IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 4 Issue 5, May 2017
ISSN (Online) 2348 – 7968 | Impact Factor (2016) – 5.264
www.ijiset.com

Collaborative Filtering Recommendation Algorithm Using Implicit Similarity in Preference Relationships

Yu Hong Zhou[*], Zhi Jie Duan, Wei Jiang

*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 using implicit similarity in preference relationships(CF-ISP). Firstly, it divides the user group based on the user attributes, and secondly, it defines the item preference in the user group and calculates the preference similarity between the items. Finally, it introduces the preference similarity to the Probabilistic Matrix Factorization recommendation algorithm.The thesis conducts the experiment to verify the validity and the accuracy of the algorithm.

Keywords: *Recommendation System, Collaborative Filtering, preference similarity, user group*

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 divide the user group based on the user attributes, and then calculates the preference similarity between the items. Finally, it introduces the preference similarity to the Probabilistic Matrix Factorization recommendation algorithm. It first uses the user's age attribute, gender attribute and preference attribute to calculate the interest distance between the different users, and utilizes the k-medoids method to divide the user group, and then the formal definition of the item preference is given to calculate preference similarity. Finally, it introduces the preference similarity to the Probabilistic Matrix Factorization recommendation algorithm and redefine the item's potential feature vector. The local optimum of the potential feature vector can be obtained according to the matrix factorization and SGD method. 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, 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. Agarval [3] proposed a collaborative filtering framework based on interaction intensity and adaptive user similarity, and explored the idea of adaptive similarity computation between users.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 SVD algorithm proposed in [5] uses explicit (ratings) and implicit (viewing history) information from users in a factorization model, called SVD++. 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 the textonly-based, hashtag-based and URL-based approaches to model user profiles and calculated the relevance between users. Work in [8] proposed a matrix factorization method to estimates unknown user-to-user rating based on given trust relation and rating records between social network users. 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. User group division

2.1 related definition

Define user set $U=\{u_1,u_2,...,u_m \mid m=|U|\}$, given a positive integer K and division method, if U is divided non-empty subset U_l of size k, and $\sum_{l=1}^{K} U_l = U$, we say U_l is one user group.

Define $a'(u)$ and $g'(u)$ represents the user age attribute value and gender attribute value, the calculation as follows:

$$a'(u)=\frac{a(u_i)-\min(a)}{\max(a)-\min(a)} \quad (1)$$

$$g'(u)=\begin{cases}0, & g(u)=\text{male}\\ 1, & g(u)=\text{female}\end{cases} \quad (2)$$

Define $d(u,v)$ represents the interest distance between the user u and the user v , the calculation as follows:

$$d(u,v)=\alpha\cdot d_{age}(u,v)+\beta\cdot d_{gender}(u,v)+\gamma\cdot d_{pref}(u,v) \quad (3)$$

Where $d_{age}(u,v)$ represents the age attribute distance, $d_{gender}(u,v)$ represents the gender attribute distance, $d_{pref}(u,v)$ represents the preference attribute distance, parameter α , β and γ is weight factor and $\alpha+\beta+\gamma=1$. $d_{age}(u,v)$, $d_{gender}(u,v)$ and $d_{pref}(u,v)$ calculates as follows:

$$d_{age}(u,v)=|a'(u)-a'(v)| \quad (4)$$

$$d_{gender}(u,v)=|g'(u)-g'(v)| \quad (5)$$

$$d_{pref}(u,v)=\frac{H(T(p(u),p(v)))}{H(T_h)} \quad (6)$$

Where $T(p(u),p(v))$ is a subtree, and the node $p(u)$ and node $p(v)$ minimum common parent node as the subtree root node, T_h is preference classification tree, $H(T)$ represents the height of the tree.

Define bipartite graph model $G<U,E,w>$, U is the user node set, E is the edge set, $e(v_u,v_i)\in E$ represents there is an edge connect the node v_u and v_i , the weight $w(v_u,v_i)$ represents the interest distance of the user node v_u and v_i . The user interest distance matrix M as follows:

$$M=[dist(i,j)]_{m\times m} \quad (7)$$

Where $dist(i,j)$ is the interest distance between the user i and user j .

2.2 user group division

After giving the relevant formula and definition, the paper uses the k-medoids method to divide the user group.

IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 4 Issue 5, May 2017
ISSN (Online) 2348 – 7968 | Impact Factor (2016) – 5.264
www.ijiset.com

The pseudocode of the collaborative filtering recommendation algorithm using implicit similarity in preference relationships as follows:

CF-ISP algorithm pseudocode

Algorithm: Collaborative Filtering Recommendation Algorithm Using Implicit Similarity in Preference Relationships

Input: user interest distance matrix M, the number of user group N, user u

Output: N user groups

1: initialize $G = \{g_1, g_2 ... g_N\}$

2: initialize $C_c = \{c_1, c_2, ..., c_k\}$

3: while $\| M_{in\_dist}[c_k^{t+1}][c_k^t \|_2^2 < \varepsilon$ **do**

4: for each user $u_i \in U$ **do**

5: for each medoid $c_i \in C_c$ **do**

6: $dis(u_i, c_i) = M_{in\_dist}[u_i][c_i]$

7: end for

8: $dis(u_i, c_m) = \min\{M[u_i][c_m], ..., M[u_i][c_k]\}$

9: $g_m = c_m \bigcup u_i$

10: end for

11: for each user $u_i \in g_i \,\&\&\, u_i \neq c_i$ **do**

12: for each user $u_j \in g_i \,\&\&\, u_j \neq c_i$

13: $dis\_sum(u_i) += M_{in\_dist}[u_i][u_j]$

14: end for

15: $u_i = \min\{dis\_sum(u_1), ..., dis\_sum(u_m)\}$

16: end for

17: $c_i = u_i$

18: return

3. Implicit similarity in preference

In the previous section, we divide the user to k user groups. For each user group, we divide the original rating matrix into $R = R_1 \bigcup R_2 \bigcup ... \bigcup R_k$, the submatrix R_i is consist of all users and rating item and rating point in i-th user group.

Define user set $U = \{u_1, u_2, ..., u_m\}$, any item i, $U(i) = \{u \in U \mid R_{u,i} \neq \varnothing\}$ represents the user set who has rated the item i, $\forall C_k \in C$

$$pref(i) = \frac{|C_k \bigcap U(i)|}{|U(i)| + \alpha} \quad (0 \leq pref(i) \leq 1) \tag{8}$$

$pref(i)$ represents the item's preference in the user group k. We construct the item-user group preference matrix by the preference value, as table 1 shows:

item-user group preference matrix

	c_1	...	c_j	...	c_k
$item_1$	$P_{1,1}$	...	$P_{1,j}$	...	$P_{1,k}$
...	...	...	...	...	...
$item_j$	$P_{j,1}$	...	$P_{j,j}$	...	$P_{j,k}$
...	...	...	...	...	...
$item_n$	$P_{n,1}$	...	$P_{n,j}$	...	$P_{n,k}$

Where $P_{n,k}$ represents the preference of the item $item_n$ in user group c_k.

Define item set $I = \{i_1, i_2, ..., i_k\}$, and $\forall i \in I$, it uses $f_i = (P_{i,1}, P_{i,2}, ..., P_{i,k})$ represents the preference feature vector of the item i, and the implicit similarity in preference relationships between the item i and item j can represent as follows:

$$sim(i, j) = \cos(f_i, f_j) = \frac{\sum_{n=1}^{k} p_{in} p_{jn}}{\sqrt{\sum_{n=1}^{k} p_{in}^2} \sqrt{\sum_{n=1}^{k} p_{jn}^2}} \tag{9}$$

4. Recommendation algorithm CF-ISP

In the previous section, it defines the implicit similarity in preference relationships, we introduce this similarity to the Probabilistic Matrix Factorization recommendation algorithm and propose a novel collaborative filtering by using implicit similarity in preference relationships(CF-ISP), as the Fig1 shows:

IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 4 Issue 5, May 2017
ISSN (Online) 2348 – 7968 | Impact Factor (2016) – 5.264
www.ijiset.com

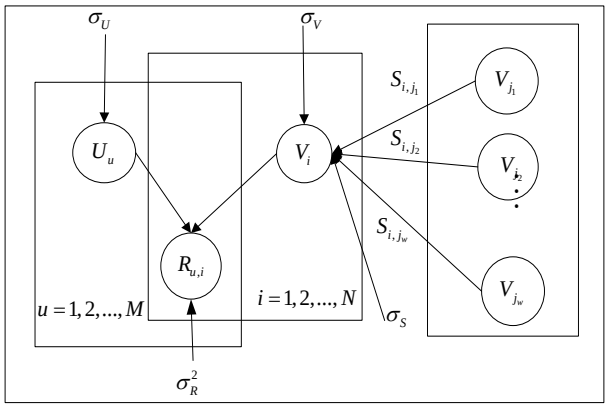

Fig 1:CF-ISP model

The thesis redefines the item's potential feature vector as follows:

$$\hat{V}_i = \sum_{j \in N_i} V_j S_{i,j} \qquad (10)$$

Where N_i is the neighbor set of the item i, $S_{i,j}$ is implicit similarity in preference relationships between item i and item j, and $\sum_{j \in N(i)} S_{i,j} = 1$. The Conditional Distributions of the rating matrix R define as follows:

$$p(R \mid U, V, \sigma_R^2) = \prod_{i=1}^{m} \prod_{j=1}^{n} [N(R_{i,j} \mid g(U_i^T V_j), \sigma_R^2)]^{I_{ij}^R} \qquad (11)$$

Where I_{ij}^R is the feature function, if the user U_i has rated the item V_j $I_{ij}^R = 1$ else $I_{ij}^R = 0$.

