Improved collaborative filtering algorithm based on heat conduction

Qiang GUO\textsuperscript{1, 4}, Jianguo LIU\textsuperscript{2, 3, 4}, Binghong WANG\textsuperscript{2, 3}

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

\textsuperscript{c} Higher Education Press and Springer-Verlag 2009

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 $\beta$ 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
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 \( O = \{a_1, a_2, \ldots, a_m\} \) and the user set as \( U = \{u_1, u_2, \ldots, u_n\} \), a recommender system can be fully described by an adjacent matrix \( A = \{a_{ij}\} \in \mathbb{R}^{m \times n} \), 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

\[
S_{ij} = \frac{\sum_{l=1}^{m} a_{il}a_{lj}}{\min\{k(u_i), k(u_j)\}}, \tag{1}
\]

where \( k(u_i) = \sum_{l=1}^{m} a_{il} \) 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

\[
s_{ij} = \frac{1}{k(o_i)} \sum_{l=1}^{m} \frac{a_{il}a_{lj}}{k(o_l)}, \tag{2}
\]

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

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

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

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