IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 4 Issue 4, April 2017
ISSN (Online) 2348 – 7968 | Impact Factor (2016) – 5.264
www.ijiset.com

Incorporating tag information to Enhance the Collaborative Filtering Recommendation Algorithm

YuHong Zhou*

*Chongqing Key Lab of Mobile Communications Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, P. R. China

Abstract

In traditional collaborative filtering recommendation algorithm, the similarity is calculated based on the common ratings and the accuracy is not well when the data is sparse. The thesis proposes a novel collaborative filtering by incorporating tags(CF-IT). Firstly, it defines the rating similarity by items' historical rated behaviors using graph theory, and secondly, it defines the tag similarity by items' tag information. Then, we combine the rating similarity and tag similarity to obtain comprehensive item similarity model. Finally, we conduct the experiment to verify the validity and the accuracy of the algorithm.

Keywords: *Recommendation System, Collaborative Filtering, multi similarity, tag information*

1. Introduction

With the explosive growth of the personalized recommendation technologies, the users can get higher quality service and better user experience, the items on the website can be displayed in front of the users who are really interested in them. Recommender systems enhances the economic efficiency of the website and the retention rate of the users on the platform[1]. For some time, the collaborative filtering technology is widely used in the industry because of it's low dependence, good applicability and simple encoding. But it also faces the great challenges such as data sparsity problem. There are at least millions of users and items in the large business recommender systems, but the user only rate some of these items and caused data sparsity problem. Moreover, personalized recommendation also faces cold start problem, extendibility problem and so on.

For the sake of those problems, we first analyze the availability of the tag data, and then incorporate the similarity calculation method based on the rating data. Finally, we propose a new similarity calculation method to search the nearest neighbor set of the target item. On the one hand, it first uses the user-item rating data to calculate the rating distance between the different items, and uses the rating distance marked by all users to calculate the rating distance vector, every column of this vector can be considered as a discrete random variable, and then according to the probability density function to calculate the information entropy of the rating distance. The smaller the information entropy is, the more similar the items are. Finally, it uses sigmoid function to standardize the information entropy of the rating distance. On the other hand, it calculates the tag similarity between the items according to the tag data. Firstly, according to the item-tag matrix, it constructs the item-tag bipartite graphs model and gives the initial weight to those tags which have marked the active item in the bipartite graphs model, the final weight can be obtained by using the resource allocation methods to make the tag weight spread in the bipartite graphs model. It calculates the item tag similarity by using the cosine value of the angle between the tag vectors. Secondly, we combine the item's rating similarity and tag similarity to obtain the comprehensive similarity computation method, Finally, we use the comprehensive similarity computation method to search the item's nearest neighbor set. It predicts the target user's preference about the item and generate more accurate recommendation according to the history behavior record. The experimental result shows that the proposed algorithm utilizes the additional item's information efficiently, enriches data source and improves the precision of the similarity computation, meets the personalized recommendation demands. In order to solve the data sparsity problem, and incorporate the tag information to the traditional recommendation algorithm problem, the industry and the academia makes deep and broad study in the recommendation area, and some solutions are proposed. The method that use the initial prediction rating fills in the rating matrix can solve data sparsity problem effectively and makes the rating matrix much denser. Work in [2] proposes to incorporate the item based and user based model to predicts the missing data in the rating matrix.

Work in [3] proposes the collaborative filtering recommendation algorithm based on cloud model LICM. First, it maps the similarity computation to the knowledge level based on interchange model from qualitative concept to the qualitative knowledge description, and then uses the rating frequency vector to calculate the rating feature vector. Finally, according to the angle cosine formula to compute the similarity and it overcomes the shortcoming of the similarity computation. The PIP similarity measure is proposed to solve the incorrect similarity computation problem caused by data sparsity and the cold start problem in current collaborative filtering recommendation algorithm[4]. This heuristic similarity measure is composed of similarity, proximity; impact and popularity. The measure performs better in accuracy than the traditional method, but the measure only considers the user's local rating information and ignores the user's whole preference. On this basis work in [5] proposes a new user similarity model to improve the accuracy of collaborative filtering. Work in [6] considers the similarity between the users that can spread just like the trust in social network, it constructs new user relationships to alleviate the data sparsity by spreading the similarity in bipartite graphs. Work in [7] proposes social graph method to introduce the user-tag, item-tag and user-user relationships to enhance the quality of the recommender systems. Chen et al. calculated the item similarity and user interest similarity using tag information, and then construct the user's trust network to generate the top-n recommendation[8]. Work in [9] considers that the tag reflects the semantic of the item rather the user's preference, it recommends according to the different role of the tags to build the user-centric tripartite graph and the item-centric tripartite graph.

