorization and R1-norm [J]. Knowledge-Based Systems, 2016, 118 (C): 177-190.

[10] Ferdousi Z V, Colazzo D, Negre E. CBPF: leveraging context and content information for better recommendations [C]. Proceedings of the 12th ACM Conference on Recommender Systems (RecSys'18), 2018, arxiv: 1810.00751v1.

[11] Kaur H, Kumar N, Batra S. An efficient multi-party scheme for privacy preserving collaborative filtering for healthcare recommender system [J]. Future Generation Computer Systems, 2018, 86: 297-307.

[12] Adomavicius G, Tuzhilin A. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions [J]. IEEE Transactions on Knowledge and Data Engineering (TKDE), 2005, 17 (6): 734-749.

[13] Camacho L A G, Alves-Souza S N. Social network data to alleviate cold-start in recommender system: A systematic review [J]. Information Processing & Management, 2018, 54 (4): 529-544.

[14] Guo G, Zhang J, Yorke-Smith N. A Novel Recommendation Model Regularized with User Trust and Item Ratings [J]. IEEE Transactions on Knowledge and Data Engineering (TKDE), 2016, 28 (7): 1607-1620.

[15] Guo G B, Zhang J, Yorke-Smith N. TrustSVD: Collaborative Filtering with Both the Explicit and Implicit Influence of User Trust and of Item Ratings [C]. Proceedings of 29th AAAI Conference on Artificial Intelligence (AAAI'15), 2015, 123-129.

[16] Guo G B, Zhang J, Zhu F D, et al. Factored similarity models with social trust for top-N item recommendation [J]. Knowledge-Based Systems, 2017 (122): 17-25.

[17] Breese J S, Heckerman D, Kadie C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering [C]. Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI'98), 1998, 461 (8): 43-52.

[18] Zhang S, Wang W H, Ford J, et al. Using singular value decomposition approximation for collaborative filtering[C]. Proceedings of the 7th IEEE International Conference on E-Commerce Technology (CEC'05), 2005, 257-264.

[19] Zhou T, Su R Q, Liu R R, et al. Accurate and diverse recommendations via eliminating redundant correlations [J]. New Journal of Physics, 2009, 11 (12): 123008.

[20] Liu J G, Zhou T, Che H A, et al. Effects of high-order correlations on personalized recommendations for bipartite networks [J]. Physica A: Statistical Mechanics and Its Applications, 2010, 389 (4): 881-886.

[21] Karimi M, Jannach, D, Jugovac M. News recommender systems–Survey and roads ahead [J]. Information Processing & Management, 2018. doi:10.1016/j.ipm.2018.04.008.

[22] Choi K, Suh Y. A New Similarity Function for Selecting Neighbors for Each Target Item in Collaborative Filtering [J]. Knowledge-Based Systems, 2013, 37 (1): 146-153.

[23] Karel C. An Entropy-Based Neighbor Selection Approach for Collaborative Filtering [J]. Knowledge-Based Systems, 2014, 56: 273-280.

[24] A. Kohrs and B. Merialdo. Clustering for Collaborative Filtering Applications. In Proceedings of CIMCA'99. IOS Press, 1999.

[25] C. Zeng, C. –X. Xing, L. –Z. Zhou. Similarity Measure and Instance Selection for Collaborative Filtering. In Proceedings. Of 12th International Word Wide Web conference, 2003.