Knowledge-Based Systems 74 (2015) 14–27

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Leveraging multiviews of trust and similarity to enhance clustering-based recommender systems

Guibing Guo [a,*], Jie Zhang [a], Neil Yorke-Smith [b,c]

[a] School of Computer Engineering, Nanyang Technological University, Singapore
[b] American University of Beirut, Beirut, Lebanon
[c] University of Cambridge, Cambridge, UK

A R T I C L E I N F O

Article history:
Received 15 April 2014
Received in revised form 25 September 2014
Accepted 28 October 2014
Available online 4 November 2014

Keywords:
Recommender systems
Multiview clustering
Collaborative filtering
Cold start
Similarity
Trust

A B S T R A C T

Although demonstrated to be efficient and scalable to large-scale data sets, clustering-based recommender systems suffer from relatively low accuracy and coverage. To address these issues, we develop a multiview clustering method through which users are iteratively clustered from the views of both rating patterns and social trust relationships. To accommodate users who appear in two different clusters simultaneously, we employ a support vector regression model to determine a prediction for a given item, based on user-, item- and prediction-related features. To accommodate (cold) users who cannot be clustered due to insufficient data, we propose a probabilistic method to derive a prediction from the views of both ratings and trust relationships. Experimental results on three real-world data sets demonstrate that our approach can effectively improve both the accuracy and coverage of recommendations as well as in the cold start situation, moving clustering-based recommender systems closer towards practical use.

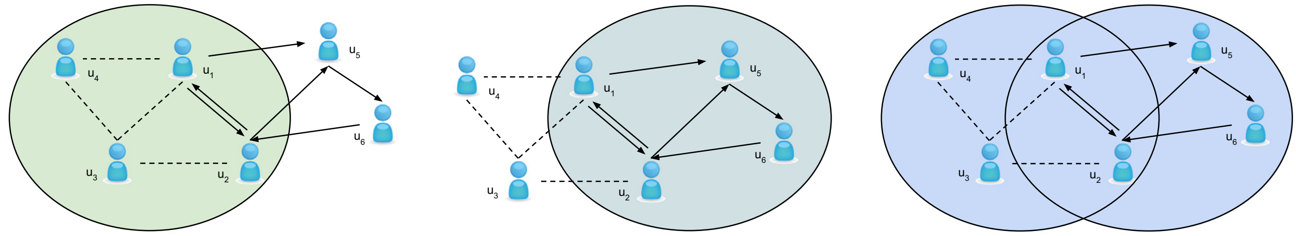
1. Introduction

Collaborative filtering (CF) [1] is a widely-exploited technique in recommender systems to provide users with items that well suit their preferences. The basic idea is that a prediction for a given item can be generated by aggregating the opinions (i.e., ratings) of like-minded users, i.e., the users with similar interest. CF has been extensively studied over decades, and many approaches [1,15,10] have been proposed in the literature. These approaches can be classified into two categories: *memory-based* and *model-based* methods. Memory-based methods [1,10] aim to find similar users (called nearest neighbors) by searching the entire user space, that is, the similarity between each user and the active user (who desires recommendations) needs to be computed using some similarity measure such as the Pearson correlation coefficient [1]. Although CF gained popularity due to its simplicity, the time-consuming procedure of searching for similar users poses a big challenge when facing large-scale data sets, which is a typical characteristic of Web 2.0. Other issues of memory-based methods include *data sparsity* and *cold start* problems [10] since the computed similarity may not be reliable due to insufficient ratings.

In contrast, model-based methods (e.g., [30,31]) can address these issues by training a prediction model offline using all the rating data (both relevant and irrelevant to the active user) rather than only based on the overlapping ratings between users. Among the various approaches, matrix factorization [15] is arguably the most popular model-based technique. It factorizes the user-item rating matrix into small ranks of user-feature and item-feature matrices. Then, the prediction is generated by the inner product of a user's feature vector and an item's feature vector. Generally, these methods can well adapt to large-scale data sets and cope with the data sparsity problem. However, a critical drawback is that the newly-issued ratings cannot be quickly involved for predictions: retraining a model is usually time-consuming and costly. This is a drawback because millions of new ratings may be reported every few hours in real applications. In addition, a lesson learned from the Netflix competition shows that the best method is a combination of hundreds of different recommendation algorithms, and none of a single algorithm can achieve the best performance over the others [2]. In this regard, it is still meaningful to develop other kinds of model-based methods. In this work, we focus on the development of a clustering-based approach based on both user ratings and trust information. In this article,

* Corresponding author.
E-mail addresses: gguo1@ntu.edu.sg (G. Guo), zhangj@ntu.edu.sg (J. Zhang), nysmith@aub.edu.lb (N. Yorke-Smith).

http://dx.doi.org/10.1016/j.knosys.2014.10.016
0950-7051/



(a) Similarity Clustering (b) Trust Clustering (c) Multiview Clustering

Fig. 1. The user clustering approaches based on similarity and trust information. Circles denote the formed user groups (clusters), and dashed lines indicate the similarity between two users, while solid lines with arrows represents user trust (social trust is directional): (a) clustering users by similarity; (b) clustering users by trust and (c) clustering users by both similarity and trust (i.e., multiviews) where more users can be grouped in one cluster.

we adopt the definition of trust as "one's belief towards the ability of others in providing valuable ratings" given by Guo [9]. By definition, trust has a much stronger correlation with user preferences than other general social connections (e.g., friendship).

