Personalized recommendation system K- neighbor algorithm optimization

Hua-Ming Wang
School of Information Science and Engineering
Hangzhou Normal University
Hangzhou, Zhejiang Province, China
whm320@126.com

Ge Yu
School of Information Science and Engineering
Hangzhou Normal University
Hangzhou, Zhejiang Province, China
yuge@hznu.edu.cn

Abstract—**This paper mainly discussed the recommended user-based collaborative filtering algorithm, and deeply studied the recommended results that generated were from the modeling system. Furthermore, we pointed out the problems in the K-nearest neighbors algorithm, and proposed improvement on it. Additionally, this paper presented the merits and demerits of personalized recommendation algorithm. Finally it notes the future development focus.**

Keywords-Personalized Recommendation; Collaborative filtering; K-nearest neighbors

I. INTRODUCTION

In the 21st century,Information has become closely akin to the living things with us.While we swim in a sea of information, but also faced with contradictory information overload of information between the individual needs of.How can the huge amount of information required to meet their network needs the world to find information is a hot research topic in recent years in the field of recommender systems.In the context of the era of big data, the demand for "personalized" more urgent.

Collaborative filtering technology appears to solve the problem that brought hope. Collaborative filtering is the more popular and sophisticated recommendation technology, numerous advantages exist, but because of the need for items based on user data, and user ratings of some of the information itself, leading to collaborative filtering are still many problems to be solved, such as cold start problem sparse and poor real-time, lack of scalability and other issues, these problems seriously affect the accuracy of the recommendation system.

This paper describes the collaborative filtering algorithm accordingly, and make improvements to the algorithm of the K-Nearest Neighbors. In order to reduce the computational complexity of similarity improve the recommendation accuracy.

II. PERSONALIZED RECOMMENDEATION SYSTEM

A. Background of the recommendation system

The 21st century, any of us can not escape the impact of the wave of information, whether as producers or consumers of information we are facing more and more challenges.

Recommended system back in the 1990s had already seen, but its real concern to the industry is subject to later began to spread in e-commerce.With the exponential growth of Internet data, we can choose from more and more information, and how to find information on in this mass of information is particularly important that we really need. It is recommended that the system will come into being.Collaborative filtering which its excellent speed and robustness in the hot field of recommender systems, the algorithm was presented in 1992 and first used in e-mail filtering system in 1994 was GroupLens for news filtering.

In fact, the collaborative filtering algorithm is simple and easy to understand, and has long been, its core idea is "Like attracts like, people in groups." For example, we often to the same laboratory and our seniors advice, ask them to recommend some of us are interested in all kinds of books. Of the case, because we recommend to them that there is a trust, and that trust comes from our common interests and research directions.

B. Collaborative filtering recommdndation

Collaborative filtering recommendation system is the most famous method, it is mainly through the historical analysis of the behavior of the user and the user's interest to make a recommendation to the user.There are a lot of collaborative filtering algorithms, more common is the K-nearest neighbors algorithm (User-CF and ItemCF etc.), matrix factorization algorithm (or Latent Factor Model as RSVD and SVD ++, etc.) and graph algorithms.Simply put, collaborative filtering technology is in the context of the Internet, we work together, through constant interaction and to filter out sites they do not like the West. A typical collaborative filtering technology which can be divided mainly based on user-based collaborative filtering and collaborative filtering project.With traditional content-based filtering recommendation different collaborative filtering analysis of user interest, find similar (interest) specified by the user in the user base of users, the combination of these similar information of a user evaluation, the formation of this system to the specified user information predict the degree of preference. Collaborative filtering has the following advantages:

(1) can be difficult to machine automatically filtered based on the information content analysis. Such as art, music;

(2) can be based on a number of complex, difficult to express the concept of (information quality, taste) filtering;

(3) recommended novelty

0105

Because of this, the collaborative filtering in commercial applications are achieved good results. Amazon, CDNow, MovieFinder, have adopted a collaborative filtering technology to improve the quality of service.

