An improvement of video recommender similarity measurement model

Deng Zhenrong[1,2], Zhang Xi[1], Deng Xing[1], Xu Liang[1], Huang Wenming[1,2]

[1]School of Computer Science and Engineering, Guilin University of Electronic Technology, Guangxi Guilin, 541004, P.R China

[2]Guangxi Key Laboratory of Trusted Software, Guilin, 541004, P.R. China

Keywords: Recommender System, Collaborative Filtering, Item Attribute, Similarity

Abstract. Collaborative recommender systems have succeeded in capturing the similarity between users and items based on ratings. However, they have rarely considered about the available information of the multimedia such as categories, delivery time and so on. Such information are valuable and feasible to solve rating bias problems in recommender systems. We found that user described their preferences directly to the item rating data is not comprehensive. In this paper, we design IBHF (Item-attribute Based Hybrid Filtering) based on movie features of the multimedia information. In the IBHF, we provide recommendation service by new Pearson method which is a similarity measure technique used to integrate movie attributes into the Item-based collaborative filtering (IBCF) framework in hopes of achieving better performance. The experiment prove that this method can make the items recommendation more ideal, and also provides a solution to solve the cold start problem to different recommendation items.

Introduction

Recently, the Internet information filtering has been changed from searching to active recommendation. Some of the information cannot be measured by the keyword, users rely on this personalized recommendations gradually. And video recommendation become a popular application field. However, the inconsistencies between user preferences description and multimedia types of items described is difficult to tell user preferences from user input. In 1992, Goldberg[1], who first proposed the "collaborative filtering" that Recommendation system is based on user involved and user history behavior data on website. And the behavior had the similarity between the users and then make recommendations based on the similarity.

The key step of collaborative filtering algorithm is find the nearest neighbor group according to the rating data from users. Neighbor group is the recommender system select users who have similar preferences of the target user[2]. In the video recommend, the common method is that different users make the evaluation method of their attention to video items, which recommendation algorithm based on different user-item overall scoring data, judging the similarity between user preferences, and ultimately giving a certain user recommendation with preferences similar users like. However, the user-item rating recommend only from the integrity of the numerical and measure of user preferences have scored is not comprehensive. That is, on the other hand, user ratings prejudice - the value measure of user preferences only from user-item scored. Numeric feedback rating system for user preference information shows that users are vulnerable to interference in many aspects: such as being affected easily by the public rating data [3], easy to influence the current popular project give score [4], the user has the item knowledge limitations [5]. So historical rating data will impact on the accuracy of recommendation in the future. Different users like an item must not be identical. The film, for instance, some users like to see a comedy movie, some usersfavor actors, some like the director, or simply it is part of the newly released popular movies etc. But the user on the film's overall score cannot reflect the information directly, so the recommendation system recommends rely on user-item rating data directly cannot recommend the best result. For example, when two users have the same history score but the item

attributes are different, items recommendation method cannot be distinguished based on the rating data. As shown in Table 1, film 3 and film 4 has the same history scoring record, so according to the item-based collaborative filtering algorithm (IBCF), which have the same opportunity to give the preference score of user ID=1. However, if the recommendation know the movie 3 belongs to the category of comedy as Movie 1 and 2, recommendation system should capture this information and recommend the movie 3 to 1 users preferred.

Table 1 User-item rating

user / movie	A	B	C	D	E
1	5	4	3		
2	4	4	3		
3		3	2		
4		3	2		
5				4	

Attribute / movie	Comedy	horror
1	98%	2%
2	90%	10%
3	98%	2%
4	4%	96%
5	98%	2%

In this paper, based on the public data set rating history, we analysis the similarityof item attributes. And modify the similarity measurement to increase the recommender accuracy.