2. Comprehensive similarity

2.1 problem description

Select the nearest neighbors for the target item is the key in collaborative filtering, it is determined by the similarity of the items. The collaborative filtering recommendation algorithm by incorporate tags is proposed and a new similarity calculation method is given to search the nearest neighbor set of the target item.

2.2 comprehensive item similarity

The comprehensive item similarity model is proposed by incorporating the item tag information and user item rating information. The comprehensive item similarity between the item i and item j $w(i,j)$ is computed as follows:

$$w(i,j) = \partial * w(i,j)_{rating} + (1-\partial) * w(i,j)_{tags} \quad (1)$$

Where $w(i,j)_{rating}$ and $w(i,j)_{tags}$ is the rating similarity and the tag similarity of the item i and item j , ∂ is the weight factor. The pseudocode of the collaborative filtering recommendation algorithm by incorporate tags CF-IT as follows:

CF-IT algorithm pseudocode

Algorithm: Collaborative Filtering recommendation algorithm by integrate tag information

Input: Item-tag mark matrix M_{tag} ,User-Item rating matrix M_{rating} ,the number of neighbor K, the length of recommended list N

Output: Top-N recommend list

1. Constructed bipartite graph mode $G(U,I,E,w)$ by user-item rating matrix M_{rating}
1. **for** $i, j \in G$ and $i \neq j$ **do**
2. Calculate $W_{rating}(i,j)$
3. **end for**
4. Constructed bipartite graph mode $G'(I,T,E,w)$ by item-tag matrix M_{tag}
5. **for** $i, j \in G'$ and $i \neq j$ **do**
6. Calculate $W_{tag}(i,j)$
7. **end for**
8. Generate $W(i,j)$ by Linear fusion $W_{rating}(i,j)$ and $W_{tag}(i,j)$
9. Search the K nearest neighbors and predicted the target user´s preference $p(u,i)$ for unrated item by $W(i,j)$
10. Generate Top-N recommend list
11. **return**

IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 4 Issue 4, April 2017
ISSN (Online) 2348 – 7968 | Impact Factor (2016) – 5.264
www.ijiset.com

2.3 rating similarity

The weight bipartite graphs modeling the user-item rating behavior effectively[10], the weight bipartite graphs $G(U,I,E,W)$ as the Fig.1 shows, where U is the set of users, I is the set of items, E is the set of edges which connect the item and user. $W: E \to Z^+$ is the mapping from E to the set of positive integers. There is an edge $e \in E$ connects the node u and node i if the user u have rated the item i, and $W(e)$ is the rating value. As shown in Fig.1, $\{X,Y,Z\}$ is the set of the users, $\{a,b,c,d\}$ is the set of items, where X rated the item $\{a,c\}$, Y rated item $\{a,b,d\}$, Z rated item $\{b,c,d\}$.

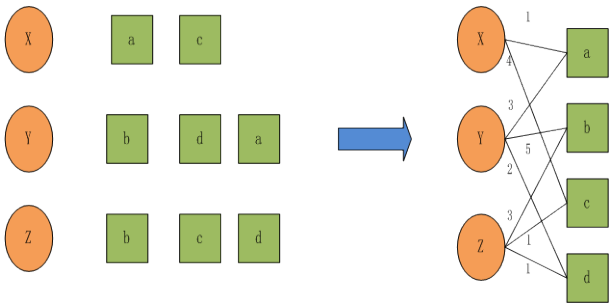

Fig.1: user-item bipartite graphs model

In the bipartite graphs model, $d^u_{i,j}$ is the rating distance between the item i and the item j marked by user u. It calculated as follows:

$$d^u_{i,j} = \begin{cases} r_{\min}, \text{if } r_{ui} = r_{uj} \\ r_{\max} - \min(r_{ui}, r_{uj}) + r_{\min}, \text{else} \end{cases} \quad (2)$$