III. USER-BASED COLLABORATIVE FILTERING ALGORITHM

A. The Recommended Principle

User-based collaborative filtering project is by far the most successful practical application of personalized recommendation technology, the basic idea is to have the same hobby of interest to the user to recommend to the target user.If the target user for the evaluation of the project and his "nearest neighbor" is similar to the target user for a comprehensive evaluation of a project can be obtained from the evaluation of his k-Nearest, as shown in Figure 1:

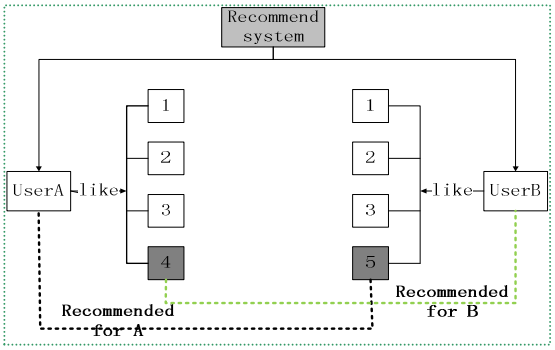

Figure 1: user-based collaborative filtering recommendation schematic

B. Algorithm model

User-based collaborative filtering typically the nearest neighbor technique, the use of historical user taste information to calculate the distance between users, and then use the target user's "nearest neighbors (k-Nearest)" Evaluation of weighted evaluation values to predict the target commodity users preferences for specific commodities extent, the system thus to make recommendations based on the preferences of the target user extent.Figure II below, showing the target user is the center of K = 4 users recently been selected as neighbors.

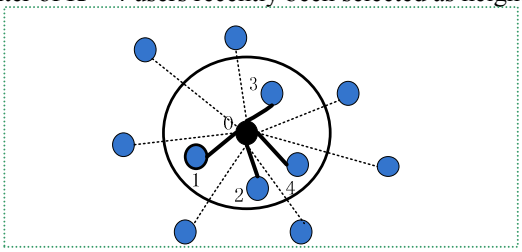

Figure 2: The nearest neighbors

User-based collaborative filtering is based on the assumption: find other users with similar interests and his target users. Recommended as the target user is then generated based on the neighbor set.According to the principle of collaborative filtering technology, user-based collaborative filtering can be divided into three stages:

1) Build a user item rating matrix. User to set the number m, the number of items is n, then the user of the item's score score data matrix can be constructed, as shown in Table 1, wherein R (m * n) indicates that the user m to n items of scores.

	I_1	. . .	I_n
U_1	r_{11}	. . .	R_{1n}
. . .	. . .	. . .	. . .
U_m	r_{m1}		r_{mn}

Table I: User scoring matrix

2) generate nearest neighbor. The scoring matrix as a user on the set of scores for all items, then each user can be expressed as a score vector. Similarity can be calculated by the formula for similarity measure between the user and the user. Throughout the user item rating matrix, calculating a similarity between each user, the size of the arrangement according to the similarity, select the highest similarity with the target user's K-Nearest users as nearest neighbor set. The following formula, for example cosine similarity measure to solve the K nearest neighbors.

Cosine similarity measure within two vectors by the cosine of the angle between the product space to measure the similarity between them. Cosine of 0 degrees is 1, and any other cosine value of the angle of not greater than 1; and the minimum value is -1. When the two vectors have the same orientation, the cosine similarity is 1; angle between two vectors is 90 °, the cosine similarity value is 0; when the two vector points to the opposite direction, the value of the cosine similarity -1. That is, the greater the similarity between users, the closer the value of their cosine 1. Set user A and user B in n-dimensional vector space vector score is expressed as:

$$a=(r_{a1},r_{a2},\ldots\ldots,r_{an})\quad b=(r_{b1},r_{b2},\ldots\ldots r_{bn})$$

Equation 3.1 shows the cosine of the acquaintance between user A and user B is calculated:

$$\mathrm{sim}(a,b)=\cos(a,b)=\frac{\sum_{i=1}^{n}R_{ai}\cdot R_{bi}}{\sqrt{\sum_{i=1}^{n}R_{ai}^2}\sqrt{\sum_{i=1}^{n}R_{bi}^2}} \qquad (3.1)$$

Where sim (a, b) represents a cosine acquaintance of the user between the user b, R_{ai}, R_{bi} represent user a, b of the project i score.

3) produce recommendations. After the target user generated nearest neighbor set, you can calculate the target user for the project did not score scores, according to recent data neighbor set score to predict the target user ratings for ungraded items typically used similarity weighted average formula. Users set up a neighbor set of users to Ma, the target users for a project i did not score prediction score recorded as P_{ai}, then P_{ai} is calculated as shown in Equation 3.2 is calculated.

$$P_{a,i}=\frac{\sum_{b\in M_a} sim(a,b)\times R_{b,i}}{\sum_{b\in M_a}|sim(a,b)|} \qquad (3.2)$$

Where sim (a, b) indicate a similarity of users of its neighbors user b,R_{bi} represents the user b project a score for item i.

