An Efficient Collaborative Filtering using New User Similarity Measure for Recommendation

Priya Agrawal[1] Tejas Kadiya[2] Ramesh Prajapati[3]
[1]M.E. Student [2,3]Assistant Professor
[1]Department of Computer Engineering
[1,2,3]Indrashil Institute of Science & Technology-Rajpur

Abstract— Collaborative filtering has become one of the most used approaches to provide personalize services for users. The key of this approach is to find similar users or items using user-item rating matrix so that the system can show recommendations for user. However, most approaches related to this approach are based on similarity algorithms and this algorithm focus on only user item rating similarity calculation. These methods are not much effective, especially when the user rating data is extremely sparse and when only few ratings are available. To solve this problem research, propose an approach to compute the user similarity with the type of user-rating item. Research improved collaborative filtering algorithm based on user similarity combination, which combines the user similarity based on user-rating item and the user similarity based on the types of user-rating items. Research has also enhanced user-rating item similarity with new hybrid similarity. Experiments on classic MovieLens datasets are implemented. The result shows the superiority of the collaborative filtering approach in recommended performance.

Key words: Collaborative Filtering, Similarity, Distance, Near Neighbor, Prediction, Movie Types, Ratings, Hybrid Model

I. INTRODUCTION

The development of Internet technology and the wide spread of information make people easily access abundant information. As people face so much information, people cannot retrieve their desired information efficiently. This phenomenon is called "information overload", which has gradually become a big challenge in people's daily life and is attracting more and more scholars' attention. With popularization of Internet information service, the demands of personalized information service have gradually stood out in people's everyday life, but traditional information retrieve techniques fail to solve the critical challenge. Recommendation techniques merge, which use machine learning, data mining algorithms to discovery users' preference from large amounts of users' history data and present the most attractive and relevant information to user to reduce their " information overload" problem.

As the e-commerce websites become an inseparable tool in our daily life, the recommendation techniques turn into a very effective way to satisfy customers' personal needs, such as: Amazon, Taobao, Dangdang. Among them, Amazon makes the best use of recommendation techniques, and also achieves a great success [1].

The rest of this paper is organized as follows: Traditional similarity calculation method is discussed in section 2. Improved user similarity calculation methods are proposed in section 3. Experiment result is presented at section 4. Finally, we make a conclusion about the paper and present some directions for future works.

II. TRADITIONAL SIMILARITY TECHNIQUES

In this section, we first analyze the drawbacks of the existing similarity measures. Then, we introduce the motivation and hypothesis of the proposed similarity measure approach.

The Pearson correlation coefficient (PCC) and cosine (COS) similarity are the most widely used similarity measures in collaborative filtering. The formulas are defined as follows [2]:

$$\mathrm{sim(u,v)}^{\mathrm{PCC}} = \frac{\sum_{p\in I}(r_{u,p}-\bar{r}_u)(r_{v,p}-\bar{r}_v)}{\sqrt{\sum_{p\in I}(r_{u,p}-\bar{r}_u)^2}\cdot\sqrt{\sum_{p\in I}(r_{v,p}-\bar{r}_v)^2}}$$

$$sim(u,v)^{COS} = \frac{\vec{u}.\vec{v}}{\|\vec{u}\|.\|\vec{v}\|}$$

However, some shortages exist in both PCC and COS (described in follows). In order to overcome these drawbacks, many improved similarity measures have been introduced. Generally, the scale of ratings is absolute in recommender systems. The system can know which ratings are positive or negative. For considering the impact of positive and negative ratings, the constrained Pearson correlation coefficient (CPCC) [3] has been presented. The CPCC is defined as follows:

$$\mathrm{sim(u,v)}^{\mathrm{CPCC}} = \frac{\sum_{p\in I}(r_{u,p-r_{med}})(r_{v,p-r_{med}})}{\sqrt{\sum_{p\in I}(r_{u,p-r_{med}})^2}\cdot\sqrt{\sum_{p\in I}(r_{v,p-r_{med}})^2}}$$

Where r_{med} is the median value in the rating scale. For example, r_{med} is 3 in a scale from 1 to 5 and, it is 4 in a scale from 1 to 7.

