Information Sciences 432 (2018) 210–230

Contents lists available at ScienceDirect

Information Sciences

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

# A novel 2D-Graph clustering method based on trust and similarity measures to enhance accuracy and coverage in recommender systems

Leily Sheugh, Sasan H. Alizadeh*

*Faculty of Computer and Information Technology Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran*

A R T I C L E  I N F O

*Article history:*
Received 17 August 2016
Revised 2 December 2017
Accepted 4 December 2017
Available online 11 December 2017

*Keyword:*
Recommender systems
2D-Graph clustering
Coverage
Accuracy
Cold start

A B S T R A C T

Various clustering approaches have been widely adopted to improve the accuracy and scalability of collaborative filtering-based recommender systems as the major objectives. Recent research has experimentally disclosed the realization of such objectives at the expense of decreasing the recommendations coverage. In this paper, we prove that the rate of coverage shows a monotonic decreasing trend as the number of clusters increases. This proof establishes a foundation for the proposal of a 2D-Graph clustering method for partitioning a novel 2D-Graph in compliance with the minimum-acceptable-rate-of-coverage criterion. To reduce the negative impact of clustering on the coverage in the proposed method, in addition to the information of trust, Pearson similarity is incorporated to construct double-weighted connections between users in the 2D-Graph. In each cluster, to address the sparsity problem, the weights corresponding to a convex combination of Pearson, trust and Jaccard similarities are determined so that the precision is optimized. The experimental results on two real-world datasets illustrate that the proposed method outperforms the state-of-the-art methods for Cold Start users in terms of accuracy and coverage.



## 1. Introduction

With the increasing utilization of World Wide Web services, the number of users experiencing information overload problems has been notably on the rise. The prime movers of recommender systems are facilitating users to fulfill their informational requirements more precisely in an automatic manner. These systems have been recently implemented in a variety of application domains, including but not limited to movies, music, news and books [36]. Among recommender system methods, Collaborative Filtering (CF) methods are the most frequently employed. These methods involve the recommendation of filtered information to target users considering the common tastes of other like-minded users [1]. The main advantage of CF methods is that they merely rely on ratings information to calculate the similarity of users and do not require any understanding of items [1].

CF-based systems are divided into model-based and memory-based subcategories. The model-based CF systems [9], in contrary to memory-based ones, require a learning phase to estimate the parameters of the model according to the training data. The process of recommendation is easy and straightforward when all of the model parameters are determined

* Corresponding author
*E-mail address:* sasan.h.alizadeh@qiau.ac.ir (S.H. Alizadeh).

https://doi.org/10.1016/j.ins.2017.12.007
0020-0255/

through training. However, memory-based approaches [2] require more complex calculations during this process. From this viewpoint, model-based approaches are more suitable for real-time large-scale online recommender systems.

Despite the advantages of the CF methods in recommender systems, these methods still suffer from several problems, including sparsity, coverage and Cold Start problems [22,25,42]. To illustrate such problems better, definitions are provided as follows. Sparsity occurs when only a small percentage of all possible ratings are available in a recommender system, which is usually caused by a lack of user tendency to rate a large number of items in a typical recommender system [22]. With this problem, the number of items that are rated by both users, called Co-rated items, decreases noticeably, leading to a decline in the reliability of regarded users' similarity measures. Consequently, poor recommendations are generated. Such sparsity provides fertile ground for the coverage problem whereby the ratings of a small percentage of items can be predicted so that they can be recommended to target users [22]. The third problem, Cold Start, is caused by the constitutional concept of new users itself. This problem shows itself when new users register and the recommender system is yet to obtain their ratings, i.e., there is not yet enough information on their ratings to empower the system to predict their items of potential interest. As the Cold Start encounter provides a new user with irrelevant recommendations, the user, especially an impatient one, might lose motivation to continue the thread of collaboration with the system and quickly abandon it [26]. Alternately, this challenge is experienced by new items because they are starving for ratings and are therefore less likely to be included in the recommendation list that users receive [25,26].

To alleviate the aforementioned problems, various clustering algorithms have been incorporated in different CF methods. The aim of a clustering method is to divide the set of all users/items into some coherent partitions. Via clustering methods, similar users/items are put together in the same group. The basic idea is that similar users have the same tastes [5,18,30]. In addition to scalability, proper clustering approaches can increase the accuracy of predictions as well. However, the main problem here is how to identify efficient clusters. In other words, poor clustering reduces not only the rate of coverage but also the prediction accuracy [11]; in particular, if the number of users in a specified cluster decreases too significantly, the information may not be deemed sufficient to provide accurate predictions [5,11,18,30].

Similarity as a measure is believed to be the core component of every clustering approach [5]. In CF-based recommender systems, the similarity of users is calculated according to just ratings information with measures such as the Pearson Correlation Coefficient (PCC) [8], Cosine [8], and Jaccard [15] or their hybrid variants [7], among others. In addition to ratings, recent studies have utilized different sources of supplementary information in some clustering approaches, including merge [15], multi-view [16], and Co-clustering [43], to address the sparsity problem and compensate for the lack of information. This includes information regarding trust [15,17], distrust [13] and friendship in a social network of users, users' demographic information [34], and the semantic [49] and ontological [28] information of items. There are primarily two major approaches developed to employ this supplementary information. In the first approach, different sources of information are combined together into a unified similarity measure for final clustering. In the second one, each different source of information constitutes a basis for a distinct similarity measure to produce a variety of clustering solutions accordingly, and a final clustering eventually emerges from the combination of so-called clustering solutions.

However, both approaches fuse information into a single solid entity by applying different arithmetic/set theory operations. The major drawback here is that, after fusion, the original information ingredients can no longer be distinguished from each other and are treated almost equally. In contrast with widely practiced ongoing approaches, in this paper, we devise a novel 2D-Graph in a manner such that all adopted similarity measures are preserved through the process of clustering, as well as recommendation, to improve the coverage while maintaining the precision as much as possible. The experimental results of the previous literature show that coverage as a measure is often compromised, unlike precision, which is occasionally degraded in clustering approaches. Our contribution, therefore, is to prove such an observation explicitly. This proof lays the foundation for the generation of various feasible clustering solutions in accordance with the Minimum Acceptable Rate of Coverage (MARC) criterion so that an optimal solution is well met by an exhaustive search method.

The remainder of this paper is organized as follows. In Section 2, we give an introduction to related works, whereas Section 3 elaborates the overall scheme of the proposed method. In Section 4, 2D-Graph construction will be presented. The clustering phase is explained in Sections 5, and 6 is devoted to the prediction. Experiments and evaluations are conducted in Section 7, followed by the results and analysis of Section 8 to investigate the efficiency of the proposed method. Finally, Section 9 addresses the conclusion and future works.

## 2. Related works

To review the literature on recommender system clustering methods, we need to bear in mind that precision and accuracy are the primary concerns of every recommender system; however, neglecting the complexity of recommendation algorithms may raise critical problems. The complexity itself brings about a significant response-time deficiency when the number of users or items greatly increases [11]. This raises an important challenge in online recommender systems because users are very impatient and can tolerate only a few seconds for the response [22,25,43]. Thus, the scope of complex recommendation algorithms is confined to various off-line applications, which are of less importance for many real-time e-commerce websites [36]. To mitigate this problem, clustering approaches are typically exploited to divide the original data set into several individual partitions [6]. Clustering-based systems are able to provide responses in a timely fashion by decreasing the size of the original data set into more manageable partitions. Indeed, clustering increases the scalability of recommender systems [46] yet preserves the accuracy and precision [4,12,24,31,32,47]. However, in regard to precision, the