IV. ADVANTAGES AND DISADVANTAGES OF THE ALGORITHM

The User-based collaborative filtering algorithm is one of the most successful recommendation algorithm, and its biggest advantage is the characteristic property does not require analysis of the project, there is no special requirements on the recommendation system that can handle unstructured projects. In general, e-commerce recommendation system is often carried out in a challenging environment and development, especially for large online shopping sites such as Amazon, Taobao. Typically, a fast and accurate recommendation system will cause the user's interest and for the benefit of companies.

But with the prosperity and the expansion of the Internet, more and more projects and users will join in, the entire user - item rating matrix dimension is growing. This not only increases the complexity of being solved with the nearest neighbors for recommendation system accuracy, timeliness bring great distress. The following four aspects of the problem is solved:

1)Sparsity of data. In fact, many large e-commerce sites often have a large number of items of information needs of the user ratings, such as Amazon Web site, but most users will assess the millions of the book of 1% -2%. So for the user - the project will be extremely sparse matrix, thus affecting the recommendation results.

2)Accuracy problem, namely to improve confidence in the quality of the user's recommendation. Users need to be worthy of their trust in a recommender system to recommend the project for them. If a recommendation system often does not meet the recommended project needs of users, users will lose trust in your site, it will cause the loss of a large number of users.

3)Cold start problems. Cold start problem in collaborative filtering techniques using the recommended system is the most prominent, as collaborative filtering techniques typically require only (user, project) to score information, without the need for any additional information. If a project that no one or very few people to evaluate it, or someone barely evaluated items, this system can not provide good personalized recommendations.

4)Scalability problem: With the expansion of the scale users and projects, most of the existing collaborative filtering techniques are faced with computational problems. Especially in the face of thousands of data, most algorithms will suffer serious scalability issues. But with the introduction of the continuous improvement in computing power and cloud computing concepts, scalability problem is relatively not particularly serious.

V. IMPROVEMENTS

Now I already know how to build a data model and generating process K-Nearest Neighbors. But in the process of solving K-Nearest Neighbors still exist in some places can be optimized.

A.*K-Nearest Neighbors improvements*

In 3.2, we according to the user - item rating matrix, for the target user via a cosine similarity calculation target user acquainted with other users. We assume there is a user A, B, C, D and project a, b, c, d, e and preference items is as follows Figure III:

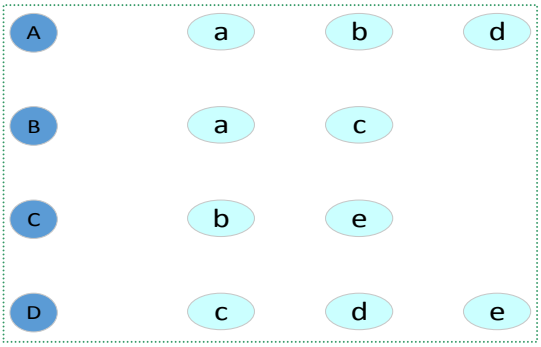

Figure 3: User preferences map

Then with the target user A is calculated in accordance with 3.2 of cosine similarity with other users B, similarity C, D is as follows:

$$\mathrm{sim}(A,B)=\frac{\sum_{i=1}^{n} R_{Ai}\cdot R_{Bi}}{\sqrt{\sum_{i=1}^{n} R_{Ai}^2}\sqrt{\sum_{i=1}^{n} R_{Bi}^2}}=\frac{1}{\sqrt{6}}$$

$$\mathrm{sim}(A,C)=\frac{\sum_{i=1}^{n} R_{Ai}\cdot R_{Ci}}{\sqrt{\sum_{i=1}^{n} R_{Ai}^2}\sqrt{\sum_{i=1}^{n} R_{Ci}^2}}=\frac{1}{\sqrt{6}}$$

$$\mathrm{sim}(A,B)=\frac{\sum_{i=1}^{n} R_{Ai}\cdot R_{Bi}}{\sqrt{\sum_{i=1}^{n} R_{Ai}^2}\sqrt{\sum_{i=1}^{n} R_{Bi}^2}}=\frac{1}{3}$$

The method compares the calculated short-answer similarity between users, but the time complexity of this method is O (|m| * |n|), which the user - a very time-consuming project is large matrix operations. In fact, many users and not to each other for the same article produced a behavior, i.e. many times the user U V and there is no intersection with the user, i.e., R (u) ∩R (v) = 0. The above algorithm will waste a lot of time on this similarity calculation between users. If you change an idea, we can first calculate R (u) ∩R (v)! = 0 the user (u, v).