Based on the Bayes theorem and likelihood function of Eq.(11) we can obtain the posterior probability of U and V:

$$p(U, V \mid R, S, \sigma_R^2, \sigma_S^2, \sigma_U^2, \sigma_V^2) \propto p(R \mid U, V, \sigma_R^2) p(U \mid \sigma_U^2)\ p(V \mid S, \sigma_V^2, \sigma_S^2) \qquad (12)$$

Where $p(V \mid S, \sigma_V^2, \sigma_S^2) \propto p(V \mid S, \sigma_S^2) \times p(V \mid \sigma_V^2)$ the thesis incorporates the implicit similarity $sim(i, j)$ into PMF, so $p(V \mid S, \sigma_S^2)$ can be defined as follows:

$$p(V \mid S, \sigma_S^2) = \prod_{i=1}^{N} N(V_i \mid \sum_{j \in N_i} V_j S_{i,j}, \sigma_S^2 I) \qquad (13)$$

We proposed by minimizing its arithmetic negation function $L_F(R, U, V, S)$, the task is transformed to the following minimization problem:

$$L_F(R,U,V,S) = \frac{\lambda_S}{2} \sum_{i=1}^{N} \left(\left(V_i - \sum_{j \in N(i)} V_j S_{i,j} \right)^T \left(V_i - \sum_{j \in N(i)} V_j S_{i,j} \right) \right) + \frac{\lambda_U}{2} \sum_{u=1}^{M} U_u^T U_u + \frac{\lambda_V}{2} \sum_{i=1}^{N} V_i^T V_i + \sum_{u=1}^{M} \sum_{i=1}^{N} I_{u,i}^R \left(R_{u,i} - U_u^T V_i \right)^2 \qquad (14)$$

Where $\lambda_U = \sigma_R^2 / \sigma_U^2$, $\lambda_V = \sigma_R^2 / \sigma_V^2$, $\lambda_s = \sigma_R^2 / \sigma_S^2$ are the regularization to avoid model overfitting. To minimize the objective function $L_F(R, U, V, S)$ in Eq.(14), we use the gradient descent on $\partial L_F / \partial U_u$ and $\partial L_F / \partial V_i$ for each pair U_u and V_i, as follows:

$$\frac{\partial L_F}{\partial U_u} = \sum_{i=1}^{N} I_{u,i}^R V_i (U_u^T V_i - R_{u,i}) + \lambda_U U_u \qquad (15)$$

$$\frac{\partial L_F}{\partial V_i} = \sum_{i=1}^{M} I_{u,i}^R U_u (U_u^T V_i - R_{u,i}) + \lambda_V V_i + \lambda_S (V_i - \sum_{j \in N(i)} V_j S_{i,j}) - \lambda_S \sum_{j \mid i \in N(j)} (V_j - \sum_{x \in N(j)} V_x S_{j,x}) \qquad (16)$$

5. Experiment

5.1 Datasets

In our experiments, we used the MovieLens10M datasets, it contains 71567 users, 10681 movies. We evaluated the completion by RMSE(Root Mean Square Error)

IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 4 Issue 5, May 2017
ISSN (Online) 2348 – 7968 | Impact Factor (2016) – 5.264
www.ijiset.com

3.2 experiment result

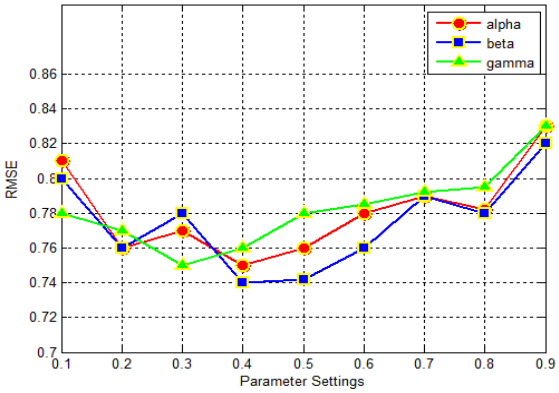

Fig.2: parameter and RMSE corresponding relation

In Figure 2, the experimental results are presented. When $\alpha = 0.4$ $\beta = 0.4 \sim 0.5$ $\gamma = 0.3$, the RMSE attain optimal value changes with the parameter. When $\beta = 0.4 \sim 0.5$, RMSE reaches minimum 0.74. So we consider the user attribute that the parameter β corresponds to has the most influential and importance to the recommender results. It reintroduces the condition $\alpha+\beta+\gamma=1$ we can draw a conclusion that when $\beta = 0.4$, $\alpha = 0.35$, $\gamma = 0.25$, RMSE reaches minimum and achieve best performance of recommendation.

