One Improved Collaborative Filtering Method Based on Information Transformation

Zhaoxing Liu
Business School, University of Shanghai for Science and Technology
YangPu Area Jungong Road 334, 200093, Shanghai, China
Tel: 86-021-139-1641-3236 E-mail: liouzhaoxing@126.com

Ning Zhang
Business School, University of Shanghai for Science and Technology
YangPu Area Jungong Road 334, 200093, Shanghai, China
Tel: 86-021-130-6168-2793 E-mail: zhangning@usst.edu.cn

The research is *supported by Natural Science Foundation of P.R. China (No. 70971089) and Shanghai Leading Academic Discipline Project(No. S30501).*

Abstract

In this paper, we propose a novel method combined classical collaborative filtering (CF) and bipartite network structure. Different from the classical CF, user similarity is viewed as personal recommendation power and during the recommendation process; it will be redistributed to different users. Furthermore, a free parameter is introduced to tune the contribution of the user to the user similarity. Numerical results demonstrate that decreasing the degree of user to some extent in method performs well in rank value and hamming distance. Furthermore, the correlation between degree and similarity is concerned to solved the drastically change of our method performance.

Keywords: Collaborative filtering, Bipartite network structure, Personal recommendation

1. Introduction

With the rapid development and spreading application of web 2.0, a large amount of personalized information yield and was dumped on the Internet in the last decade (Solomon, 2004). There are so many alternatvies that people sometimes frustrated to get what they really want from the Internet. There are millions of books on amazon.com that you can buy, and thousands of web pages on dilicious.com that you can collect and browser. This is so called informationoverload: we face too much data and sources to be able to get the relevant things. The benchmark of the informaiton filtering is search engine. Search engine provides users with what they care by the keyword (keywords), however, the same result to all the user cannot cater for users' different taste. And sometimes what you want maybe hard to expressive, even the user do not qualified. Therefore, the search engine can't work as we expect in those circumstances (liu, 2009).

As a consequence, how to find out the relevant things for our users becomes an urgent problem. The emergence of the recommendation system brings more convenience to users. Many diverse recommendation techniques have been developed, including collaborative filtering (Hill,1994, Resnick,1994,Linden,2003,Adomavicius,2005),content-basedanalysis(Pazzani, 2007),spectral analysis, latent semantic models (Hofmann,2004) and Dirichlet allocation and hybird algorithm (Zhou,2010) and so on. Collaborative Filtering has been proved to be get great success in many domains--especially in entertainment domain, such as book recommendation in amazon.com (Linden,2003), movie recommendation in netflix.com, and so on. Despite its success, the performance of the CF is limited by the sparsity of the data. Recently, some physics dynamics methods, including mass diffusion (Zhou, 2007), heat conduction (Zhang,2007), and trust-based recommendation (Zhou,2007), show good perfromance in recommendation system.

In our paper, we integrated CF and information transformation together, which is extendible CF (ECF). In order to make our method more available, a free parameter β was introduced to keep our method more scalable.

2. Method and numerical results

Recommendation systems (RS) predict ratings of the item or suggest a list of items that is unknown to the user. A RS generally consists of users and items.We can model this RS as bipartite network. In this bipartite network, node sets consist of user set and item set. There is a link if and only if the user collects the item. In other words, there are no links between nodes in the same set. Meanwhile, some notations are introduced, which we use throughout the rest of the paper. Denote item set $O = \{o_1, o_2, \ldots, o_m\}$ and user item set $U = \{u_1, u_2, \ldots, u_n\}$. And a_{ij} is an $m \times n$ adjacent item-user matrix of the bipartite network, where $a_{il} = 1$ if user l collect item i and $a_{il} = 0$ otherwise.