B. K-Nearest Neighbors improvements

To do this, you can create items to a user's inverted list for each item is stored in the article produced a list of users behavior. So sparse matrix C [u] [v] = R (u) ∩N (v). So, assuming that the user u and v belong to the user a list of users

in the K inverted list items correspond, there is C [u] [v] = K.Thus, you can scan down the list of users for each row in the table items corresponding to the user list corresponding to the user pairwise C [u] [v] plus one, eventually you can get all the users is not between 0 C [u] [v] matrix.

As in the instance of Figure III, for the article a, the W [A] [B] and W [B] [A] plus 1, for the article b, the W [A] [C] and W [C] [A] plus one, and so, after all items have been scanned, we can get the final matrix W, where W is the cosine similarity of the molecular part, and then divided by the denominator W can get the end user interest similarity.

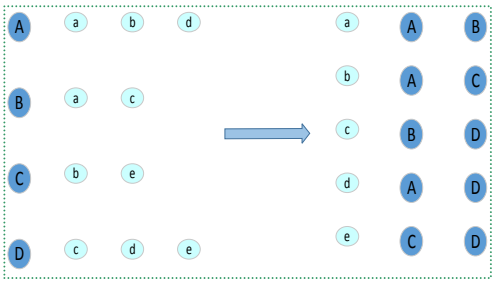

Figure 4: In the article centered inverted list

After the similarity between users get interested, recommendation algorithm will give users recommend the most similar interests and his K users favorite items. 5-4 above formula to measure the degree of user u i items of interest recommendation algorithm:

$$p(u,i) = \sum_{w \in S(u,K) \cap N(i)} w_{uv} r_{vi} \qquad (5\text{-}4)$$

Wherein, S (u, K) contains the closest interest and user u K users, N (i) Is the item i had a set of user behavior, w_{uv} similarity is interested user u and v of the user, r_{vi} behalf of the user v level of interest in items i, where we use a single act of implicit feedback data, so all $r_{vi} = 1$.

Learned from Figure III, user A can be calculated for items c, e degree of interest to determine whether c, e recommend to the A user. 5.2 According to the improved algorithm, the user A on items c, e of interest are:

$$p(A,c) = w_{AB} + w_{AD} = 0.7416$$
$$p(A,e) = w_{AC} + w_{AD} = 0.7416$$

VI. CONCLUSION

Personalized recommendation is currently in the field of e-commerce has been widely used, but with the advent of the era of big data the original recommendation algorithm has been a great challenge. This paper details the recommendation system, while K- neighbor algorithm is proposed for improvement. The improved algorithm to improve the recommendation accuracy while reducing the calculation K- nearest neighbor on the cost. However, with the development of the information age, the number of explosive growth of information, a large user of project information to make collaborative filtering technology is facing a great challenge, which will be the next focus of our research.

REFERENCES

[1] PeiYong Xia,Research on collaborative filtering algorithm in personalized recommendation technology. Ocean University of China. 2011.

[2] An A, Fang. Research commerce personalized recommendation service information.University of International Business and Economics.2006.

[3] Cukier K. Data, data everywhere:A special reort on managing information[N].Economist Newspaper,2010,(1).

[4] Goldberg P,Varian H R.Recommender systems[J].Communications of the ACM,1997,40(3):56-58.

[5] 许海玲，吴潇，李晓东，等.互联网推荐系统比较研究[J].软件学报，2009,20(2):350-362.

[6] Schafer J B,Konstan J,Riedl J.Recommender systems in e-commerce[C]//Proceedings of the 1st ACM Conference on Electronic Commerce.Denver,New York:ACM Press,1999:158-166.

[7] 林鸿飞,战学刚.文本特征区域与文本过滤的匹配机制.北京:计算机工程与应用,2000 年,36(7):7 一 9.

[8] Nicholas J. Belkin, W. Bruce Croft. Information Filtering and Information Retrieval:Two sides of the same coin. Communications of the ACM,1992,35(12):29-38

[9] Prahalad, C. K., Beyond CRM:CK Prahalad predicts customer context is the next big thing. American Management Association McWorld,2004.

[10] Herlocker, J. L., Konstan,J.A.,Content-independent task-focused recommendation. IEEE Internet Computing,pages 40-47,2001.

[11] Han J, Kamber M,Pei J.数据挖掘:概念与技术[M]。范明,孟小峰译.北京:机械工业出版社，2001:232-233.