Neurocomputing 285 (2018) 94–103

Contents lists available at ScienceDirect

Neurocomputing

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

A dynamic trust based two-layer neighbor selection scheme towards online recommender systems☆

Ziyang Zhang[a], Yuhong Liu[b,*], Zhigang Jin[a,*], Rui Zhang[a,c]

[a] Department of Communications Engineering, Tianjin University, Tianjin, PR China
[b] Department of Computer Engineering, Santa Clara University, Santa Clara, USA
[c] Department of Communications Engineering, Tianjin Sino-German University of Applied Sciences, Tianjin, PR China

ARTICLE INFO

Article history:
Received 27 June 2017
Revised 15 November 2017
Accepted 30 December 2017
Available online 31 January 2018

Communicated by Prof. Yicong Zhou

Keywords:
Recommender system
Collaborative filtering
Neighbors
Availability
Trust
Temporal information
Forgetting factor

ABSTRACT

Collaborative filtering has become one of the most widely used methods for providing recommendations in various online environments. Its recommendation accuracy highly relies on the selection of appropriate neighbors for the target user/item. However, existing neighbor selection schemes have some inevitable inadequacies, such as neglecting users' capability of providing trustworthy recommendations, and ignoring users' preference changes. Such inadequacies may lead to drop of the recommendation accuracy, especially when recommender systems are facing the data sparseness issue caused by the dramatic increase of users and items. To improve the recommendation accuracy, we propose a novel two-layer neighbor selection scheme that takes users' capability and trustworthiness into account. In particular, the proposed scheme consists of two modules: (1) capability module that selects the first layer neighbors based on their capability of providing recommendations and (2) a trust module that further identifies the second layer neighbors based on their dynamic trustworthiness on recommendations. The performance of the proposed scheme is validated through experiments on real user datasets. Compared to three existing neighbor selection schemes, the proposed scheme consistently achieves the highest recommendation accuracy across data sets with different degrees of sparseness.

1. Introduction

The rapid growth of the Web provides huge potentials for individual users to contribute and access diverse information online every day. For example, YouTube users were uploading 300 h of new videos every minute in the year 2014, three times more than one year earlier [2]; 500 million tweets are generated on Twitter every day, bringing around 30% growth in volume every year [3]; and Wikipedia and its sister projects receive over 10 edits per second, more than 800 new articles per day from editors all over the world [2]. In such context, people are often overwhelmed by the vast amount of information and have to spend much more time and energy looking for their favorite items. The item can be a video clip on YouTube, a piece of news on social media, a post on Wikipedia, or a book on Amazon. Fortunately, online recommender systems are developed, which provide item recommendations to users by recording and analyzing users' behavior data explicitly (e.g. through ratings) or implicitly (e.g. through web browsing history). By utilizing such data, recommender systems cannot only help users in finding their desired items in a reasonable time, but also make it more convenient to show items to users [4].

Among different approaches implementing recommender systems, the collaborative filtering (CF) algorithm has been most widely adopted, due to the fact that it does not require domain related feature extractions, which makes the processing of unstructured data very convenient. Specifically, the CF algorithm can be divided into two categories: (1) item-similarity-based method, which recommends items that are similar to the items previously liked by the target user, and (2) user-similarity-based method, which provides recommendations based on the preferences of users who are similar to the target user in previous item preferences. Approaches in both of these two categories calculate similarities based on a user-item matrix that holds the previously observed user preferences on items. The calculated similarities are then used to select the most appropriate neighbors of the target user/item to provide accurate recommendations. Therefore, how to calculate user/item similarities, or one step further, how to select

☆ The preliminary version [1] of this paper has been accepted for publication in the proceedings of the 7th IEEE Annual Computing and Communication Workshop and Conference (IEEE CCWC 2017), Hotel Stratosphere, Las Vegas, USA, 9–11 January 2017.

* Corresponding authors.
E-mail addresses: zhangziyang@tju.edu.cn (Z. Zhang), yhliu@scu.edu (Y. Liu), zgjin@tju.edu.cn (Z. Jin), zhangrui0113@tju.edu.cn (R. Zhang).

https://doi.org/10.1016/j.neucom.2017.12.063

appropriate neighbors is the key of the CF algorithm that dramatically influences the recommendation accuracy.

The CF algorithm is facing new challenges in the big data era. One of these challenges is the data sparseness issue. As discussed above, recent online recommender systems often contain vast amount of items. And the number of items is still rapidly increasing. Compared to the total amount of items available in the system, the number of items rated by an individual user becomes very limited, leading to high uncertainty in estimating this user's preferences.Although the total number of users involved in recommender systems is also increasing, how to appropriately select capable and reliable (i.e. trustworthy) neighbors to predict the target user's preferences for accurate recommendations is very challenging, which leads to the so-called data sparseness problem.

