

Research on Recommendation Technology based on Collaborative Filtering

Yi Ren^{1, a*} and Cuirong Chi^{2, b}

¹Shenyang Jianzhu University, Liaoning, China

²Shenyang Jianzhu University, Liaoning, China

^{a*}1181134429@qq.com, ^b726733255@qq.com

Keywords: Collaborative Filtering; Recommender Systems; Recommendation Algorithms

Abstract. Collaborative filtering is the most widely used technology in recommender systems. Researchers have done a lot of work on how to improve the collaborative filtering technology. However, there are few summaries about collaborative filtering technology which is reviewed in this paper. First of all, the different methods of collaborative filtering technology are studied and analyzed. Secondly, we compare the advantages and disadvantages of different methods. Then, the problems about collaborative filtering algorithm are proposed. Finally, the recommendation algorithms are summarized and some work will be done.

Introduction

With the rapid development of internet and information technology, the volumes of global data have exploded. People get more resources and have to spend more time and energy searching for available information.

Recommender system is an important way to solve this problem. As a new way of intelligent information service, it is widely used in e-commerce and social network, which can greatly improve the sales ability of e-commerce. For example, the Amazon personalized recommender system boost nearly 35% of the sales of its company.

Recommendation technology currently used by recommender system is mainly Content-Based Recommendation, Collaborative Filtering Recommendation and Hybrid Recommendation. The most successful recommendation algorithm is collaborative filtering recommendation proposed in KDD Cup and Netflix competition [1,2].

This paper analyzes collaborative filtering recommendation algorithms. The second section mainly introduces the collaborative filtering algorithm and its advantages and disadvantages. The third section proposes the problems with collaborative filtering. The fourth section gets some conclusions.

Collaborative Filtering Recommendation

The term "collaborative filtering" was first proposed by Goldberg et al. in document [3], which introduced the collaborative filtering recommender system *Tapestry* for the first time in detail. Collaborative filtering algorithm is a recommendation algorithm based on user behavior data. In the academic field, these algorithms have been deeply studied, and many methods have been proposed, including neighborhood-based method and model-based method. The following is an introduction to the collaborative filtering algorithms.

Neighborhood-based Method. The neighborhood-based algorithm is the most basic algorithm in the recommender systems. The algorithm is not only studied in the academic world, but also widely used in the industry. Different from basic assumptions, which can be divided into two categories: User-based collaborative filtering and Item-based collaborative filtering. The following details:

User-based Collaborative Filtering(UserCF). The idea of collaborative filtering algorithm based on users is to recommend other users' favorite items similar to their interests. It mainly includes two steps:(1)Find a user set that is similar to the target user's interest;(2)Find items that users like in this

collection and target users with no historical behavior to recommend to the target users. The crux of the algorithm is to calculate the interest similarity among two users. Collaborative filtering algorithm mainly uses user behavior similarity to calculate the interest similarity. u and v represent given users, $N(u)$ represents a set of historical data positive feedback items for user u , $N(v)$ represents a set of historical data positive feedback items for user v . The common formula for similarity calculation is as follows.

Jaccard:

$$S_{uv} = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|} \tag{Eq.1}$$

Cosine similarity:

$$S_{uv} = \frac{|N(u) \cap N(v)|}{\sqrt{|N(u)| |N(v)|}} \tag{Eq.2}$$

User interest similarity is figured by UserCF, and user behavior records are shown in Figure 1. User A has acted on the items $\{a, b, c\}$, user B has acted on the items $\{a, d\}$, and cosine similarity is used to

figure the interest similarity between user A and user B: $S_{AB} = \frac{|\{a, b, c\} \cap \{a, d\}|}{\sqrt{|\{a, b, c\}| |\{a, d\}|}} = \frac{1}{\sqrt{6}}$.

A	a	b	c
B	a	d	
C	b	d	e

Figure 1. UserCF

Item-based collaborative filtering (ItemCF). The Item-based collaborative filtering algorithm is currently widely used in Amazon, Netflix and YouTube platforms, and its algorithm idea is an item that recommends to the target user similar to the item he liked before. It mainly consists of two steps: (1) Calculating the similarity between items; (2) Generating a recommendation list for the user based on the similarity of the items and the historical behavior of the user.

