Front. Comput. Sci. China 2009, 3(3): 417–420
DOI 10.1007/s11704-009-0050-2

RESEARCH ARTICLE

Improved collaborative filtering algorithm based on heat conduction

Qiang GUO[1, 4], **Jianguo LIU**(✉)[2, 3, 4], **Binghong WANG**[2, 3]

1 Business School, University of Shanghai for Science and Technology, Shanghai 200093, China
2 Research Center of Complex Systems Science, Shanghai University of Science and Technology, Shanghai 200093, China
3 Department of Modern Physics, University of Science and Technology of China, Hefei 230026, China
4 Department of Physics, University of Fribourg, Chemin du Musée 3, CH-1700, Switzerland

Abstract In this paper, we present an improved collaborative filtering (ICF) algorithm by using the heat diffusion process to generate the user correlation. This algorithm has remarkably higher accuracy than the standard collaborative filtering (CF) using Pearson correlation. Furthermore, we introduce a free parameter β to regulate the contributions of objects to user correlation. The numerical simulation results indicate that decreasing the influence of popular objects can further improve the algorithmic accuracy and diversity.

Keywords recommendation algorithm, collaborative filtering, heat conduction

1 Introduction

With the advent of the Internet [1], the exponential growth of the World-Wide-Web [2] and routers confront people with an information overload. We face too much data and sources to be able to find out those most relevant for us. Consequently, how to efficiently help people obtain information that they truly need is a challenging task nowadays [3]. A landmark for information filtering is the use of the search engine [4, 5], by which users could find the relevant web pages with the help of properly chosen keywords. However, the search engine has two essential disadvantages. First, it does not take into account personalization and returns the same results for people with far different habits. Second the search engine is a tool helping users to find the web pages at least containing some content known to them. Being an effective tool to address this problem, the recommender system has caught increasing attention from researchers to engineers, and has become an essential issue in Internet applications such as e-commerce systems and digital library systems [6]. For example, *Amazon.com* uses one's purchase record to recommend books [7], *AdaptiveInfo.com* uses one's reading history to recommend news [8], *Recipefinder.com* uses one's stated interests to recommend restaurants [9], and so on. Motivated by its significance in economy and society, the design of an efficient recommendation algorithm becomes a joint focus from engineering science to marketing practice. Various kinds of recommendation algorithms have been proposed, including the correlation-based methods [12, 13], content-based methods [14, 15], spectral analysis [16, 17], iteratively self-consistent refinement [18], principal component analysis [19], bipartite-network-based methods [20−23], and so on (see the review article [10, 11] and the references therein).

One of the most successful recommendation algorithms, called *collaborative filtering* (CF), has been developed and extensively investigated over the past decade [12, 13, 24]. The main idea of CF could be demonstrated in two steps. First, CF identifies a set of similar users from the past records, and then makes a prediction based on the weighted combination of those similar users' opinions. Despite its wide applications, collaborative filtering suffers from several major limitations including system scalability and accuracy [25]. Recently, some physical dynamics have been successfully introduced in CF algorithm. By using the diffusion process to compute the user similarities, Liu et al. proposed a modified CF algorithm by using the diffusion process to generate the user correlation [26], which has higher

Received December 12, 2008; accepted June 2, 2009

E-mail: liujg004@ustc.edu.cn

accuracy than the standard one. Furthermore, by considering the second-order correlations, Liu et al. designed an effective algorithm by depressing the influences of mainstream preferences [27]. It should be emphasized that two traditional physical approaches have been demonstrated to be of both high accuracy and low computational complexity, including mass diffusion [21, 22, 26, 27] and heat conduction [20]. Inspired by the heat conduction process [20], we introduce a improved collaborative filtering (ICF) method, which has remarkably higher accuracy than the standard CF.

