Knowledge-Based Systems xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Grey Forecast model for accurate recommendation in presence of data sparsity and correlation

Feng Xie [a,b,*], Zhen Chen [b,c], Jiaxing Shang [a], Geoffrey C. Fox [d]

[a] Department of Automation, Tsinghua University, Beijing 100084, China
[b] Research Institute of Information Technology, Tsinghua University, Beijing 100084, China
[c] Tsinghua National Lab for Information Science and Technology, Beijing 100084, China
[d] School of Informatics and Computing, Indiana University, IN 47408, USA

ARTICLE INFO

Article history:
Received 21 August 2013
Received in revised form 25 March 2014
Accepted 5 April 2014
Available online xxxx

Keywords:
Recommender systems
Collaborative filtering
Grey Forecast model
Data sparsity
Data correlation

ABSTRACT

Recently, recommender systems have attracted increased attention because of their ability to suggest appropriate choices to users based on intelligent prediction. As one of the most popular recommender system techniques, Collaborative Filtering (CF) achieves efficiency from the similarity measurement of users and items. However, existing similarity measurement methods have reduced accuracy due to problems such as data correlation and data sparsity. To overcome these problems, this paper introduces the Grey Forecast (GF) model for recommender systems. First, the Cosine Distance method is used to compute the similarities between items. Then, we rank the items, which have been rated by an active user, according to their similarities to the target item, which has not yet been rated by the active user; we use the ratings of the first k items to construct a GF model and obtain the required prediction. The advantages of the paper are threefold: first, the proposed method introduces a new prediction model for CF, which, in turn, yields better performance of the model; second, it is able to alleviate the well-known sparsity problem as it requires less data in constructing the model; third, the model will become more effective when strong correlations exist among the data. Extensive experiments are conducted and the results are compared with several CF methods including item based, slope one, and matrix factorization by using two public data sets, namely, MovieLens and EachMovie. The experimental results demonstrate that the proposed algorithm exhibits improvements of over 20% in terms of the mean absolute error (MAE) and root mean square error (RMSE) when compared with the item based method. Moreover, it achieves comparative, or sometimes even better, performance when compared to the matrix factorization methods in terms of accuracy and F-measure metrics, even with small k.

1. Introduction

Recommender systems help users cope with the information overload experienced in a wide range of Web services and have been widely adopted in various applications, such as e-commerce (e.g., Amazon[1]), online video sharing (e.g., YouTube[2]), and online news aggregators (e.g., Digg[3]). Recommender systems have also been successfully developed for e-business and e-government applications [1–3]. They can be used to present the most attractive and relevant items to the user based on the individual user's characteristics. As one of the most promising recommender techniques [4], collaborative filtering (CF) predicts the potential interests of an active user by considering the opinions of users with similar preferences. As compared to other recommender techniques (e.g., content based methods [5]), CF technologies have the capability to recommend unanticipated items to users, which are not similar to those they have seen before; this could work well in domains where the attribute content of items is difficult to parse. Generally, the representative CF technique, namely, the memory based CF technique [6], has been widely used in many commercial systems due to its simplistic algorithm and reasonably accurate recommendations. It obtains the user's ratings on different items by explicitly asking the user or by implicitly observing the user's interactions with the systems; these ratings are stored into a table known as the user-item rating matrix. Then, the memory based CF methods use similarity measurement methods to filter out the users (or items) that are similar to the active user (or the target item) and calculate the prediction from

* Corresponding author at: Department of Automation, Tsinghua University, China. Tel.: +86 15801438286.
E-mail address: xiefeng10@gmail.com (F. Xie).

[1] www.amazon.com.
[2] www.youtube.com.
[3] www.digg.com.

http://dx.doi.org/10.1016/j.knosys.2014.04.011
0950-7051/© 2014 Elsevier B.V. All rights reserved.

the ratings of these neighbors. Memory based methods can be further classified into user based methods [7] or item based methods [8] depending on whether the process of defining neighbors follows the process of finding similar users or similar items.

Despite its widespread use, memory based CF techniques still suffer from several major problems, including the data sparsity problem [4,9], data correlation problem [10], and cold start problem [11,12]. The cold start problem can be regarded as a data sparsity problem. Hence, in this paper, we focus on the first two issues. In most recommender systems, each user rates only a small subset of the available items, and therefore, most of the entries in the rating matrix are empty. In such cases, determining similar users or items becomes a considerable challenge. Consequently, the similarity between two users or items cannot be calculated and the prediction accuracy becomes very low. Furthermore, the active users always tend to consume similar commodities, and the ratings for these items will be close, which indicates that there are strong correlations among the ratings. However, the existing similarity measurement methods, such as Cosine Distance and Pearson Correlation, suffer from such issues. Therefore, we cannot directly use similarities for rating prediction. To overcome these shortcomings, some researchers have developed algorithms that use models employing pure rating data to make predictions, such as clustering CF models [13,14], Bayesian belief nets (BNs) CF models [15,16], Markov decision process based (MDP-based) CF models [17], and latent semantic CF models [18]. However, some of these models are extremely complicated, require estimation of multiple parameters, and are sensitive to the statistical properties of data sets. In practice, many of these theoretical models have not been used in recommender systems due to the high costs involved.