In a standard CF (SCF), what you need to do is to obtain the user-pair similarity matrix (Konstan, 1997, Sarwar,2001, Herlocker,2004,Bobadilla,2010). It reflects the correlation between user-pair according to the overlap items of the two users, defined as:

$$s_{ij} = \frac{\sum_{k=1}^{m} a_{ki} a_{kj}}{\min(k(u_i), k(u_j))} \tag{1}$$

Where $k(u_i) = \sum_{k=1}^{m} a_{ki}$ denotes the degree of the user i. In paper (Liu, 2009), which introduced method SA-CF, it evaluated the user similarity by information transformation like this:

$$s_{ij} = \frac{1}{k(u_j)} \frac{\sum_{k=1}^{m} a_{ki} a_{kj}}{k(o_j)} \tag{2}$$

where $k(o_j) = \sum_{l=1}^{m} a_{jl}$ denotes the degree of the item j. s_{ij} indicates that the resource diffuses from user i to user j. After this step, what the user i gets from other users (including itself) weighs the importance of the user in this system. Denote $Si = \sum_{k=1}^{n} s_{ik}$, which is seemed as the popularity and the user i is very active in this recommendation system and has more significance when we recommends items to other users if the vaule of Si is big enough. We can view this resource Si as the initial resource of user in this recommendation system. So, the rewrited predictation value for uncollected item j of user i is:

$$v_{ij} = \frac{1}{\sum_{l=1, l=i}^{n} s_{il}} \sum_{l=1, l \neq i}^{n} \frac{s_{il} a_{jl} Sl}{k(u_l)} \tag{3}$$

Now we can predict each uncollected item for users and sort them descending order of predication value and recommend those at the top. In a word, the framework of this algorithm is organized as follows:

(1) calculate the user similarity matrix $\{s_{ij}\}$ based on infromation transfrom ;

(2) calculate the similarity $\{Si\}$ one with any other users ;

(3) predicate the score of each uncollected item for each user according to formulation above ;

(4) sort the score of uncollected item for each user in descending order and recommended item in the top;

The benchmark dataset to which we applied our algorithm and others can download from www.grouplens.org. The total dataset contains 943 users and 1682 movies. Users rate the movie with different scale from 1 to 5, 1 denoting the user dislikes the movie most and 5 denoting the user prefers it and 3 is the middle. Before we applied the method, we preprocesses the dataset: we only keep the record with rate no less than 3 (representing the user does not dislike the movie). There is 10^5 records in the original dataset, 82.25% of which with rate more than 2. After that, we have 82520 records in our dataset. In order to test the recommendation methods, we randomly split the records in two separated parts. One is train set, with the information we know clearly and containing 90% of the records. The other is test set and we know nothing about it. The measurement in our method to test the algorithm is rank value. In our context, if the item is in the test set, it represents the user like it. So the position of this item in the recommendation list should high. The average rank value, averged all items in the test set, can evaluate the performance of the three method. The average rank value are 0.130, 0.121 and 0.110 for SCF, SA-CF and ECF repectively. Apparently, our method shows us good performance.

3. Improved ECF method

In order to extend our method, we introduce a free parameter β to tune the perfromance of the method. It is common that information of the popular things is comparatively easy to obtain in our daily life. So even the

user-pair has a big overlap, it does not mean the two users have the same taste. On the contrary, even a small overlap may indicate the two users share common taste if the things is not so popular. We introduce β to regulate this effect. Now the predication formulation is :

$$v_{ij} = \frac{1}{\sum_{l=1,l=i}^{n} s_{il}} \sum_{l=1,l\neq i}^{n} \frac{s_{il} a_{jl} Sl}{k^{\beta} k(u_l)} \tag{4}$$

Figure 1 demonstrates that the rank value is function of β. It is clearly observed that the accuracy of the our method reaches the best performance when $\beta = 1$, which is different from the value obtained by paper (liu,2009). Average degree of this method K, averaged over all items with a certain length of recommendation list, indicates that whether items recommended by RS is popular or not. A good method should give user more unpopular items, because that the popular item is easy to get and it is not necessary for us to recommend any more. Hamming distance h_{ij}, is defined as difference between any two users in recommendation list, which reads $h_{ij} = 1 - \frac{Q}{L}$, where Q is the overlap of the user i and user j in recommendation list L. H, averaged over all hamming distance between all user-pairs in this RS, can reflect the average difference between user's taste. The bigger of the H, the better of the performance.

From figures 1, 2 and 3, it is clearly observed that when the free parameter is no greater than 3, the performance of our method works well, which means the method has good rank value and high hamming distance. However, the performances of this method lose it's effect when the $\beta \geq 3$. It seems like that 3 is a critial point for our method, which is not found in any other research before. And there is no monotonous correlation between rank value, K, H and β. Figure 4 demonstrates the correalation between the degree and Sk. Sk, averaged all the similarity Si of users with the degree k. The similairty decreases drastically at the begining and when the degree surpasses 25, there is almost no change. Probably, that is why there exists this critical point though the cirital point does not reach the same value in above three figures.