2 Method

Denoting the object set as $O = \{o_1, o_2, \ldots, o_m\}$ and the user set as $U = \{u_1, u_2, \ldots, u_n\}$, a recommender system can be fully described by an adjacent matrix $A = \{a_{ij}\} \in R^{m,n}$, where $a_{ij} = 1$ if o_i is collected by u_j, and $a_{ij} = 0$ otherwise. For a given user, a recommendation algorithm generates an ordered list of all the objects he/she has not collected before. In the standard CF, the correlation between u_i and u_j can be evaluated directly by a Pearson-like form as

$$s_{ij}^c = \frac{\sum_{l=1}^m a_{li}a_{lj}}{\min\{k(u_i), k(u_j)\}}, \tag{1}$$

where $k(u_i) = \sum_{l=1}^m a_{li}$ is the degree of user u_i. In this paper, we assume each user is a heat resource. The target user would distribute his/her temperature to all the objects he/she has collected, and then each object sends the heat back to all the users who have collected it, the user correlation s_{ij} (the final temperature of user u_j) can be expressed as

$$s_{ij} = \frac{1}{k(u_i)} \sum_{l=1}^m \frac{a_{li}a_{lj}}{k(o_l)}, \tag{2}$$

where $k(o_l) = \sum_{i=1}^n a_{li}$ denotes the degree of object o_l.

In the standard CF algorithm, for the user-object pair (u_i, o_j), if u_i has not yet collected o_j (i.e., $a_{ji} = 0$), the predict score, v_{ij}, is given as

$$v_{ij} = \frac{\sum_{l=1}^n s_{li}a_{jl}}{\sum_{l=1}^n s_{li}}. \tag{3}$$

Based on the definitions of s_{ij} and v_{ij}, given a target user u_i, the ICF algorithm could be given.

3 How to evaluate the algorithmic performance?

The algorithmic accuracy is measured by *average ranking score* [22]. Indeed, a recommendation algorithm should provide each user with an ordered list of all its uncollected objects. For an arbitrary user u_i, if the entry u_i-o_j is in the probe set (according to the training set, o_j is an uncollected object for u_i), we measure the position of o_j in the ordered list. For example, if there are $L_i = 100$ uncollected objects for u_i, and o_j is the 10th from the top, the position of o_j is 10/100, denoted by $r_{ij} = 0.1$. Since the probe entries are actually collected by users, a good algorithm is expected to give higher average ranking scores, leading to small r_{ij}. Therefore, the mean value of the position r_{ij}, $\langle r \rangle$ (called *average ranking score* [22]), averaged over all the entries in the probe, can be used to evaluate the algorithmic accuracy: the smaller the average ranking score, the higher the algorithmic accuracy, and vice versa.

Besides accuracy, the mean value of Hamming distance, S, is taken into account to measure the algorithmic diversity [23]. The personal recommendation algorithm should present different recommendations to different users according to their different tastes and habits. The diversity can be quantified by the average Hamming distance, $S = \langle H_{ij} \rangle$, where $H_{ij} = 1 - Q_{ij}/L$, L is the length of recommendation list, and Q_{ij} is the overlapped number of objects in u_i and u_j's recommendation lists.

4 Numerical results

We use a benchmark data set, namely *MovieLens*[1], which consists of 1682 movies (objects) and 943 users. The users vote movies by discrete ratings from one to five. We suppose a movie is set to be collected by a user only if the giving rating is larger than 2. The user-object (user-movie) bipartite network after the coarse gaining contains 85250 edges. The data set is randomly divided into two parts: the training set contains 90% of the data, and the remaining 10% of data constitutes the probe.

Implementing the ICF and CF, the average value of ranking score are 0.1051 ± 0.0132 and 0.122 ± 0.0274. Clearly, under the simplest initial configuration, subject to the algorithmic accuracy, the ICF algorithm outperforms the standard CF.

