Computer Physics Communications 185 (2014) 489–494

Contents lists available at ScienceDirect

Computer Physics Communications

journal homepage: www.elsevier.com/locate/cpc

Redundant correlation effect on personalized recommendation

Tian Qiu [a,*], Teng-Yue Han [a], Li-Xin Zhong [b], Zi-Ke Zhang [c,d], Guang Chen [a]

[a] School of Information Engineering, Nanchang Hangkong University, Nanchang, 330063, China

[b] School of Finance and Coordinated Innovation Center of Wealth Management and Quantitative Investment, Zhejiang University of Finance and Economics, Hangzhou, 310018, China

[c] Institute of Information Economy, Hangzhou Normal University, Hangzhou, 310036, China

[d] Web Sciences Center, University of Electronic Science and Technology of China, Chengdu, 610054, China

ARTICLE INFO

Article history:
Received 30 January 2013
Received in revised form
30 August 2013
Accepted 3 October 2013
Available online 10 October 2013

Keywords:
Infophysics
Heat conduction
Mass diffusion
Redundant correlation

ABSTRACT

The high-order redundant correlation effect is investigated for a hybrid algorithm of heat conduction and mass diffusion (HHM), through both heat conduction biased (HCB) and mass diffusion biased (MDB) correlation redundancy elimination processes. The HCB and MDB algorithms do not introduce any additional tunable parameters, but keep the simple character of the original HHM. Based on two empirical datasets, the *Netflix* and *MovieLens*, the HCB and MDB are found to show better recommendation accuracy for both the overall objects and the cold objects than the HHM algorithm. Our work suggests that properly eliminating the high-order redundant correlations can provide a simple and effective approach to accurate recommendation.

1. Introduction

People have to face a great amount of data with the flourish of internet technique. How to quickly find out what they actually want, e.g. select a preferred book from online bookstores, has become an important and practical problem. As an emerging powerful tool, the recommendation engine greatly helps people from the confusion brought about by the overload of information [1]. Seeking for the effective recommendation algorithm therefore attracts a central interest of scientists from different disciplines [2] including physics [3,4].

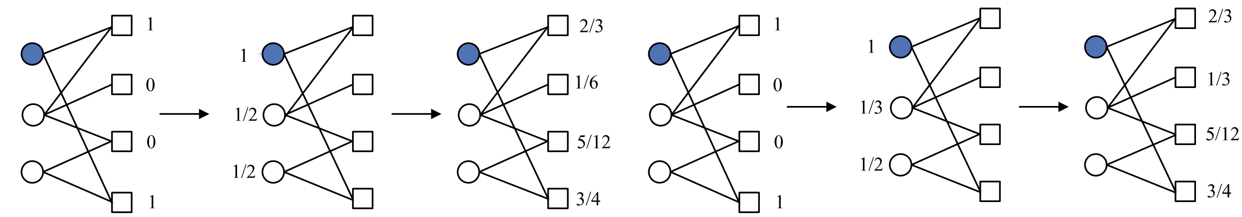
Different recommendation algorithms have been proposed. Those widely investigated algorithms include the collaborative filtering [5,6], the content-based method [7], and their relevant extensive studies [8–13]. Recently, thanks to fruitful achievements of complexity theory, complex-network based recommendation algorithms have been proposed [3,14–22]. Meanwhile, concepts from statistical physics have also been introduced into the network-based recommendation algorithms. Considering the heat conduction analogous process in a bipartite network, Y C Zhang et al. proposed a heat-conduction algorithm, which can generate a highly personalized recommendation [3]. Enlightened by this work, different diffusion patterns are incorporated into the network-based recommendation systems, e.g., the mass diffusion pattern [14].

Among these algorithms, an excellent algorithm is a hybrid method of heat conduction and mass diffusion (HHM), which successfully resolves the long-standing dilemma between recommendation accuracy and diversity [19]. Generally speaking, an excellent recommendation algorithm should at first be accurate. To be applicable, the algorithm should also not be time-consuming. The HHM algorithm covers these two advantages, with only one tunable parameter. Based on the HHM method, some improved methods are proposed. For example, a heterogeneous HHM method [23] is reported to show better performance than the HHM, by introducing heterogeneous initial configurations into the HHM method. However, it has two tunable parameters to optimize the algorithm, which therefore increases the algorithm complexity and computation time.

Aiming at developing an algorithm which is not only accurate but also simple, in this article, we investigate the high-order correlation effect in the HHM method, and propose a heat conduction biased (HCB) and a mass diffusion biased (MDB) method by eliminating the correlation redundancies of the system. In the network-based recommendation systems, if two objects α and β are collected by the same user i, the two objects are taken as directly correlated by the user i, i.e., the low-order correlation. However, objects might also have correlations by indirect collections. For example, besides the case that objects α and β are both collected

* Corresponding author. Tel.: +86 13576100249.
E-mail addresses: tianqiu.edu@gmail.com (T. Qiu), zhangzike@gmail.com (Z.-K. Zhang).

0010-4655/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cpc.2013.10.003



(a) Mass diffusion. (b) Heat conduction.