A formula used to figure the items similarity is as follows.

$$S_{ij} = \frac{|N(i) \cap N(j)|}{|N(i)|} \tag{Eq.3}$$

Denominator $|N(i)|$ is the number of users who prefer item i . Molecular $|N(i) \cap N(j)|$ is the number of users who prefer i and item j at the same time. Therefore, the above formula can be understood as the proportion of users who prefer item i to the number of users who also prefer item j .

Table 1 Comparison of advantages and disadvantages between UserCF and ItemCF

	UserCF	ItemCF
Performance	Suitable for less users.	Suitable for less items.
Field	Areas with stronger timeliness and less personalized user interest.	The long tail items are rich, users with strong personalized needs.
Real-time	New behaviors of users do not necessarily cause real-time changes in recommendation results.	New behaviors of users will lead to real-time changes in recommendation results.
Cold-start	New users have fewer ratings on items and cannot personalize them immediately. As long as the user has scored the item, the new item can be recommended to the target user.	As long as a new user has a rating on an item, he recommends a similar item to the project. The new item is recommended to users only if the item similarity table is updated offline.

Model-based Method. Although the basic idea of neighborhood-based collaborative filtering is clear, simple to implement, and effective, it requires a search on the entire database. This causes a huge burden on the system when users and item volumes increase. For the problems of neighborhood-based collaborative filtering algorithms, such as poor scalability and sensitivity to data sparseness, collaborative filtering based on models can mitigate to a large extent. Another method of collaborative filtering recommendation is model-based collaborative filtering. This method uses probability statistical model or machine learning method to construct user feature model (cluster model, matrix factorization model, graph model, etc.) on training set, and recommends it.

Matrix Factorization Model (MF). In recent years, the matrix factorization method [4] is the core algorithm, which has been able to factorize the user-items rating matrix into the item of two or more low dimensional matrices, in order to implement the specification of the dimension, and to study the properties of the high dimensional data with the low dimensional spatial data. It mainly includes singular value decomposition, nonnegative matrix factorization (NMF) and probability matrix factorization (PMF).

Nonnegative matrix factorization [5] is the most successful matrix factorization method. In this method, the user rating matrix $R_{n \times m}$ of the items can be factorized into two real valued nonnegative matrices $U_{n \times k}$ and $V_{k \times m}$, making $R \approx U^T V$.

$$U = \begin{bmatrix} U_{1,1} & \dots & U_{1,n} \\ \dots & U_{k,x} & \dots \\ U_{k,1} & \dots & U_{k,n} \end{bmatrix}, V = \begin{bmatrix} V_{1,1} & \dots & V_{1,m} \\ \dots & V_{k,i} & \dots \\ V_{k,1} & \dots & V_{k,m} \end{bmatrix} \tag{Eq.4}$$

Among them, $k < \min(m, n)$, The item $U_x^T V_i$ of the i column vector V_i of the x column vectors U_x , and $V_{k \times m}$ of the matrix $U_{n \times k}$ represents the user's prediction rating for the items, call U_x as the latent feature vector of the user x , V_i is the latent feature vector of the item i . Finally, the optimal solution $U_{n \times k}$ and $V_{k \times m}$ can be obtained by gradient descent method.

Graph model. The recommendation method based on graph model mainly uses the topology structure and basic feature information of user-item two sub graph or user social relation map, and uses graph topology model analysis, sub graph division, other network analysis methods and semi supervised learning methods to design recommendation algorithm. In document [6], which introduced that it is used to express the relationship between users and items in the form of graphs. Because of the intuition and image expressed by graph model, new relationships can usually be mined. The shortcomings of the graph model method are similar to the new similarity method, and it does not change the number of preference data, and does not solve the sparsity problem in essence.

When studying graph models, firstly, we need to represent user behavior data into graphs. Let $G(V, E)$ represents the two sub diagram of the user-item, $V = V_U \cup V_I$ is composed of user vertex set V_U and item vertex set V_I . For each two-tuples (u, i) of a data set, there is a $e(v_u, v_i)$ corresponding edge in the graph, in which $v_u \in V_U$ is the vertex of the user u . $v_i \in V_I$ is the vertex of the item i .