Clustering-based approaches [22] offer an alternative to model-based methods. Instead of decomposing the rating matrix into matrices with small ranks, these approaches reduce the search space by clustering similar users or items together. For example, as illustrated in Fig. 1, users can be clustered by either similarity (a) or trust (b) such that the search space for nearest neighbors can be effectively narrowed down (to smaller clusters). In this way, new ratings of clustered users or items can be timely made use of to make predictions. However, clustering-based methods have not been widely exploited in recommender systems. Although demonstrated to be efficient and scalable to large-scale data sets, they are recognized to suffer from relatively low accuracy and coverage [22,28,3]. This is mainly because similar users can only be selected from the fixed size of cluster members, and in general a fewer number of similar users can be identified (than searching the whole space). In addition, the recommendation performance is also sensitive to the quality of the clustering methods. As a consequence, relatively low accuracy and coverage are observed, and these issues severely hinder the practical use of clustering-based approaches in recommender systems. To sum up, as dimension reduction models, clustering-based approaches retain the advantages of low computational cost (for searching candidate users) over memory-based approaches, and are capable of integrating newly-issued ratings for up-to-date recommendations relative to matrix factorization-based models. However, clustering-based approaches are less exploited in the literature, and suffer from relatively low accuracy and coverage.

To cope with the aforementioned issues, we develop a multiview clustering method by making use of both the view of rating patterns and the view of social trust relationships. Specifically, users are iteratively clustered from the two views using a classic partitional clustering algorithm, and clusters generated from different views are then combined together (e.g., as illustrated in Fig. 1(c)). The underlying assumption is that similarity and trust provide different views of user correlations.

Multiview-based clustering methods have not been well exploited in recommender systems and most previous works only function in a single view, namely, the user similarity. The proposed multiview clustering method has several advantages relative to single view clustering methods. First, since the clusters of users from different views will be integrated together, there are more candidate users from which similar users can be identified. Hence intuitively, both the recommendation accuracy and coverage will be improved, as we will demonstrate. Second, to accommodate users who appear in two different clusters simultaneously, we employ a support vector regression (SVR) model [7] to determine a proper prediction for a given item based on user-, item- and prediction-related features, described in Section 4.2. By doing so, the recommendation performance can be further improved. Third, to accommodate (cold) users[1] who cannot be clustered due to insufficient data, we propose a probabilistic method in Section 5 to derive a prediction from the viewpoints of both ratings and trust relationships. A series of experiments are conducted in Section 6 based on three real-world data sets, namely Flixster, FilmTrust and Epinions. The results confirm that our approach can effectively improve both the accuracy and coverage in comparison with the other counterparts, and function significantly better in handling cold users than trivial strategies (such as the average of all cluster predictions) used in previous approaches.

In summary, the main contributions of this article are:

1. We propose a multiview clustering method to cluster users from both the views of user similarity and trust. To our best knowledge, we are the first to propose a multiview clustering method based on both kinds of information.
2. We propose a support vector regression (SVR) model to handle the situation where two predictions are generated from two clusters. A number of user-, item-, prediction-related features are identified for this purpose.
3. We propose a probabilistic method to resolve the cold start problem that has not been addressed previously. Both ratings and trust information are adopted in the method.
4. We conduct a series of experiments on three real-world data sets to verify the effectiveness of the proposed multiview clustering method in comparison with other methods.

Our work takes the first step to cluster users from multiple different views of user preference rather than a single view, and verifies the ability to mitigate the issues of low accuracy and coverage using real-world data sets, moving clustering-based recommender systems closer towards practical use.

The rest of this article is organized as follows. Section 2 gives an overview of the related research on trust-based and clustering-based recommender systems. Then, our approach is elaborated in the threefold: formulating the multiview clustering algorithm in Section 3, generating predictions by support vector regression in Section 4, and handling the cold start problem in Section 5. After that, experiments based on three real-world data sets are conducted in Section 6. Finally, Section 7 concludes our present work and outlines the future research.

[1] The cold-start or cold users refer to those who rated only zero or a small number of items, e.g., less than 5 items.