5 The improved algorithm

In order to further improve the algorithmic accuracy, we propose a modified method. Taking into account the potential role of object degree may give better performance. Accordingly, instead of Eq. (2), we introduce a more complicated way to get user correlation

[1] http://www.grouplens.org

$$s_{ij} = \frac{1}{k(u_i)} \sum_{l=1}^{m} \frac{a_{li} a_{lj}}{k^{\beta}(o_l)}, \tag{4}$$

where β is a tunable parameter. When $\beta = 1$, this method degenerates to the algorithm mentioned in the above section. The case with $\beta > 1$ weakens the contributions of large-degree objects to the user correlation, while $\beta < 1$ will enhance the contributions of large-degree objects. According to our daily experience, if two users u_i and u_j have simultaneously collected a very popular object (with very large degree), it does not mean that their interests are similar. On the contrary, if both of the two users collected an unpopular object (with very small degree), it is very likely that they share some common and particular tastes. Therefore, we expect a larger β (i.e. $\beta > 1$) will lead to higher accuracy than the routine case $\beta = 1$.

Figure 1 reports the algorithmic accuracy as a function of β. The curve has a clear minimum around $\beta = 1.9$, which strongly support the above statement. Compared with the routine case ($\beta = 1$), the ranking score can be further reduced by 5.0% at the optimal value. Fig. 2 shows the positive correlation between S and β, which indicates that depressing the influences of high-degree objects makes the recommendations more personalized. The above simulation results indicate that ICF outperforms CF from the viewpoints of accuracy and diversity.

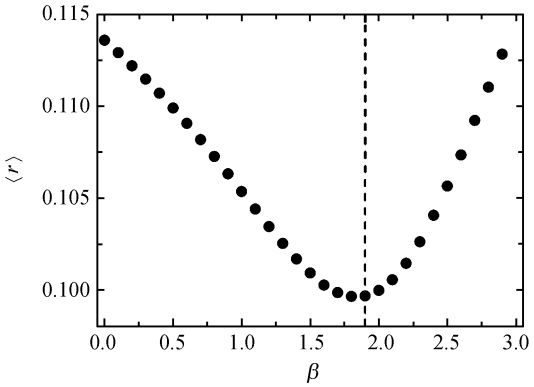

Fig. 1 The average ranking score $\langle r \rangle$ vs. β for the improved algorithm. The optimal β, corresponding to the minimal $\langle r \rangle = 0.0998$, is $\beta_{\rm opt} = 1.9$. All the data points are averaged over ten independent runs with different data-set divisions.

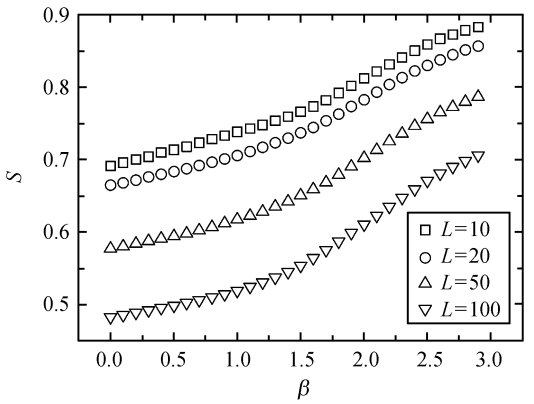

Fig. 2 S vs. β. Squares, circles, upper and lower triangles represent the lengths $L = 10$, 20, 50 and 100, respectively. All the data points are averaged over ten independent runs with different data-set divisions.

6 Conclusion

In this paper, we present a modified collaborative filtering algorithm based on a new user similarity definition, named heat conduction process. The ICF has obviously higher accuracy than the standard CF. The algorithmic complexity of the presented algorithm is the same with the one presented by Liu [26], $\mathbf{O}(m^2\langle k_u \rangle + mn\langle k_o \rangle)$, therefore, the computational complexity of ICF is much less than that of the standard CF. Furthermore, we presented an improved algorithm by considering the effect of object degree. The improved algorithm could further enhance the accuracy and personality by weakening the contribution of large-degree objects to user correlations.

How to automatically find relevant information for diverse users is a long-standing challenge in modern information science. We believe the current work can enlighten readers in this promising direction.

Acknowledgements We acknowledge *GroupLens Research Group* for providing us the data set. This work was partially supported by the National Natural Science Foundation of China (Grant Nos. 60744003 and 10635040), the National Basic Research Program of China (973 Program) (2006CB705500), the Swiss National Science Foundation (project 205120-113842), and GQ acknowledges the Research Fund of the Education Department of Liaoning of China (20060140).