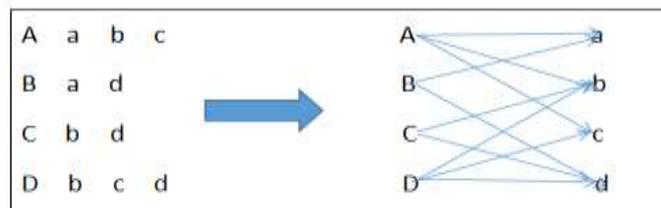


Figure 2. Two sub graph model for user-item

Figure 2 is a user-item graph model. A, B, C and D represent users, a, b, c, d, e represent items. The node A is connected with a, b and d to represent user A and act on items a, b and d. The weight of the final recommendation list is the access probability of the item node.

Table 2 Comparison of matrix decomposition model and graph model

	Advantages	Disadvantages
Matrix Factorization model	1 Low complexity 2 Good prediction effect 3 Solving the problem of data sparsity	Poor interpretability
Graph model	1 Express intuition and image 2 New relationship can be excavated	Data sparsity problem

Existing Problems of Collaborative Filtering Recommendation Algorithm

With the development of Web 2.0 and the explosive growth of data volumes, recommender system has become an indispensable tool in the industry. However, by analyzing collaborative filtering recommendation algorithm, it can be seen that the application of collaborative filtering algorithms in the recommendation system cannot bring better results, the reasons are as follows:

(1) Data sparsity

The recommended techniques that perform well off-line testing on small-scale dataset usually do not work when implemented on actual large-scale dataset. At the same time, the actual system not only has a large amount of data, but new users and new products will continue to enter the system, and users and items will constantly generate new connections (favorites, reviews, purchases, etc.). When there are many items, the rating matrix of the user-items and their sparsity make it difficult to discriminate similarities between users.

(2) Cold-start problem

There is a cold-start problem for new users and new items. If the new user registration does not have historical behavior, the recommender systems can not recommend the items that users prefer correctly. At the same time, without clicking the new item, it is impossible to judge whether the item is recommended to specific target users.

(3) Extensibility

With the increasing number of users and items, the complexity of collaborative filtering algorithm is increasing rapidly, and it will lack scalability. Many online systems need to recommend users in real-time after considering users' purchase and evaluation history, which cannot meet the real-time requirements.

(4) Cross-domain Recommendation

Cross-domain recommendation refers to the integration of data from different systems to implement recommendations. Everyone in real life will enter many systems and leave information in them. It is possible to use a behavioral record in one system to predict user expectations in another system [7] or to integrate information provision recommendations for different systems. How to reasonably integrate data from different systems to achieve personalized recommendation in the context of big data era is a hot spot, and it is also a difficult point [8].

Summary

Some of the classic collaborative filtering algorithms summed up in this paper can understand some advantages of collaborative filtering algorithm, such as strong model generality and simple algorithm. Collaborative filtering algorithm still has some problems in data sparsity and cold-start. The future work will take into account the integration of multielement information (labels, social information, timing, etc.) on the basis of a collaborative filtering algorithm, in order to improve the recommendation quality of the recommender systems.

References

- [1] G Dror, N Koenigstein, Y Koren and M Weimer. *Journal of Machine Learning Research*. Vol.18(2012),p.3.
- [2] RM Bell and Y Koren. Lessons from the Netflix prize challenge. *Acm Sigkdd Explorations Newsletter*. Vol.9(2007)No.2,p.75.
- [3] Goldberg D, Nichols D, and Oki B M, et al. Using Collaboration Filtering to Weave an Information Tapestry. *Communications of the Association for Computing Machinery*, Vol. 35(1992) No.12, p.61.
- [4] X.W. Meng, S.D. Liu, and Y.J. Zhang, et al. Social recommendation system research[J]. *Journal of Software*, Vol. 26(2015) No.6, p.1356-1372 .(In Chinese).
- [5] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Journal of Nature*, Vol.401(1999) No.6755,p.788.
- [6] Aggarwal C C, Wolf J L, Wu K L, et al. *Proc.of KDD-99 :the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (San Diego, Calif, USA, Aug 15-18, 1999). p.201.
- [7] Nozomi Nori, Danushka Bollegala and Mitsuru Ishizuka. *Proc.of International AAI Conference on Weblogs and Social Media (ICWSM 2011) 5th*.(2011). p.241.
- [8] Gediminas Adomavicius and YoungOk Kwon. Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques. *IEEE Transactions on Knowledge and Data Engineering*. Vol.24(2012)No.5,p.896.