Similarity measurement and analysis

Collaborative filtering recommend a list to the target user recommendation rely on other users' view. It is that if the user rating of some items is quite similar, they score on other items are relatively similar [6]. Many collaborative filtering algorithms based on users or items are focusing on calculating the neighbor. The accuracy of neighbors' similarity will influence the final recommendation algorithm result. Kong [7] and others pointed out the irrationality of the traditional algorithm to produce the neighbor and introduced the association rules, they modified Pearson correlation coefficient to improve the accuracy of similarity. Hu[8] proposed to measure the time data and item data based on two dimensions to calculate similarity when the user's interest changed. Hu [9] proposed a collaborative filtering algorithm based on item attribute value matrix, which is the attribute score of the items by multiplying the value of the similarity between users. Through the user-item after attribute preference matrix to calculate the similarity between items, which are the target user item rating prediction.JiXiaosheng [10] through a combination of the collaborative filtering algorithms based on the user and the items.In the prediction of the target user to commodity score, According to the forecast of the nature of items to the finding of neighbor users. They introduced the distance of users and items related to the nearest neighbor users as the weight calculating target user prediction score for commodities. This method can improve the recommendation accuracy to a certain extent. Chang]joined the time attribute in the user behavior records, Higher weights was given to user current interest keywords, setting up the user interest keywords update tree. Those methods solved the problem of updating, and the user model users interested in short-term and long-term interests to a certain extent. Because of Pearson does not fit to apply to make the text similarity analysis. In this paper, this method is used to analyze the correlation similarity between users and the film items. Literature using the Pearson correlation coefficient measurement, it points out that the selection of collaborative filtering in Pearson metric method is helpful to improve recommendation accuracy.

The above methods mentioned to measure the similarity of users or items from different dimensions. When the average value is too high or too low in user rating dispersion, we can also get more appropriate recommendation results.In this paper, according to the literature [12] Pearson method, the improvement was made on the weight distribution of similarity metric model.

Pearson similarity measurement and its improvement

Pearson similarity measurement. Pearson similarity is a value between -1 and 1 which is used to describe the change trend of two sets of linear data.

$$\mathrm{sim}(i,j)=\frac{\sum_{c\in I(u_i,u_j)}(r_{u_{i*c}}-\overline{r_{u_i}})(r_{u_{j*c}}-\overline{r_{u_j}})}{\sqrt{\sum_{c\in I(u_i,u_j)}(r_{u_{i*c}}-\overline{r_{u_i}})^2}\sqrt{\sum_{c\in I(u_i,u_j)}(r_{u_{j*c}}-\overline{r_{u_j}})^2}} \quad (1)$$

In the above, $r_{u_{i*c}}$, $r_{u_{j*c}}$ describes respectively the score of the user on item C. $\overline{r_{u_i}}$, $\overline{r_{u_j}}$ describes respectively the average on the user who has been scoring items, $I(u_i,u_j)$ is u_i,u_j scoring together。Two objects are more similar, the value of sim(i,j) is small, two items are more similar.

Improved measurement model. After the analysis of experimental film data，we let sim(i, j)as the calculation value of the similarity of original items i and j. This value is obtained by the Pearson formula. We add the audiences have main concern (film type and the time of release) into this formula. The calculation of sim(i,j) is modified as:

$$\mathrm{sim}(i,j)=\mathrm{sim}(i,j)*e^{(-\alpha)*(c_i-c_j)}+\mathrm{sim}(i,j)*e^{(-\beta)*(t_i-t_j)}+\ldots+\mathrm{sim}(i,j)*e^{(-k)*(p_k-p_{k+1})} \quad (2)$$

In the above, the c_i,c_j are the film category, t_i,t_j are the film publishing time, p_k， p_{k+1} are the property to other dimensions of a movie. The bigger of two film types and publishing time difference are, the smaller of their similarity, even the given movie items are the same score data by users. By adjusting the parameters α，β, we can decide what factors affect the similarity in the weighted formula. In the fourth chapter, the experiment will verify the rating prediction precision improvement situation by adjusting the appropriate αand β.

Hybrid filtering recommendation algorithm based on item attributes

According to Pearson similarity algorithm, we expand the parameters of the item attribute in its formula and have improved similarity measurement algorithm to solve the measurement of user rating bias, called Item-attribute Based Hybrid Filtering(IBHF). It selects attributes concentrated by usersfrom actual movie item, and solve the problems mentioned in the introduction. The film attributes presented in this paper are starting from the reality of people's attention preference, that when choosing between different user made in the items mainly evaluate several films' attribute. Besides, the users rating after watching a movie was mainly reflects several properties of the film that users have personalized preference, instead of all film attribute. Combined with the characteristics of traditional similarity measure methods, the algorithm proposed in this paper put forward in the basic premise of the following:

(1) In the selection of new movies, the user preference of movie attributes still have consistent with the former attribute preference;

(2) Users' explicit data ratings, can be used to the research on several important attributes, and we believe the attributes affect preference.

The following introduces the key steps of items-attributes recommended algorithm: find the nearest neighbor and generate a recommendation list.