References

1. Pastor-Satorras R, Vespignani A. Evolution and Structure of the Internet. Canbridge: Cambridge University Press, 2004
2. Broder A, Kumar R, Moghoul F, et al. Graph structure in the web. Computer Networks, 2000, 33: 309–320
3. Resnick P, Varian H R. Recommender systems. Communications of the ACM, 1997, 40: 56–58
4. Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 1998, 30: 107–117
5. Kleinberg J M. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999, 46: 604–632
6. Schafer J B, Konstan J A, Riedl J. E-commerce recommender applications. Data Mining and Knowledge Discovery, 2001, 5: 115–152
7. Linden G, Smith B, York J. Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Computing, 2003, 7: 76–80
8. Billsus D, Brunk C A, Evans C, et al. Adaptive interfaces for ubiquitous web access. Communications of the ACM, 2002, 45: 34–38
9. Burke R. Hybrid recommender systems: survey and experiments. User Modeling and User-Adapted Interaction, 2002, 12: 331–370
10. Adomavicius G, Tuzhilin A. Toward the next generation of recom-

mender systems: a survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 2005, 17: 734–749

11. Liu J -G, Chen M Z Q, Chen J, et al. Recent advances in personal recommendation systems. International Journal of Information and Systems Science, 2009, 5(2): 230–247
12. Herlocker J L, Konstan J A, Terveen K, et al. Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems, 2004, 22: 5–53
13. Konstan J A, Miller B N, Maltz D, et al. GroupLens: applying collaborative filtering to usenet news. Communications of the ACM, 1997, 40: 77–87
14. Balabanović M, Shoham Y. Learning information retrieval agents: experiments with automated web browsing. Communications of the ACM, 1997, 40: 66–70
15. Pazzani M J. Detecting change in categorical data: mining contrast sets. Artificial Intelligence Review, 1999, 13: 393–408
16. Billsus D, Pazzani M. Learning collaborative information filters. In: Proceedings of the Fifteenth International Conference on Machine Learning. Madison, 1998
17. Sarwar B, Karypis G, Konstan J, Riedl J. Application of dimensionality reduction in recommender system–A case study. In: Proceedings of the WebKDD 2000 Web Mining for E-Commerce Workshop at ACM SIGKDD. Boston, 2000, TR 00-043
18. Ren J, Zhou T, Zhang Y C. Information filtering via self-consistent refinement. Europhysics Letters, 2008, 80: 58007
19. Goldberg K, Roeder T, Gupta D, et al. Eigentaste: a constant time collaborative filtering algorithm. Information Retrieval, 2001, 4: 133–151
20. Zhang Y C, Blattner M, Yu Y K. Heat conduction process on community networks as a recommendation model. Physical Review Letters, 2007, 99: 154301
21. Zhang Y C, Medo M, Ren J, et al. Recommendation model based on opinion diffusion. Europhysics Letters, 2007, 80: 68003
22. Zhou T, Ren J, Medo M, et al. Bipartite network projection and personal recommendation. Physical Review E, 2007, 76: 046115
23. Zhou T, Jiang L L, Su R Q, et al. Effect of initial configuration on network-based recommendation. Europhysics Letters, 2008, 81: 58004
24. Huang Z, Chen H, Zeng D. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. ACM Transactions on Information Systems, 2004, 22: 116–142
25. Sarwar B, Karypis G, Konstan J, et al. Analysis of recommendation algorithms for E-commerce. In: Proceedings of the ACM Conference on Electronic Commerce. ACM, New York, 2000, 158–167
26. Liu J G, Wang B H, Guo Q. Improved collaborative filtering algorithm via information transformation. International Journal of Modern Physics C, 2009, 20: 285–293
27. Liu J G, Zhou T, Wang B H, et al. Highly accurate recommendation algorithm based on high-order similarities. 2008, arXiv: 0808.3726