Different people have different preferences of rating. Some people like to rate high, even they do not like the item very much. However, some people tend to rate low, even they like the items very much. The traditional cosine similarity does not account for the preference of the user's rating. For considering the preference of the user's rating, the adjusted cosine measure (ACOS) [4] has been introduced. The ACOS is defined as follows:

$$\mathrm{sim(u,v)}^{\mathrm{ACOS}} = \frac{\sum_{p\in P}(r_{u,p}-\bar{r}_u)(r_{v,p}-\bar{r}_v)}{\sqrt{\sum_{p\in P}(r_{u,p}-\bar{r}_u)^2}\cdot\sqrt{\sum_{p\in P}(r_{v,p}-\bar{r}_v)^2}}$$

Where P is the set of all items. If user u has not rated the item $p \in P$, the rating $r_{u,p}$ is zero.

Jaccard [5] and mean squared difference (MSD) [6] are another two widely used measures. Jaccard only considers the number of common ratings between two users. The basic idea is that users are more similar if they have more common ratings. The drawback is that it does not consider the absolute ratings. MSD only considers the absolute ratings, but not consider the number of common ratings. The drawback is that it ignores the credibility of the similarity.

$$sim(u,v)^{Jaccard} = \frac{|I_u| \cap |I_v|}{|I_u| \cup |I_v|}$$

$$\mathrm{sim(u,v)^{MSD}} = 1 - \frac{\sum_{p \in I}(r_{u,p} - r_{v,p})^2}{|I|}$$

Where r_u and r_v represents the set of items by user u and v rated respectively.

Although, many similarity measures have been proposed and they make up some drawbacks of the traditional similarity methods. These Similarity measures still have some drawbacks. In this section, we show the shortages of these methods.

The main drawbacks are described as follows:

1) Low similarity regardless of the similar ratings by two users.
2) High similarity regardless of the difference between the two user's ratings.
3) Ignoring the proportion of common ratings will lead low accuracy.
4) Discarding the absolute value of rating will become difficult to distinguish different users.

III. Improved Collaborative Filtering Technique

Research achieve higher similarity and accurate prediction result using following steps which is improved collaborative filtering technique

1) Step 1: Download MovieLens (User Movie Rating) Dataset and (Movie Type) Dataset from GroupLens website.
2) Step 2: After Preprocessing generate user/movie rating matrix and movie type matrix.
3) Step 3: Applying Proximity on two users' pair.

$$\mathrm{Proximity}(r_{u,p}, r_{v,p}) = 1 - \frac{1}{1 + \exp(-|r_{u,p} - r_{v,p}|)}$$

4) Step 4: Applying Significance on two users' pair.

$$\mathrm{Significance}(r_{u,p}, r_{v,p}) = \frac{1}{1 + \exp(-|r_{u,p} - r_{med}|.|r_{v,p} - r_{med}|)}$$

5) Step 5: Applying Singularity on two users' pair.

$$\mathrm{Singularity}(r_{u,p}, r_{v,p}) = 1 - \frac{1}{1 + \exp(-\left|\frac{r_{u,p}+r_{v,p}}{2} - \mu_p\right|)}$$

6) Step 6: Combine PSS.

i. $\mathrm{PSS}(r_{u,p}, r_{v,p}) = \mathrm{Proximity}(r_{u,p}, r_{v,p}) \times \mathrm{Significance}(r_{u,p}, r_{v,p}) \times \mathrm{Singularity}(r_{u,p}, r_{v,p})$

7) Step 7: Applying Modified Jaccard on two users' pair which is Jaccard'.

$$\mathrm{sim(u,v)^{Jaccard'}} = \frac{|I_u \cap I_v|}{|I_u| \times |I_v|}$$

8) Step 8: Combining Jaccard' and PSS which is JPSS.

$$\mathrm{sim(u,v)^{JPSS}} = \mathrm{sim(u,v)^{PSS}}.\mathrm{sim(u,v)^{Jaccard'}}$$

9) Step 9: Calculation of User Preference Base Similarity which is URP.

$$\mathrm{sim(u,v)^{URP}} = 1 - \frac{1}{1 + \exp(-|\mu_u - \mu_v|.|\sigma_u - \sigma_v|)}$$