Many existing neighbor selection mechanisms, however, have intrinsic inadequacies in handling the data sparseness issue. First, most of existing user similarity calculation mechanisms, such as Pearson correlation coefficient, cosine-based similarity and adjusted cosine-based similarity [5–7], calculate the similarity between a pair of users as a symmetric value, while ignoring these two users' asymmetric capabilities in recommending items to each other. Second, when comparing two neighbors, their total number of commonly rated items with the target user is often ignored, leading to a weird scenario where a neighbor sharing only one commonly rated item with the target user may yield a higher similarity score than the neighbor who shares 100 commonly rated items with the target user. Third, most current neighbor selection schemes do not consider the consistency of users' preferences on different items, not to mention the dynamic changes of users' preferences.

This paper aims to improve the recommendation accuracy of the user-similarity based CF, which, as discussed above, highly relies on precisely selecting neighbors for the target user, and resolve the problems caused by sparse data through the optimization of neighbor selection. To achieve this goal, we propose a novel two-layer neighbor selection scheme that selects capable and trustworthy neighbors based on two modules: (1) a capability module that selects the first layer neighbors by considering the asymmetric capabilities as well as the total number of commonly rated items between a pair of users, and (2) a dynamic trust module that performs the second layer neighbor selections by considering users' preference consistencies on different items. Experiments on real user datasets verify that the proposed scheme consistently achieves high recommendation accuracies across datasets with different sparseness degree.

The rest of the paper is organized as follows. Section 2 reviews the related work; Section 3 introduces the proposed scheme; Section 4 describes the experiments and results, followed by conclusion in Section 5.

2. Related work

2.1. Memory-based and model-based CF algorithms

User-similarity-based CF algorithms can be conducted through either memory-based methods or model-based methods. Memory-based CF methods calculate similarities directly based on the user-item matrix. Users with higher similarities to the target user are identified as the neighbors, whose preferences will then be utilized to predict preferences of the target user [8]. As a result, the recommendation accuracy of such methods highly relies on the precise neighbor selection. Although memory-based CF algorithms have been widely adopted by many recommender systems, they still have some significant limitations when handling data sparseness, cold start and scalability. On the other hand, model-based CF methods model user's behavior patterns based on the user-item matrix. Such patterns will then be used to predict this user's preferences. The most popular model-based CF approaches are based on clustering [9], co-clustering [10,11], matrix factorization [12–15], mixtures models [16,17], and transfer learning approaches [18,19]. Compared to memory-based CF methods, model-based CF methods can handle large-scale datasets well and provide faster predictions once the model has been established. However, the modeling process itself is usually time-consuming and often causes information loss, which may lead to the drop of recommendation accuracy.

2.2. User similarity calculations

Due to the wide adoption of the memory-based CF, many researchers are attracted to improve its accuracy. Consequently, a number of methods are proposed. Pearson correlation coefficient, cosine-base similarity and adjust cosine-base similarity, are the most popular methods to calculate user similarity, which serves as the foundation of selecting appropriate neighbors for recommendations. Some extension methods are proposed to further improve the accuracy of user similarity computations. In [20], the authors propose a method to detect and correct unreliable ratings to ensure the availability of the data set. In [21], a significance-based similarity measure is proposed to compute user similarities based on three types of significances. A new similarity function, proposed in [22], achieves higher recommendation accuracy by (1) assigning different weights to each individual item and (2) selecting different sets of neighbors for each specific user. This scheme, however, also significantly increases computational complexity. A new information entropy-driven user similarity measure model is proposed in [23] to measure the relative difference between ratings and a Manhattan distance-based model is then developed to address the fat tail problem by estimating the alternative active user average rating, which improves the accuracy of the similarity computation. In [24], a multi-level collaborative filtering method is proposed to assign a higher similarity score to a pair of users if their Pearson correlation coefficient or the number of commonly rated items exceeds a certain threshold.

These methods, however, calculate user similarities without considering the size of their commonly rated items and their asymmetric capabilities in recommending items to each other, not to mention their recommendation trustworthiness. More important, the data sparseness issue makes such inadequacies even worse, which may then cause significant drop of recommendation accuracy.

2.3. User trustworthiness evaluations

Another trend is to improve recommendation accuracy by introducing trust values among users in collaborative filtering. For instance, the trust values can be computed based on the transitivity rules for similarities among users [25]. Some researchers propose trust models based on users' social network trust relationships or the propagation effect of online word-of-mouth social networks [26–29]. In [30], the authors propose an innovative Trust-Semantic Fusion (TSF)-based recommendation approach within the CF framework which incorporates additional information from the users' social trust network and the items' semantic domain knowledge in order to deal with the data sparsity,user and item cold-start problems.However, these trust models may not be applicable in many recommender systems due to the lack of information about users' social relationship or behavior. Some researchers propose trust models based on subjective logics or belief theory such as [31] and [32]. In [33], the model calculates direct and indirect trust among users by considering one-hop or multiple-hop distances among items. The authors in [34] propose