The rating distance vector V^d_{ij} as follows:

$$V^d_{ij} = (d^{u_1}_{i,j}, d^{u_2}_{i,j}, ..., d^{u_m}_{i,j}) \quad (3)$$

According to the rating distance vector we can define the set of the rating distance D_{ij} and the rating distance frequency vector V^f_{ij}, as follows:

$$D_{ij} = \{d_1^{ij}, d_2^{ij}, ..., d_m^{ij}\} \quad (4)$$

$$V^f_{ij} = (N_{d_1}, N_{d_2}, ..., N_{d_m}) \quad (5)$$

Where the kth item N_{d_k} is the occurrence of the rating distance d_k between the item i and item j with the secondary users. The information entropy theory can measure the uncertainty of the rating distance quantitatively. The higher the uncertainty is, the less the similarity is. The information entropy of the rating distance between the item i and the item j $E(i,j)$ as follows:

$$E(i,j) = -\sum_{d_i \in D_{ij}} p(d_i) \cdot \log p(d_i) \quad (6)$$

Where $p(d_i) = \frac{N_{d_i}}{N_d}$ ($N_d = \sum_{i=1}^{m} N_{d_i}$) is the probability of the rating distance d_i. The impact factors of the rating distance as follows:

$$\gamma_{ij} = \frac{1}{1 + d^{med}_{ij}} \quad (7)$$

Where d^{med}_{ij} is the median of the rating distance between item i and item j. In conclusion, the calculation formulas of the rating similarity between the item i and item j as follows:

$$W(i,j)_{rating} = \gamma_{ij} \times (1 - \frac{1}{1 + e^{-E(i,j)}}) \times \frac{|N_i \cap N_j|}{|N_i| \times |N_j|} \quad (8)$$

2.4 tag similarity

Tag is seen as the resources which have some initial concentrations, the processes of the resource allocate as the Fig.2 shows,

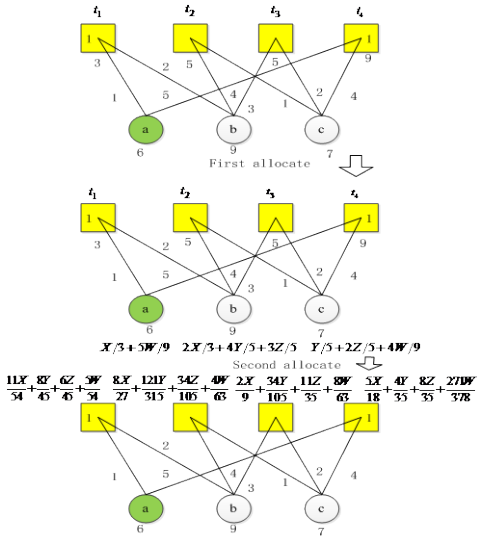

Fig.2: the process of the resource allocates

where item a is the target item. We give the initial resource of the tag which have marked the item a, the value that under the tag node is the total numbers of the marked tags, the value that under the item node is the mark times of the item which is marked by all the tags. It assume that X,Y,Z,W is the initial weight of the tag t_1, t_2, t_3, t_4. We can obtain the final weight of the tags by spreading the resources as the Fig.2 shows.

For example, the final weight of the tag is after the resource allocated. The thesis defines the tag feature vector according to the final weight of the tags as follows:

$$\vec{P}_{it} = (\frac{n_{it_1}}{n_i} \times w(t_1), \frac{n_{it_2}}{n_i} \times w(t_2), ..., \frac{n_{it_k}}{n_i} \times w(t_k), ..., \frac{n_{it_n}}{n_i} \times w(t_n)) \quad (9)$$

Where n_i is the marking times of the item i, n_{it_j} is marking times marked by the tag t_j of the item i. $w(t_j)$ is the weight of the tag t_j after resources allocated. $\frac{n_{it_j}}{n_i}$ is frequency of the item i marked by tag t_j.

The thesis defines the tag similarity of the target item i and item j by calculating the cosine angle of the tag feature vector. The calculation formula as follows:

$$w(i,j)_{tag} = \cos ine(\vec{P}_{it}, \vec{P}_{jt}) = \frac{\vec{P}_{it} \cdot \vec{P}_{jt}}{\|\vec{P}_{it}\| \cdot \|\vec{P}_{jt}\|} \quad (10)$$

3. Experiment

3.1 experimental design

We conducted experiments using datasets from MovieLens100k and Last.fm, The original MovieLens 100k dataset contains 943 users, 1682 movies. The Last.fm dataset contains 1892 users, 11946 tags, 17632 items. In this section, we design two set of experiments to evaluate performance of the CF-IT algorithm.