4. Conclusion

In this paper, we integrate the user-pair similarity and SA-CF and view Si as recommendation power. The ECF show us good performance than SA-CF and SCF. When a free parameter β is introduced, a critical point is observed which was never discovered before. Through the numerical result, there is no monotonous correlation between rank value, K, H and this tunable parameter β. The appearence of the critical point of our method probably is unveiled by the distribution of Sk. The value of the β in one recommendation system counts on the target you want to achieve.

References

Adomavicius, G & Tuzhilin, A. (2005). Toward the next generation of the recommender system: a survey of the state of the art and possbile extensions. IEEE *Trans Knowledge Data Eng* 17:734-749.

Bobadilla, J & Serradilla., F and Bernal, J. (2010). A new collaborative filtering metric that improves the behavior of recommender systems, *Knowledge-Based Systems*, v.23 n.6, p.520-528.

Cai-nicolas , Ziegler & Sean , M. Mcnee. (2005). Improving recommendation lists through topic diversification, *Proceedings of the 14th international conference on World Wide Web*, May 10-14, Chiba, Japan.

Hill., W & Stead, L and Rosenstein, M. (1994). Recommendation and evaluating choices in a virtual community of use.*Proc Conf Human Factors in Computing System*. Dever.194-201.

Hofmann, T. (2004). Latent semantic models for collaborative filtering. ACM *Trans lnf Syst*.22:89-115.

Herlocke, .J.L & Konstan, J.A & Terveen, L.G and Riedl, J.T. (2004). Evaluating Collaborative Filtering Recommender Systems, *ACM Trans. Information Systems,* vol. 22, no. 1, pp. 5-53.

Konstan, J.A & Miller, B.N and Maltz, D. (1997). Grouplens:Applying collaborative to usernet news. *Comm ACM*.40(3) :77-87.

Linden, G & Smith, B and York, J. (2003). Amazon.com recommendations: item-to-item collaborative filtering. IEEE *Internet Comput* 7:76 80.

Liu, J.G & Zhou, T and Wang,G.H. (2009). the research progress of the personlized recommendation system,*the progress of natural science*. 19.1-15 (in Chinese).

Liu, J.G & Wang, B.H and Guo, Q. (2009). Improved collaborative filtering algorithm via information transformation. *International Journal of Modern Physics C* Vol. 20.No. 2,285–293.

Pazzani, M.J & Billsus, D. (2007). Content-based recommendation systems. *Lect Notes Comput Sci* 4321:325-341.

Resnick, P & Iakovou, N and Sushak, M. (1994). Group: An open architecture for collaborative filtering of netnews. *Proc 1994 Computer Supported Cooperative Work Conf*,chapel Hill.175-186.

Sarwar, B & Karypis, G & Konstan, J and Riedl, .J. (2001). Item-Based Collaborative Filtering Recommendation Algorithms, Proc. 10th Int'l WWW Conf.

Solomon, G & Schrum, L. (2007). *Web 2.0: New Tools, New Schools. International Society for Technology in Education*,(chapter 1).

Zhou, T & Kuscsik, Z & Liu, J.G & Medo, M & Wakeling, R.J and Zhang, Y.C. (2010). Solving the apparent diversity-accuracy dilemma of recommender systems, 10.1073/pnas.1000488107.

Zhou., T & Ren.J & Medo.M and Zhang, Y.C. (2007). Bipartite network projection and personal recommendation. *Phys. Rev. E* 76, 046115.

Zhou, T & Ren, J & Medo, M and Zhang, Y.C. (2007). *Phys. Rev. E* 76. 046115.

Zhang, Y.C & Blattner., M and Yu, Y.K. (2007). Heat conduction process on community networks as a recommendation model. *Phys Rev Lett* 99:154301.