1. Where,

$$\mu_u = \sum_{p \in I_u} \frac{r_{u,p}}{|I_u|} \quad , \quad \mu_v = \sum_{p \in I_v} \frac{r_{v,p}}{|I_v|}$$

$$\sigma_u = \sqrt{\sum_{p \in I_u} \frac{(r_{u,p} - \bar{r}_u)^2}{|I_u|}} \quad ,$$

$$\sigma_v = \sqrt{\sum_{p \in I_v} \frac{(r_{v,p} - \bar{r}_v)^2}{|I_v|}}$$

10) Step 10: Combining URP and JPSS which is NHSM.

$$\mathrm{sim(u,v)^{NHSM}} = \mathrm{sim(u,v)^{JPSS}}.\mathrm{sim(u,v)^{URP}}$$

11) Step 11: User rating type statistics table for two user u and v.
12) Step 12: Finding cosine similarity of two user u and v.

$$sim(u,v)^{cos} = \frac{\vec{u}.\vec{v}}{\|\vec{u}\|.\|\vec{v}\|}$$

13) Step 13: Combine cosine and nhsm similarity of user u and v.

$$sim = cos * nhsm$$

14) Step 14: Calculate List of target item for target user.
15) Step 15: Top neighbor algorithm on list of target item.
16) Step 16: Calculate Prediction for target user and target item.

$$p(a,i) = \bar{r}_a + \frac{\sum_{u=1}^{n} sim(a,u)(r_{u,i} - \bar{r}_u)}{\sum_{u=1}^{n} |sim(a,u)|}$$

17) Step 17: Recommending Item based on prediction score.

IV. Experiment

In order to validate the recommendation algorithms proposed by this study, this paper chooses the classic Movielens data sets to do experiments. Movielens dataset is provided by Grouplens Group for free, which can be obtained from Grouplens official website, and it also provides various datasets for collaborative filtering research. This dataset contains 943 users 100,000 ratings for 1682 movies, and the rating score value is between 1 and 5.

This study chooses mean average errors [1] (MAE) and (RMSE) as the measure indicator to evaluate the prediction accuracy. So the smaller MAE and RMSE, and the better accuracy prediction achieves. finally compare the different between predicting ratings and actual ratings.

The calculation of Mean Average Error is the average of the absolute differences between prediction rating $r_{p,i}$ and actual rating$r_{a,i}$. The equation is as follow:

$$MAE = \frac{1}{n}\sum |r_{p,i} - r_{a,i}|$$

Where n is the size of dataset

The calculation of Root Mean Square Error is the average of the absolute differences between prediction rating $r_{p,i}$ and actual rating$r_{a,i}$. The equation is as follow:

$$RMSE = \frac{1}{n}\sum |r_{p,i} - r_{a,i}|^2$$

Where n is the size of dataset

Research compares the MAE and RMSE among the traditional user-based collaborative filtering method improved user-based collaborative filtering method by similarity fusion with different near neighbors. The experiment results are shown in TABLE.

	N=5	N=7	N=10	N=15	N=25
PCC	0.80641	0.80539	0.80288	0.79941	0.79741
Hybrid Similarity	0.75978	0.72816	0.70367	0.68056	0.66982

Table 1: MAE in different number of nearest neighbors and techniques

	N=5	N=7	N=10	N=15	N=25
PCC	1.01232	1.02054	1.01471	1.00617	1.00156
Hybrid Similarity	0.98236	0.90477	0.84391	0.78725	0.76036

Table 2: RMSE in different number of nearest neighbors and techniques

The following figure gives fairly brief comparisons.