Getting the nearest Neighbors. In the recommended process similarity on item attributes calculation, we found that the calculation of similarity exists some problems. For example, calculation of the user u on item I score, where the similar items I for is I1, I2, I3. Similarity between the three similar items is strong at initial, so in the weighted calculation of recommended process, the similarity information will be repeated calculation. And when the user rating received multiple interference, the score prediction will not be accurate. How to increase or decrease the similarity value in the nearest neighbor computation is the key step of our algorithm.

Based the formula2.2, IBHF can control the similarity of neighbors when they involved into repeated calculation by adjusting the parameters like αand β. For instance, I1,I2, I3 have strong similar at initial, so we set the value of α to decrease the weight of similarity calculation.

Recommendation. The essence in recommendation is display the score data which unknowns in the prediction. The above method to calculate similarity between items is the first step. For each item, according to the similarity with the item i, the most similar k data set collection, named N(I, k). The prediction of user u to items i, is the method:

$$r_{ui} = \frac{\sum_{j \epsilon N(i,k)} sim(i,j) * r_{u_j}}{\sum_{j \epsilon N(i,k)} sim(i,j)} (3)$$

This method is the basic method of collaborative filtering, the similarity of sim(i,j) based on formula 2.1, and then get results from 2

Experimental results and analysis

Datasets analysis. The experiments data is from the datasets of MovieLens. The selected datasets, including 943 users, 1682 films and 100000 scoring records.

In the data file of MovieLens, the data format of data sets of u.item is {movie_id | movie_title | release_date | IMDB URL | unknown | Action | Adventure | Animation | Children 's | Comedy | Crime | Drama | Horror |...}. Among those attributes, the vast majority of the data given by the film type and distribution of time, which is available for this experiment combined score matrix.

Recommendation accuracy. Currently personalized recommendation research field and paper use the root mean square error (RMSE) as evaluation standard. The experiments of this chapter also use RMSE to make the evaluation and the calculation method of RMSE is as follows. $r_{u_i}$is the actual ratings of user u to item i, $\overline{r_{u_1}}$ is the score caculated by recommendation algorithm. The RMSE defined as:

$$RMAE = \sqrt{\frac{\sum_{u,i \epsilon T} (r_{u_i} - \overline{r_{u_1}})^2}{|T|}} (4)$$

This calculates the mean square of error square of the evaluation score and the actual score, so the smaller the value is, the more accurate the prediction results are.

The contrast of two algorithms

Item-based Collaborative filtering based on Pearson method

Item-based Collaborative filtering based on Pearson method is that the neighbor similarity measurement is Pearson method. The result as follows:

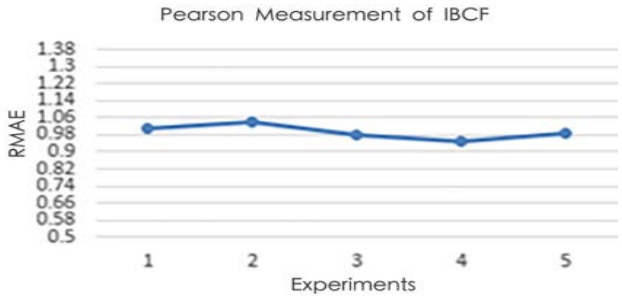

Fig1. RMAE of Item-based Collaborative filtering based on Pearson method

After five experiments, the average of RMAE is 0.9914.

Item-attribute-Based Hybrid Filtering Recommendation Result. In the improved similarity measure method, the value of α is a film type property to calculate the influence degree to the size of the item similarity score, which will affect the prediction accuracy.

In the experiments, in view of the similarity calculation of two weights αand β. And conditions: $\alpha + \beta > 0.5$, which according to the actual test data and the least squares fitting method of linear planning, continuously adjust the value. In order to get the best results. The $\{\alpha=0.1,\beta=0.5;\alpha=0.15,\beta=0.4;\ \alpha=0.2,\beta=0.4;\ \alpha=0.25,\beta=0.35;\alpha=0.28,\beta=0.35\}$ 5 different group. After the experimental observation, the IBHF algorithm recommendation shows an striking improvement.