In addition, dimensionality reduction techniques, such as singular value decomposition (SVD) [19], have been investigated to alleviate the data sparsity problem, where the unrepresentative users or items in the user-item rating matrix are removed to reduce the dimensionalities. However, useful information may be lost when certain users or items are discarded, and it is difficult to factor the matrix due to the high portion of missing values caused by its sparseness. Koren et al. [20] proposed a matrix factorization model, which is closely related to SVD. The model learns by only fitting the previously observed ratings. Its excellent performance enables it to be considered a state-of-the-art approach in rating prediction, but it also faces parameter estimation problems and high time complexities. Luo et al. [21,22] improved the matrix factorization based method by including incremental computations and applying an adaptive learning rate.

In this paper, we present novel approaches that aim at overcoming data sparsity limitations and benefiting from the data correlations existing among the ratings rather than eliminating them altogether. In particular, the proposed algorithm calculates the similarities between the items using the simplest method, namely, the Cosine Distance measurement method. It is worth noting that we do not directly use the exact value of the similarities, but rather rank the items according to their similarities. Then, a Grey Forecast (GF) model is constructed for rating prediction. This model has been successfully adopted for forecasting in several fields, such as finance [23], integrated circuit industry [24], the market for air travel [25], and underground pressure for working surface [26]. We compare the performances of the proposed algorithm with several other CF methods, including item based methods, slope one, and the state-of-the-art matrix factorization based method. Extensive experiments were conducted on two public data sets, namely, MovieLens and EachMovie. The results provide empirical evidence that the GF model can indeed cope effectively with data sparsity and correlation problems.

The remainder of this paper is organized as follows. Section 2 provides a detailed description of conventional user based CF (UCF) methods, item based CF (ICF) methods, the definition of existing problems, and our contributions. Section 3 presents the proposed GF model based algorithm in detail. Section 4 describes the experimental study, including experimental data sets, evaluation metrics, methodology, analysis of results, followed by a final section on conclusions and future work.

2. Related work

The CF technique is one of the most successful recommender techniques [27]: it can be classified into memory based CF techniques [7,8] such as similarity or neighborhood based CF algorithms, model based CF techniques such as clustering CF algorithms [13,14], and hybrid CF techniques such as personality diagnosis [28], hybrid fuzzy-based personalized recommender system [1], and hybrid semantic recommendation system [29]. As a representative memory based CF technique, the similarity based method represents one of the most successful approaches for recommendation. They have been extensively deployed into commercial systems and been comprehensively studied [4,30]. This class of algorithms can be further divided into user and item based methods. The former is based on the basic assumption that people who share similar past preferences tend to agree in their future preferences. Hence, for the target user, the potential interest for an object is predicted according to the ratings from the users who are similar to the target user. As opposed to the user based method, an item based method recommends the items that are similar to what the active user has consumed before. In a typical memory based CF scenario, there is a set of n users $U = \{u_1, u_2, \ldots, u_n\}$, a set of m items $I = \{i_1, i_2, \ldots, i_m\}$, and the $n \times m$ user-item rating matrix. The ratings can either be explicit indications, such as an integer number from 1 to 5 (The integer number represents the rating a user gives to the item. Usually, number 1 means that the user does not like the item, while number 5 indicates the user is very satisfied with the item.), or implicit indications, such as purchases or click-throughs [31]. For example, implicit user behaviors (Table 1a) can be converted into a user-item rating matrix R (Table 1b). When the kth user has purchased the lth item, $R(k,l)$ for the kth row and the lth column of the matrix is assigned to rating 1. If the kth user has not purchased the lth item yet, a *null* value is assigned to$R(k,l)$. Therefore, the recommendation problem is reduced to predicting the *null* entries (*Lily* is the active user for whom we want to make recommendations for in Table 1b). Generally, the procedure for this type of CF method consists of two steps: similarity measurement and rating prediction.

2.1. Similarity measurement

The critical step in memory based CF algorithms is the similarity computation between users or items [32–35]. In UCF methods, the similarity $s(u_x, u_y)$, between the users u_x, and u_y is determined

Table 1
An example of a user-item rating matrix.

User	Purchase	Not Purchase
(a)		
Alice	Milk, Bread, Cake	Beer
Lily	Milk, Bread	Cake, Beer
Lucy	Milk, Cake	Bread, Beer
Bob	Bread, Beer	Milk, Cake

	Bread	Beer	Cake	Milk
(b)				
Alice	1		1	1
Lily	1		?	1
Lucy			1	1
Bob	1	1		