Knowledge-Based Systems 83 (2015) 51–57

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Exploiting matrix factorization to asymmetric user similarities in recommendation systems

Parivash Pirasteh [a], Dosam Hwang [a,1], Jason J. Jung [b,*,1]

[a] Department of Computer Engineering, Yeungnam University, Republic of Korea
[b] Department of Computer Engineering, Chung-Ang University, Republic of Korea

ARTICLE INFO

Article history:
Received 15 September 2014
Received in revised form 3 February 2015
Accepted 6 March 2015
Available online 14 March 2015

Keywords:
Collaborative filtering
Matrix factorization
Recommender systems
User similarity
Asymmetry

ABSTRACT

Although collaborative filtering is widely applied in recommendation systems, it still suffers from several major limitations, including data sparsity and scalability. Sparse data affects the quality of the user similarity measurement and consequently the quality of the recommender system. In this paper, we propose a novel user similarity measure aimed at providing a valid similarity measurement between users with very few ratings. The contributions of this paper are twofold: First, we suggest an asymmetric user similarity method to distinguish between the impact that the user has on his neighbor and the impact that the user receives from his neighbor. Second, we apply matrix factorization to the user similarity matrix in order to discover the similarities between users who have rated different items. Experimental results show that our method performs better than commonly used approaches, especially under cold-start condition.

1. Introduction

In the age of data overload, the enormous amount of information makes users to spend more time and energy to select an item. The item can be a book for an Amazon user or a place to visit for a tourist or a course to pass for a learner. In order to overcome information overload problem, recommender systems aid users to find their desired content in a reasonable time, by analyzing their behavior data related to user activity [20].

Content-based filtering is based on the hypothesis that users are able to formulate queries that express their interests or information needs in terms of intrinsic features of the desired items. The user profile is a structured representation of user interests, extracted to recommend new items. The recommendation process basically consists of matching the characteristics of the user profile against the characteristics of a content object [14,22].

Collaborative filtering (CF) is an alternative to content-based techniques. Instead of recommending new items that are similar to items the user has liked in the past, this method recommends items that similar users have liked [2,3]. CF techniques are more often implemented than content filtering and often result in better predictive performance. The main reason is that they are independent of data used by content filtering, which are invasive and time consuming to collect. Generally, CF outperforms content-based techniques except in special cases, such as when user ratings of a certain item are highly varied (i.e. controversial items) or for cold-start situations, where the users did not provide enough ratings to compute similarity with other users [30].

A key factor in the quality of the recommendations obtained in a CF-based recommender system lies in its capacity to determine which users have the most in common with a given user. Traditional methods of similarity suffer from three drawbacks. First, usually all items are treated equally, regardless of the various amounts of information that can be extracted from different items. This is addressed by [16], where weighting schema are defined in order to capture the importance of an item and giving distinct items a higher coefficient in assigning a correlation.

The second problem occurs under cold-start condition where recommender systems need to predict preferences for a user with a small number of ratings. Pure CF methods which rely on calculating similarities between users based on their co-rated items, fail to provide similarity between users who do not have any co-rated items. To illustrate this limitation, consider the example of Fig. 1. While John and Alex share the same neighbors, the similarity coefficient between them cannot be computed based on traditional methods because they do not have any common item. However, content information can help bridge the gap between existing items and new items by inferring similarities among them [32]. Several hybrid methods (combinations of content-based and CF techniques) have been proposed to improve the performance of

* Corresponding author. Tel.: +82 2 820 5316; fax: +82 2 820 5301.
E-mail address: j2jung@gmail.com (J.J. Jung).
[1] These authors contributed equally to this work as the first author.

http://dx.doi.org/10.1016/j.knosys.2015.03.006
0950-7051/

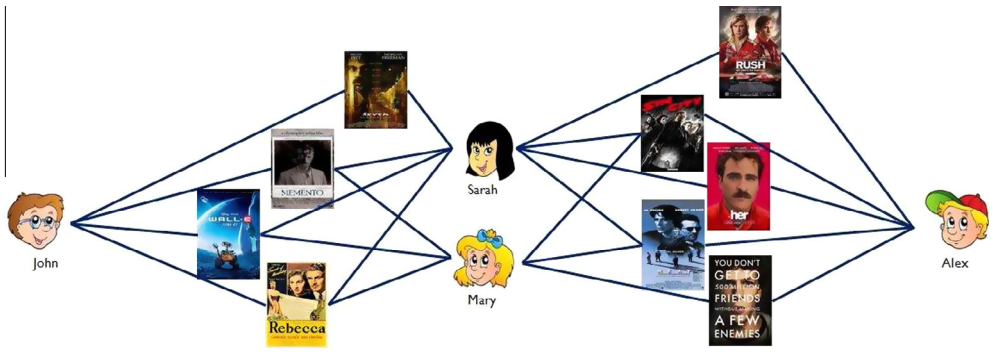

Fig. 1. Although John and Alex do not have a direct overlap between their rated items, the similarity ratio between them and their joint neighbors can express their agreement on and disagreement over on different movies.

recommender systems when it comes to cold-start prediction. Although content-based methods are not very sensitive to data sparsity, they suffer from other limitations. For example, the content information is hard to automatically extracted out, and is not always available for privacy reasons.