Table 3. The RMAE based on the film type and publishing time

	$\alpha = 0.1$, $\beta = 0.5$	$\alpha = 0.15$, $\beta = 0.4$	$\alpha = 0.2$, $\beta = 0.4$	$\alpha = 0.25$, $\beta = 0.35$	$= 0.28$, $\beta = 0.35$
u1	0.9006	0.8878	0.9707	0.953	0.9938
u2	0.9469	0.9449	0.8901	0.8743	0.9712
u3	0.9199	0.9223	0.8432	0.9057	0.9094
u4	0.9264	0.8909	0.9404	0.9353	0.9445
u5	0.8828	0.8947	0.8901	0.8721	0.9332
Average	0.91532	0.90812	**0.9069**	0.90808	0.95042

The above 5 experiments, α=0.2,β=0.4 in IBHF method compared with IBCF algorithms: the recommendation accuracy RMSE value change trend shown in fig 2, the RMAE value changed from 0.9914 in 4.3.1 to 0. 9069 (as table 3).

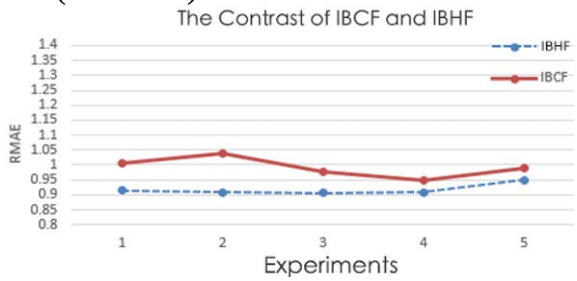

Fig 2. The contrast of IBCF and IBHF

The experiments fully proved that the traditional collaborative filtering based on the item rating, which considering the film itself attribute information to calculate the similarity, will enhance the personalized recommendation scores in rating forecast accuracy.

Conclusion

This paper comes up with the Item-attribute Based Hybrid Filter (IBHF), which based on the traditional collaborative filtering algorithm, and analysis of the project on collaborative filtering similarity calculation and its scalability contribution. The author proposed algorithm on real data sets,compared its accuracy with the results based on collaborative filtering recommendation algorithm items.Experimental results show that IBHF algorithm not only slightly better than the classic IBCF algorithm, but also provides a guideline to solve the cold start problem in the future research.That is, when the historical data is too small, enlarging the impact factor to make the property as the main basis for calculating neighbors similarity. This paper has reason to believe that fully exploring original information of the recommended items is a key mean to advance personalized techniques.

Acknowledgments

This work is supported by Guangxi key Laboratory of Trusted Software (No: kx201317), by the Postgraduate's Innovation Project of Guilin University of Electronic Technology under (No: GDYCSZ201470), by the 2014 Guangxi University of Science and Technology Research Projects (NO: LX2014149) ,by the Nature Science Foundation of Guangxi (No: 2013GXNSFAA019350).

References

[1] Goldberg D, Nichols D, Oki B M, et al. Usingcollaborative filtering to weave an informatio-n t apestry[J]. Communications of the ACM, 1992, 35(12): 61-70.
[2] Wen You, shui-sheng Ye. Collaborative filtering recommendation in e-commerce recommendder systems [J]. Computer technology and development，2006, 16(9): 70-72.
[3] Hui-guirong, shengxuHuo, etc. The collaborative filtering recommendation algorithm based on user similarity [J]. Journal of communications, 2014, 35(2) : 16-24.
[4] Ang Lee, jiangZhong. Combined with the collaborative filtering model research of potential

properties [D]. Chongqing university, a master's degree in theory, 2013, 4
[5] Chun-xiaoxing, GaoFengrong etc. The Adaption changes in user interest in the collabor-ative filtering recommendation algorithm [J]. Journal of computer research and development, 2007, 44 (2) : 296-301.
[6] Deng Ailin yang-yongzhu, bole. The collabor-ative filtering recommendation algorithm based on items score predicts [J]. Journal of software, 2003, 14 (9).
[7] Kong Weiliang. The research of key problems in Collaborative filtering recommendation system [D]. Central China normal university, 2013.
[8] Jun Hu,BingWang,Yu. Personalized Tag Recommendation Using Social Influence[J]. Journal of Computer Science & Technology 2012-05(20):218-223
[9] Hu Xinmin, jin-long zhang, etc. The electronic commerce recommendation system based on commodity attribute research [D]. Huazhonguniversity of science and technology, Ph.D. Thesis, 2012, 4.
[10] JiXiaoSheng, Liu Yan soldier, lai-mingluo. In collaborative filtering based on user interest degree of similarity measure method [J]. Journal of computer applications, 2010, 30 (10) : 2618-2620.