Fig. 1. An illustration of the resource reallocation process for (a) the mass diffusion and (b) the heat conduction.

by user i, if objects α and γ are collected by user i, and β and γ are collected by user j, then objects α and β have the higher-order correlation through the media object γ. The correlation redundancies therefore refer to such higher-order correlations in real systems. So far, the correlation redundancy elimination effect has only been investigated in some simple algorithms, e.g., the mass diffusion algorithm [22]. The reason might be that the redundancy-eliminated algorithm usually needs to introduce an additional tunable parameter to optimize the algorithm, which would make the algorithm much more complex and time-consuming. In our study, from both the theoretical and numerical simulation points of view, we show that the HCB and MDB algorithms successfully eliminate the high-order correlation redundancies without introducing any additional parameters, which leads to a better recommendation accuracy than the original HHM algorithm.

2. Hybrid algorithm of heat conduction and mass diffusion

A recommender system can be described by a bipartite network composed of a user set $U = \{u_1, u_2, \ldots, u_m\}$ and an object set $O = \{o_1, o_2, \ldots, o_n\}$. The links of the network are characterized by an adjacent matrix A, with $a_{\alpha i} = 1$ for the linked user–object pairs, and $a_{\alpha i} = 0$ for the unconnected pairs.

In the following algorithms, a so-called "resource" is introduced to objects. Firstly, for a specific user, an initial resource is assigned to each object, labeled as $\mathbf{f}^i = \{f_1^i, \ldots, f_\alpha^i, \ldots, f_n^i\}$. Here we assume the $f_\alpha^i = a_{i\alpha}$. Then the resources are reallocated according to a transformation matrix $\mathbf{W}$, and objects can obtain a final resource $\mathbf{f}'$. The resource reallocation process is formulated as,

$$\mathbf{f}' = \mathbf{W}\mathbf{f}. \tag{1}$$

For a particular user, rank the final resource $\mathbf{f}'$ according to the decreasing order, and recommend the top L objects to the user. Henceforth, the reallocation process plays a key role in the network-based recommendation algorithm.

We begin with the hybrid method of heat conduction and mass diffusion (HHM) proposed in [19], which takes the advantages of the mass diffusion method [14] and the heat conduction method. In the mass diffusion (MD) method, at first, each object distributes the resource to its neighboring users with an equal probability. Afterwards, the user redistributes its total level of resource with an equal probability to its neighboring objects. By summing up all the resources from their neighboring users, the objects then obtain their final level of resources. An example of the process of resource reallocation is illustrated in Fig. 1(a). The resource transformation matrix of the MD is formulated as,

$$W_{\alpha\beta}^{\mathrm{MD}} = \frac{1}{k_\beta}\sum_{j=1}^{m}\frac{a_{\alpha j}a_{\beta j}}{k_j}, \tag{2}$$

where k_β is the degree of object o_β, and k_j is the degree of user u_j, with the degree denoted as the number of links owned by the object and the user, respectively. We assume an object to be popular if the object has a high degree, otherwise, the object is to be cold. Due to assigning more priority to the popular objects with high degrees in the last diffusion step, the MD method has a high recommendation accuracy but a relatively low diversity.

In the heat conduction (HC) method illustrated in Fig. 1(b), the particular user i firstly receives an average level resource from its neighboring objects. Then the object again gets the average resource from all its neighboring users. The transformation matrix then reads,

$$W_{\alpha\beta}^{\mathrm{HC}} = \frac{1}{k_\alpha}\sum_{j=1}^{m}\frac{a_{\alpha j}a_{\beta j}}{k_j}, \tag{3}$$

where k_α is the degree of object o_α. Due to assigning more priority to the cold objects with low degrees in the last diffusion step, the HC method has a high recommendation diversity, but at the cost of the recommendation accuracy.

The HHM method [19] combines the HC and MD methods by,

$$W_{\alpha\beta} = \frac{1}{k_\alpha^{1-\lambda}k_\beta^{\lambda}}\sum_{j=1}^{m}\frac{a_{\alpha j}a_{\beta j}}{k_j}. \tag{4}$$

At a optimal value of the tunable parameter $\lambda \in [0, 1]$, the HHM shows a great advantage in both the recommendation accuracy and diversity.

3. Data

We employ two empirical datasets to test the performance of the algorithms, i.e., the *Netflix* and *MovieLens*, which are both movie rating systems with a five-level rating. The *Netflix* dataset is randomly selected from the huge dataset of the *Netflix* Prize, and the *MovieLens* is downloaded from the web site of GroupLens Research.[1] To construct the bipartite network, we add a link between a user and a movie if the rating of the user to the movie is no less than three. The *Netflix* contains 9999 users, 5870 objects and 815 917 links, and the *MovieLens* contains 943 users, 1682 objects and 100 000 links. To depict the characteristic of the dataset, we define its sparsity as the number of links proportional to the total number of the user–object links. The sparsity of the *Netflix* is calculated to be 1.39%, and of the *MovieLens* is calculated to be 6.30%, respectively. The total links of the user–object network are randomly divided into two subsets. That is, randomly delete 10% links as the test set to test the performance of the algorithms, and keep the remaining 90% links as the training set to make predictions.

4. Metrics

Whether an algorithm can make an accurate prediction for users or not is essential to evaluate the algorithm performance. In our study, we use three widely adopted indicators to measure the recommendation accuracy, i.e., the ranking score ($\langle RS\rangle$), Recall (R), and Precision (P).

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