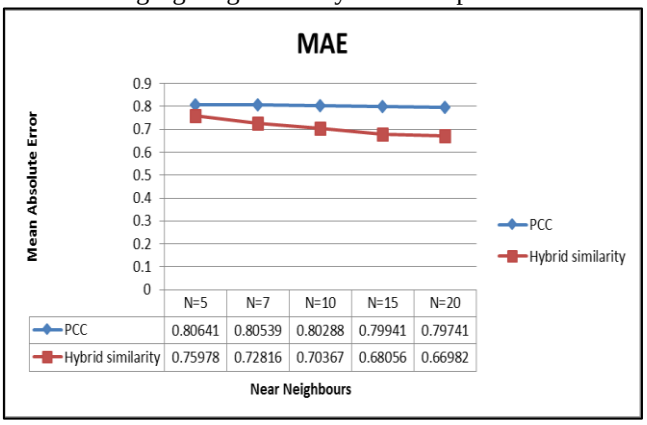

Fig. 1: near neighbors Vs MAE

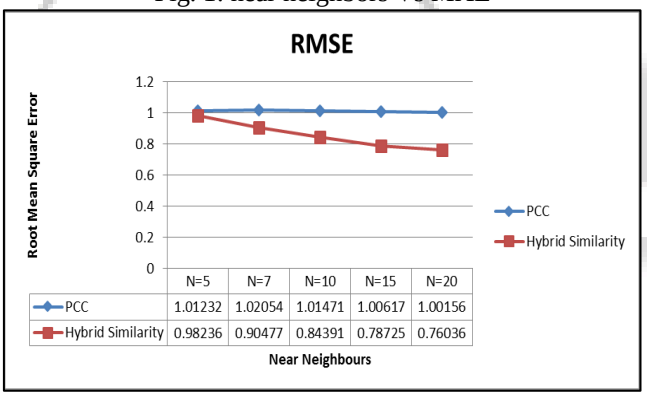

Fig. 2: near neighbors Vs RMSE

The experimental result shows following conclusion:

Improved user-based collaborative filtering approach by similarity fusion is superior to the traditional user-based collaborative filtering algorithms with the same number of nearest neighbors, that is, the hybrid similarity calculation method is more accurate than the Pearson correlation and user-rating item similarity calculation method.

V. Conclusion and Future Work

Research points out the inaccurate similarity calculation methods in traditional collaborative filtering algorithm. Proposes a new user-rating similarity which takes the proportion of the common rating between two users into account. Considering different users have different rating preferences. Also uses the mean and variance of the rating to describe the rating preferences of user. Research puts forwards the user-rating item type similarity method. Experiment shows this method obviously outperforms the traditional method at accuracy. As Number of item types are far less than that of items, so it can easily avoid rating data sparsity problems of similarity calculation. In order to yield even better recommendation result, this work comes up with improved user-based collaborative filtering algorithm by similarity fusion.

Furthermore, goods have more attributions than the type in e-commerce or in the film, so user-rating information also includes their preferences on other attributions, in e-commerce it can be brands, producing companies, and colors etc. and in the film, it may be the directors, actor, and studio etc., so using that information to improve the recommendation quality is advisable.

References

[1] K. Zhao and P. y. Lu, "Improved collaborative filtering approach based on user similarity combination," *2014 International Conference on Management Science & Engineering 21th Annual Conference Proceedings*, Helsinki, 2014, pp. 238-243.

[2] Haifeng Liu, Zheng Hu, Ahmad Mian, Hui Tian, Xuzhen Zhu, A new user similarity model to improve the accuracy of collaborative filtering, Knowledge-Based Systems, Volume 56, January 2014, Pages 156-166, ISSN 0950-7051

[3] U. Shardanand, P. Maes, Social information filtering: algorithms for automating word of mouth, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1994, pp. 210–217.

[4] H.J. Ahn, A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem, Inform. Sci. 178 (1) (2008) 37–51.

[5] G. Koutrica, B. Bercovitz, H. Garcia, FlexRecs: expressing and combining flexible recommendations, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, 2009, pp. 745–758.

[6] F. Cacheda, V. Carneiro, D. Fernández, V. Formoso, Comparison of collaborative filtering algorithms: limitations of current techniques and proposals for scalable, high-performance recommender system, ACM Trans. Web 5 (1) (2011) 1–33.

[7] Priya Agrawal, Tejas Kadiya, Ramesh Prajapati, "A Comparative Survey of Collaborative Filtering Similarity Measures: Limitations of Current Similarity and Formalization of New Similarity Measure" , IJIRT-144111, International Journal of Innovative Research in Technology, NOV-2016, Volume 3, Issue 6, 94-97