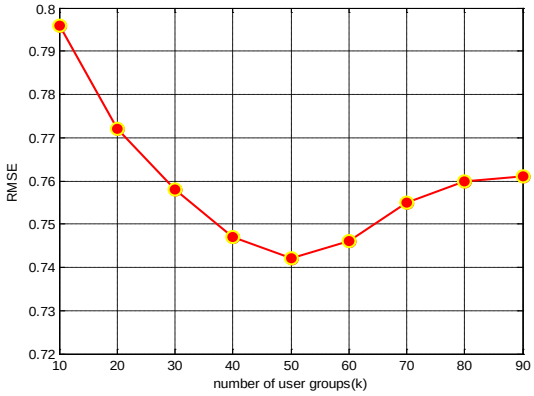

Fig.3: k and RMSE corresponding relation

In Figure 3, the experimental results are presented. The RMSE decreases with increasing of user group k, when $k = 50$ RMSE reaches minimum 0.74. Continue increasing k, RMSE increases and it reduces the recommendation quality.

Table 1 : λ_s and RMSE corresponding relation

λ_S	$f = 5$	$f = 10$
0.001	0.7655	0.7632
0.01	0.7476	0.7459
0.1	0.7459	0.7433
0.5	0.7572	0.7569
1	0.7667	0.7661
5	0.7702	0.7698
10	0.7798	0.7729

In Table 1, it is easy to find that in the same dimension, changes in parameter λ_s has the great influential to RMSE. RMSE decreases with increasing of λ_s, when $\lambda_s \in [0.01, 0.1]$, RMSE descends to its lowest level.

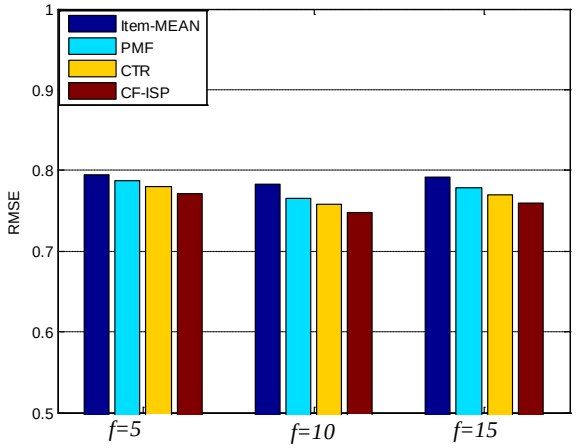

Fig.4: feature number and RMSE corresponding relation

In Figure 4, the experimental results are presented. RMSE decreases with increasing of feature number f, when $f = 10$ RMSE descends to its lowest level. Increasing f, the performance of the recommender system starts to slide, But the CF-ISP algorithm still has the lowest RMSE compared to the other three algorithms.

6. Conclusions

In this thesis, we have proposed a new collaborative filtering recommendation algorithm CF-ISP which leverages preference information to improve the recommendation result. The thesis first divide the user group based on the user attributes, and then calculates the

preference similarity between the items. Finally, it introduces the preference similarity to the Probabilistic Matrix Factorization recommendation algorithm. The experiment result shows that CF-ISP 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] V. Agarwal and K. K. Bharadwaj. A collaborative filtering framework for friends recommendation in social networks based on interaction intensity and adaptive user similarity.Social Network Analysis & Mining, 3(3):359-379, 2013.

[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] Y. Koren. Factor in the neighbors: Scalable and accurate collaborative filtering. ACM Transactions on Knowledge Discovery from Data (TKDD), 4(1):1–24,Jan. 2010.

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

[7] M. Chechev and P. Georgiev. A multi-view content-based user recommendation scheme for following users in twitter.In Social Informatics, pages 434-447. ACM, 2012.

[8] Jamali, M. and Ester, M. 2010. A matrix factorization technique with trust propagation for recommendation in social networks. In Proceedings of the fourth ACM Conference on Recommender Systems(Sept 26-30, 2010,Barcelona, Spain), RecSys'2010, ACM, 135-142.DOI=http://doi.acm.org/10.1145/1864708.1864736.

[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.

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..

Zhi Jie Duan, was born in Jian of JiangXi province in 1992. He is now a graduate student in Chongqing University of Posts and Telecommunications. His research concerns Mobile communication techniques.

Wei Jiang was born in ZheJiang province in 1992. He is now a graduate student in Chongqing University of Posts and Telecommunications. His research concerns Mobile communication techniques.