The proposed method in the paper allows generating high quality recommendations even on sparse dataset. It also helps to learn dependencies between all the pairs of users by projecting the existing similarity values in a latent space model. In this paper, first, we propose an asymmetric user similarity measurement based on mean square difference and cosine similarity. In this way, we incorporate into the proposed model the ability to distinguish between two users with a different proportion of common ratings. Second, we apply matrix factorization to characterize the user's interests by a vector of factors derived from the proposed similarity measure in order to predict similarity among users with few immediate neighbors.

The rest of this paper is organized as follows. In Section 2, we present some of the most relevant works on the topic and describe advantages and disadvantages. Section 3 presents our proposed method for user similarity measurement. The user similarity matrix factorization framework is presented in Section 4. The results of experimental analysis are presented in Section 5, followed by the conclusion and future directions in Section 6.

2. Related works

The most commonly used measurement techniques for similarities between users are the Pearson correlation coefficient (PCC) [29] and cosine similarity (COS) algorithm [5]. PCC defines user similarity as the linear correlation between them. It is well recognized that PCC and COS only consider the direction of rating vectors and ignore the length [26].

COS assumes that the rating of each user is a point in a vector space and then evaluates the cosine angle between the two points. It considers the common rating vectors $X = \{x_1, x_2, x_3, \ldots, x_n\}$ and $Y = \{y_1, y_2, y_3, \ldots, y_n\}$ of users X and Y, represented by a dot product divided by magnitude. COS has frequently been used for performance comparisons in CF [19]. The constrained Pearson correlation coefficient is a slightly modified version of the Pearson correlation that increases the correlation only when both users have rated an item positively or negatively [33].

Although the two popular similarity measures, PCC and COS have proven to be successful in many studies, they have some drawbacks. The main limitation of these approaches is their inability to consider the size of the set of common items between users. To overcome this limitation, a combination of Jaccard similarity with Pearson correlation coefficient has been proposed [34].

Jaccard similarity does not suffer from this limitation because it measures the overlap that two vectors share with their attributes. On the other hand, such a measure does not take into account the difference in ratings between the vectors. In this case, if two users watch the same movies but have completely opposite opinions of them, the users are considered to be similar regardless of their differing opinions [6].

Like PCC, the Spearman rank is correlation coefficient based and computes a measure of correlation between ranks instead of actual preference scores [36]. Shardanand et al. proposed a measure based on mean square difference (MSD), which evaluates the similarity between two users as the inverse of the average squared difference between the ratings given by those users on the same item [33].

Other experts proposed new similarity measures to substitute for the traditional one. Konstan et al. suggested a concordance-based measure that helps users with privacy concerns and those who do not want to reveal their ratings history [17]. On the data sparsity problem, one noteworthy study presented a similarity method using proximity-impact-popularity (PIP). Ahn discussed the problems of widely used similarity methods resulting in decreased prediction accuracy and proposed PIP similarity [1]. Using genre information to circumvent the cold-start problem for new items has been used in [28]. Hence, when a new item enters the system, genre correlation aids in finding similar items.

Luo et al. [23] divide user similarity into two parts: local user similarity and global user similarity. Local similarity is determined based on surprisal-based vector similarity (SVS). Global similarity measures the similarity between two users by further considering the extent to which their neighbors are locally similar (using the local similarity). Therefore, two users become more similar if they can be connected through a series of locally similar neighbors. From this point of view, the rest of the mentioned methods are categorized as global similarity measures.

Moreover, Jamali et al. introduced a similarity measure using the Markov-chain model of a random walk. This approach can weaken the similarity of small common items among users [15]. In graph-based approaches, the data are represented in the form of a graph, where nodes are users, items or both, and edges encode the interactions or similarities among the users and items. The transitive associations captured by graph-based methods can be used to recommend items. Fouss et al. suggested applying Euclidean commute time distance, which is one of the random-walk-based methods to compute similarities between nodes. The SimRank algorithm is another graph-based model that is utilized to compute the similarity [12]. For example [11] applied the SimRank-based algorithm to construct a general recommendation system, while [7] proposed combining SimRank with clustering to match users in online dating networks.