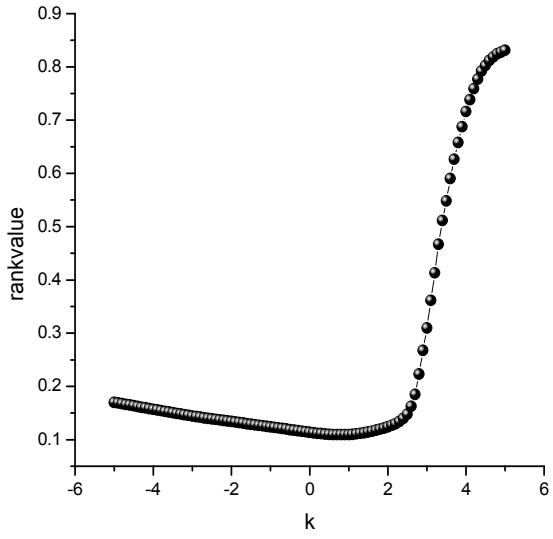

Figure 1.

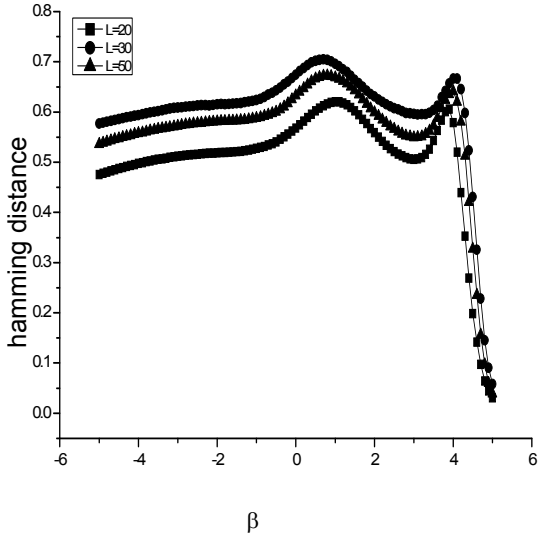

Figure 2.

Figure 1. The correalation between rank value and β. Different from the SA-CF, the curve is very smooth when β is no greate than 3. When β surpasses this value, the performance of our method decreases rapidly. Figure 2. hamming distance vs β. The square line, circle line and triangle lines present cases with recommendation list of length 20,30,50 repectively

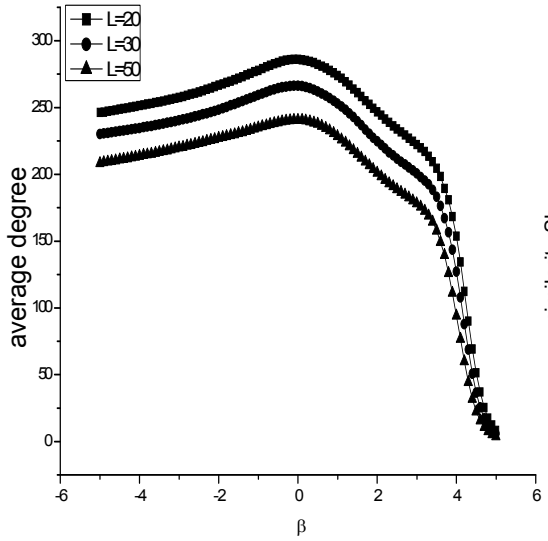

Figure 3.

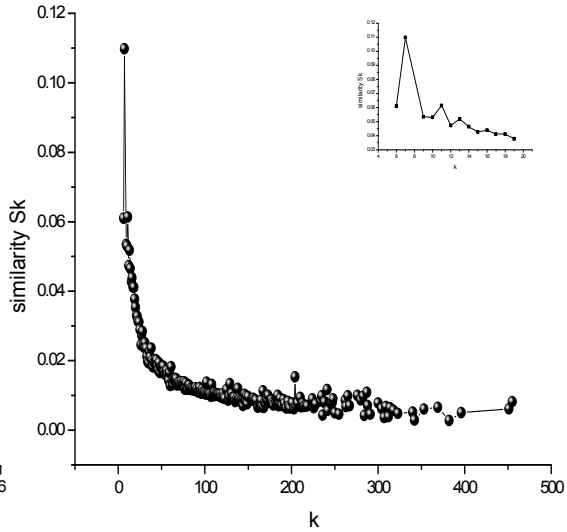

Figure 4.

Figure 3. average degree K as a function of β. Three lines from top to bottom represent the cases with recommendation list of length 20,30and 50 repectively. Figure 4. the correlation between degree and Sk. The inset shows the correalation with degree no greater than 25.