1. Using the MovieLens 100k dataset, compare CF-IT algorithm with item-based collaborative filtering algorithm to see which one has a higher comprehensive measurement F1.

2. Using the Last.fm dataset, observe the performance of the CF-IT algorithm to determine if the addition of tag information into a collaborative filtering algorithm can improve the quality of the algorithm. If so, go ahead and pursue the values of the parameters K, N, αwhich may enable the best performance of the algorithm.

3.1 experiment result

experiment 1： In this experiment, compare the algorithm CF-IT with one of the traditional collaborative filtering algorithms ItemCF [11] in terms of the comprehensive evaluation metric F1 . The comparison result is shown in Figure 6 with the number of neighbors ranging from 5 to 100. Given the fact that a larger F1 indicates a superior algorithm, it can be seen clearly that the proposed algorithm outperforms ItemCF, and would be a preferred choice when a special requirement such as finding an equilibrium between Precision and Recall is needed.

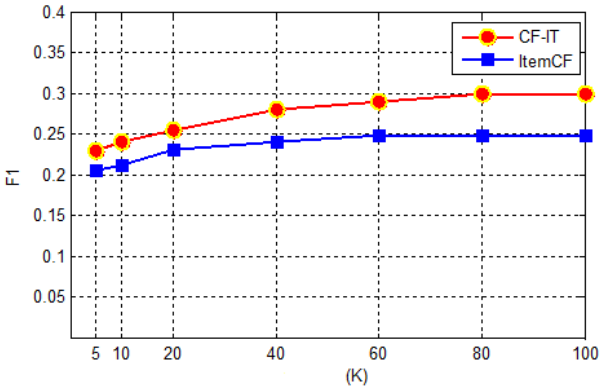

Fig.4: Comparison between CF-IT and ItemCF with respect to F_PR

In order to see the effectiveness (or non-effectiveness) of different similarities on the evaluation metrics Precision , Recall , F1 , and Coverage , we in this experiment conduct three sub-experiments: (1) CF-IT@1: implement the recommendation solely by the attraction similarity, (2) CF-IT@2: implement the recommendation solely by the interaction similarity, and (3) CF-IT@3: implement the recommendation by the combination of the attraction similarity and the interaction similarity.

Figures 5 – 8 show the comparisons of CF-IT@1, CF-IT@2, and CF-IT@3 in terms of Precision , Recall , Coverage.

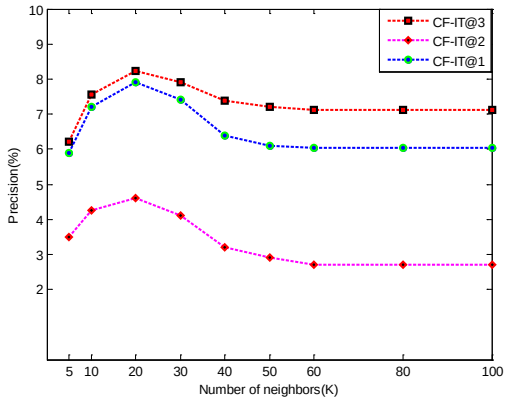

Fig.5: Correlation between K and Precision

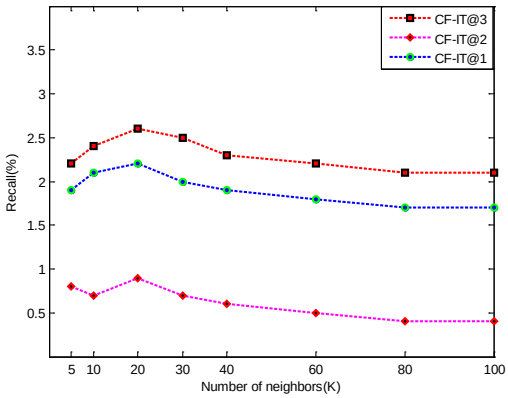

Fig.6: Correlation between K and Recall

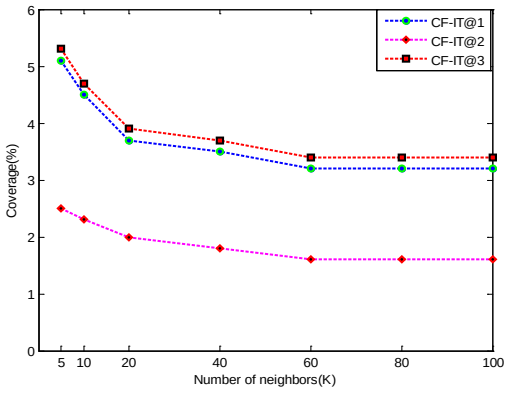

Fig.7: Correlation between K and Coverage

All experiments shown in Figures 5 – 8 are conducted for some given and fixed sparsity. Figure 5 illustrates the correlation between K and Precision ; see that CF-IT@3 outperforms CF-IT@1 slightly, but beats CF-IT@2 to a large extent. Figure 6 shows the correlation between K and Recall .

