Physica A 408 (2014) 72–84

Contents lists available at ScienceDirect

Physica A

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

A pattern mining approach to enhance the accuracy of collaborative filtering in sparse data domains

Mohsen Ramezani, Parham Moradi *, Fardin Akhlaghian

Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran

HIGHLIGHTS

- We seek to find a group of similar users based on their preferences for CF systems.
- The non-redundant subspaces are extracted to show the user interest patterns.
- A tree structure is created by mining the common patterns with the active user.
- Experiments conducted on Movielens and Jester datasets.
- The results show the better performance for the proposed than the other methods.

ARTICLE INFO

Article history:
Received 2 January 2014
Received in revised form 31 March 2014
Available online 15 April 2014

Keywords:
Recommender systems
Collaborative filtering
Clustering
Similarity measure
Pattern mining

ABSTRACT

Recommender systems seek to find the interesting items by filtering out the worthless items. Collaborative filtering is one of the most successful recommendation approaches. It typically associates a user with a group of like-minded users based on their preferences over all the items and recommends the items which are welcomed by others in the group to the user. But, many challenges like sparsity and computational issues still arise. In this paper, to overcome these challenges, we propose a novel method to find the neighbor users based on the users' interest patterns. The main idea is that users who are interested in the same set of items share similar interest patterns. Therefore, the non-redundant item subspaces are extracted to indicate the different patterns of interest. Then, a user's tree structure is created based on the patterns he has in common with the active user. Moreover, a novel recommendation method is presented to predict a new rating value for unseen items. Experimental results on the Movielens and the Jester datasets show that in most cases, the proposed method gains better results than already widely used methods.

1. Introduction

The Internet has become an important part of people's lives since its introduction in the 20th century. Also, developments in the Internet have made different types of data accessible online [1]. Therefore, people almost feel that they are surrounded by a number of distinct items to choose from. To overcome this problem, one can recommend items to people according to their preferences [2–6]. To this end, the Recommender System (RS) was introduced [6,7]. The RS is applied to different domains such as tourism [8], movies [3,9], music [10,11], and news [2,6].

Based on the modeling of the RSs, these systems can be categorized into content-based and collaborative-filtering approaches [4,5,10,12–14]. Content-based recommendation approaches analyze a set of descriptions of items previously

* Corresponding author. Tel.: +98 8716668513; fax: +98 8716668513.
E-mail addresses: m.ramezani@uok.ac.ir (M. Ramezani), pmoradi@gmail.com, p.moradi@uok.ac.ir (P. Moradi), f.akhlaghian@uok.ac.ir (F. Akhlaghian).

http://dx.doi.org/10.1016/j.physa.2014.04.002
0378-4371/

rated by a user and then create a model or profile of user interests based on the features of the items rated by the user [15]. On the other hand, a collaborative-filtering approach relies on the assumption that if some users have similar interesting items up to now, they will have similar interests in future [16,17]. Therefore, this approach takes advantage of information about users' habits to recommend potentially interesting items [18]. Thus, the main goal of the CF is to search for a set of users (which are called neighbor users) similar to a user who is being recommended to (which is called an active user).

One of the classic ways to find the neighbor users is using similarity. There are several well-known similarity measures in the literature including Pearson coefficient [19,20], cosine [20], Spearman [21] correlation and Kendall's Tau [3,22,23]. Most of these measures consider ratings of the common seen items between the active user and the other users. Therefore, they are not enough for finding the effective similar users especially for the sparse cold start user, who rates only a small number of items. Moreover, the performance of these measures will decrease when faced with very sparse domains with a small number of rated items common to users. Furthermore, these measures are computationally complex even with a relatively small number of items. Moreover, to overcome these problems, several similarity measures have been recently proposed based on fuzzy clustering and genetic algorithm [24,25] and statistical physics with the development of network science [26–29]. Most of these measures are based on bipartite networks [26,27]. Some of these methods are innovative and effective in improving not only accuracy but also diversity and novelty [28,30]. On the other hand, clustering methods can be used as another way to group the users into several clusters [31–33]. Therefore, the users in the cluster which the active user belongs to, are identified as his similar users [4,31,34,35]. The use of clustering methods to pre-determine groups of similar users leads to a significant increase in the performance of collaborative filtering systems. The generic clustering methods such as K-Means [4,31] and SOM [34] rely on nearest neighbor schemes to map a new user to the existing user groups. Therefore, the accuracy of these methods will decrease for sparse data domains.

In this paper, to overcome the high-dimensionality and sparsity issues, we propose a new approach to find neighbor users which is well suited to collaborating filtering systems with sparse data. In the proposed method, the non-redundant items subspaces are recognized, and the system seeks to find a list of users who share a common interest on each subspace. The users who belong to each subspace have the same interest patterns, like all corresponding items of the subspace. It is clear that a specific user can belong to several lists with different interest patterns. These lists are used to construct a new structure tree of the users. The active user will be placed on the root of the tree, and the other users are placed on the different levels of the tree. The level of each user on the tree is determined based on his interest pattern and the pattern similarity to the active user. According to this structure, a user with no seen items in common with the active user can be placed in the tree. This is due to the common interesting items with the others, that are used as references to connect the user to the active user. Therefore, the proposed method will be effective in sparse data domains. Finally, a new recommendation method is presented to use the tree structure to recommend items to the active user.

The rest of the paper is as follows. Section 2 reviews the collaborative filtering recommendation and methods of finding neighbor users. Then, in Section 3, details of this paper's research method are described. In Section 4, the dataset and experimental results are presented, and finally, the last section provides the conclusion.

2. Background

Researchers have studied broad dimensions of the CF such as improving the quality of recommendation. The recommendation relies on the interesting items of the neighbor users as recommender users [3,36,37]. In the following sections, we present a brief description of the CF approach.

2.1. Basics of collaborative filtering

Many successful RSs on the Web use the CF approach. The main goal of a CF approach is to suggest items to the active user based on his neighbor users. The CF approach can be classified into two categories including model-based and memory-based [20,23,37]. The model-based category includes methods that learn a model in an offline learning phase and then apply this model online to perform recommendation [38,39]. On the other hand, the memory-based category includes heuristic methods for recommendation. In these methods, recommendation is based on collecting available ratings of the items. In the memory-based CF approach, a set of m users $U = \{u_1, u_2, \ldots, u_m\}$ and a set of n items $I = \{i_1, i_2, \ldots, i_n\}$ are available. Each user imputes his preferences to seen items by a rating value in a defined scale. These preferences are represented as an $M \times N$ pattern matrix which is called the rating matrix.

The main idea of the CF approach is as follows: if two users have similar ratings on common items, then they have similar preferences. Therefore, in this approach, the recommendation to the active user will be performed based on the neighbor users [32]. Thus, the set of the neighbor users of the active user $u_a \in U$ which is shown as U_{u_a} is defined as follows:

$$U_{u_a} = \{u_i \in U | sim(u_a, u_i) \geq sim_{0\leq j\leq m \& j\neq i}(u_a, u_j)\} \tag{1}$$

where $sim(u_a, u_i)$ denotes the similarity of preferences between users u_a and u_j.

To recommend items to the active user, first of all, a numerical value for an unseen item will be calculated, and then a list of top N high-valued items to be recommended to the active user is prepared.