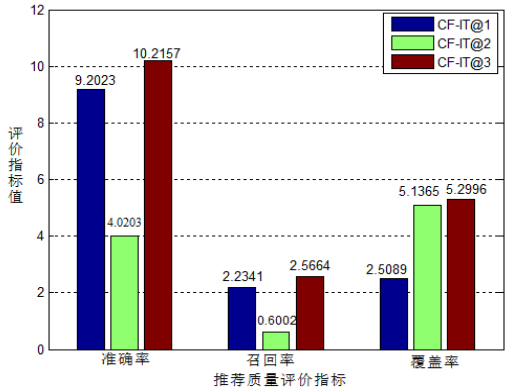

Fig.8: Comparison with respect to evaluating indicator

Again, we are able to observe that CF-IT @3 is superior to both CF-IT@1 and CF-IT@2. Figure 7 exhibits the correlation between K and Coverage,and clearly indicates that SuMu@3 has a stronger long-tail item mining capability than both CF-IT @1 and CF-IT @2. Figure 8 shows the CF-IT@3algorithms exceeds CF-IT@1 and CF-IT@2 in terms of performance metrics Precision, Recall and Coverage.

Figures 5-8 clearly and heuristically indicate that the algorithm tends to be stable for all aspects when is sufficiently large, although the rigorous such argument needs to be proved mathematically. Also, we can determine by these figures that the optimal values for K, N, α, β are K=20, N=18, α=0.899.

4. Conclusions

In this thesis, we have proposed a new collaborative filtering recommendation algorithm CF-IT which leverages tag information to improve the recommendation result. The thesis first defines the rating similarity by items' historical rated behaviors using graph theory, and secondly, defines the tag similarity by items' tag information. Finally, we combine the rating similarity and tag similarity to obtain comprehensive item similarity model. The experiment result shows that CF-IT excels other peer algorithms in terms of recommendation evaluation metrics. As the future work, we plan to parallelize the algorithm, and increase the amount of experimental data.

References

[1] Groh G, Ehmig C. Recommendations in taste related domains: collaborative filtering vs. social filtering[C]//Proceedings of the 2007

international ACM conference on Supporting group work. Sanibel: ACM Press, 2007: 127-136.

[2] Ma H, King I, Lyu M R. Effective missing data prediction for collaborative filtering[C]// SIGIR 2007: Proceedings of the, International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, the Netherlands, July. DBLP, 2007:39-46.

[3] Wang S, Xie Y, Fang M. A collaborative filtering recommendation algorithm based on item and cloud model[J]. Wuhan university journal of natural sciences, 2011, 16(1): 16-20.

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

[5] Liu H, Hu Z, Mian A, et al. A new user similarity model to improve the accuracy of collaborative filtering ☆[J]. Knowledge-Based Systems, 2014, 56(3):156-166.

[6] Satsiou A, Tassiulas L. Propagating Users' Similarity towards Improving Recommender Systems[C]// Web Intelligence. ACM, 2014:221-228.

[7] Zheng L, Yang S, He J, et al. An optimized collaborative filtering recommendation algorithm[C]//Cloud Computing and Internet of Things (CCIOT), 2016 2nd International Conference on. IEEE, 2016: 89-92.

[8] in J, Chen Q. A trust-based Top-K recommender system using social tagging network[C]// International Conference on Fuzzy Systems and Knowledge Discovery. IEEE, 2012:1270-1274.

[9] Firan C S, Nejdl W, Paiu R. The benefit of using tag-based profiles[C]//Web Conference, 2007. LA-WEB 2007. Latin American. IEEE Press, 2007: 32-41.

[10] Xiang Liang. Recommender Systems in Action[M]. Beijing: Posts & Telecom Press, 2012:25-79.

YuHong Zhou, was born in Tongling of Anhui province in 1989. He is now a graduate student in Chongqing University of Posts and Telecommunications. His research concerns Mobile communication techniques..