Knowledge-Based Systems 56 (2014) 273–280

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

An entropy-based neighbor selection approach for collaborative filtering

Cihan Kaleli [*]

Department of Computer Engineering, Anadolu University, 26555 Eskisehir, Turkey

ARTICLE INFO

Article history:
Received 23 April 2013
Received in revised form 29 November 2013
Accepted 29 November 2013
Available online 12 December 2013

Keywords:
Collaborative filtering
Nearest neighbors
Entropy
Optimization
0–1 Knapsack problem

ABSTRACT

Collaborative filtering is an emerging technology to deal with information overload problem guiding customers by offering recommendations on products of possible interest. Forming neighborhood of a user/item is the crucial part of the recommendation process. Traditional collaborative filtering algorithms solely utilize entity similarities in order to form neighborhoods. In this paper, we introduce a novel entropy-based neighbor selection approach which focuses on measuring uncertainty of entity vectors. Such uncertainty can be interpreted as how a user perceives rating domain to distinguish her tastes or diversification of items' rating distributions. The proposed method takes similarities into account along with such uncertainty values and it solves the optimization problem of gathering the most similar entities with minimum entropy difference within a neighborhood. Described optimization problem can be considered as combinatorial optimization and it is similar to 0–1 knapsack problem. We perform benchmark data sets-based experiments in order to compare our method's accuracy with the conventional user- and item-based collaborative filtering algorithms. We also investigate integration of our method with some of previously introduced studies. Empirical outcomes substantiate that the proposed method significantly improves recommendation accuracy of traditional collaborative filtering algorithms and it is possible to combine the entropy-based method with other compatible works introducing new similarity measures or novel neighbor selection methodologies.

1. Introduction

In parallel with the current innovations in the Internet technologies, people face with much more information to be needed handling. Recommender systems have been developed for overcoming such problem for online applications, e.g., e-commerce [1], e-learning [17], and social networks [6]. Collaborative filtering (CF) is a well known recommendation technique, which has the capability of extracting useful information from a huge amount of user data. Inside a CF recommendation process, there is a user-item matrix which holds a great deal of users' preferences on various items and the goal is to produce personalized predictions for items which are not yet rated by users.

CF has a basic assumption that the customers agree in the past, tend to agree in the future, as well [8]. CF has also another principle that entities of a user-item matrix do not have the same utility for a particular user's recommendation process [15]. In order to form neighborhoods, such systems utilize degree of similarities among those entities, i.e., users or items. As a result, correlations between entities are one of the decisive factors of predicting a user's future trend. Thus, the central aspect of neighborhood-based CF algorithms is to determine the similarity between entities either considering commonly rated items by users or focusing on users having rated both corresponding items. Those like-minded customers or similar items are referred to as nearest neighbors (NN) and CF estimates predictions for a target item of a customer by employing only such neighborhood instead of whole existing user preferences. Therefore, accuracy of predictions excessively depends on selecting NN of entities successfully.

Broadly speaking, *k*-NN-based CF algorithms form neighborhood of an entity either by utilizing correlation based similarity function, i.e., Pearson's correlation coefficient (PCC) [15] or adjusted cosine-based similarity (ACS) measure [28]. Although these two similarity measures are successful at obtaining entity correlations, they can have challenges on working with data collected for recommendation purposes. Since user preferences are very sparse, similarities between entities might be computed from only a small number of common ratings and might end up with unreliable neighborhoods. To date, researchers introduce several studies in order to improve NN selection capability of CF algorithms and they enhance accuracy of predictions either by proposing new similarity measures [2,30,9,12] or evolving new perspectives for NN formation [4,5,20,13,10].

Each individual has her own interpretations about a rating domain in order to express her preferences. For example, in a 5-star rating scale, a customer might express her preferences in a binary manner and she might always give 1 for items either she

[*] Tel.: +90 222 321 35 50; fax: +90 222 323 95 01.
E-mail address: ckaleli@anadolu.edu.tr

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.11.020

strictly dislikes or almost disfavors, conversely, she might rate her both favorite and likeable products by 5. So, it would not be wrong to infer that this user does not regard all members of a rating domain to express her tastes on products. On the other hand, another customer might take all rating values into account in order to differentiate her preferences according to her tastes. Therefore, she might rate her disliked items as 1 or 2 according to her degree of dislike, similarly, 4 or 5 rating values for her favorite and likeable products, and 3 to express an average liking. Thus, it can be concluded that this user is motivated on utilizing each member of the rating domain in order to differentiate degree of her admiration. In addition, a user might be someone who can be either easily satisfied or an exacting person. Herewith, personal preferences of these type of customers are affected by their personal characteristic and it is expected that while latter type of user employs 4 and 5 rating evincing her thoughts on products, former one mostly utilizes either 1 or 2. Eventually, although all customers employ members of the same rating domain to express their preferences, their rating usage trends are likely to differentiate. If a user prefers to utilize every member of a rating domain, since all rating values take place in her preference vector, her rating vector's uncertainty increases. Consequently, we can assign a degree of uncertainty (DU) of a user which can be estimated by computing entropy [29] of her preference vector. In CF systems, if any two users have close DU values, then it can be inferred that these two users utilize the rating domain in a similar way. If their correlation-based similarity is also high, intrinsically, it can be concluded that these two users have strong neighborhood relationship. Note that, the same scenario is also valid for items. Likewise, an item's DU value can be calculated by estimating entropy of its ratings vector.

Since user-item matrix in recommender systems are extremely sparse, it is very important to use all possible information for the sake of improving prediction quality. In this paper, we aim to combine DU and similarity information in order to form a more qualified NN of an entity. In our proposed method, during formation of a neighborhood we aim to minimize differences of DU values between entities while maximizing similarities. As a result, we face with an optimization challenge similar to 0–1 knapsack problem [34] which is one of the most popular combinatorial optimization problems. In 0–1 knapsack problem, there is a set of items having different weights and profits, and the aim is determining the items to be included in a collection with minimum total of weights, which is less than or equal to a given criteria, and maximum total profit. In our optimization problem, we consider size of knapsack associated with DU values, so that, while forming NN of an entity, we try to fill neighbor container with the entities having minimum DU difference and maximum similarity. Afterwards, we employ all users/items in the container as NN of the active entity in recommendation process. We evaluate success of our method on two conventional CF algorithms, i.e., user- and item-based methods, using two benchmark data sets. We also investigate how combination of our method with other works introducing new similarity measures or novel neighbor selection methodologies improves accuracy of recommendations. Experimental results substantiate that the proposed method significantly improve recommendation accuracy of CF algorithms and it is possible to integrate it with prior related solutions in order to boost their accuracy.

Major contributions of the work are, as follows:

1. CF is considered as an optimization problem.
2. A new neighborhood selection approach is proposed.
3. Accuracy of traditional user- and item-based CF algorithms are significantly improved.
4. Proposed entropy-based optimization approach can be combined with other compatible related works.

The rest of the paper is organized, as follows. We briefly describe previous related research in Section 2. We then explain two traditional CF algorithms, entropy concept in information theory, and 0–1 knapsack problem in the following section. We deeply describe our proposed method in Section 4. After demonstrating and discussing experimental results in Section 5. Finally, we conclude the paper and give future research directions in Section 6.

2. Related work

With the increasing attention to k-NN-based CF at the second half of nineties, researchers introduce several studies on forming more appropriate NN in recommendation process. The first comprehensive study about neighbor selection in CF is performed by Herlocker et al. [15]. The authors analyze designing neighborhood-based prediction systems in detail and according to their results, PCC is an effective choice to compute correlations between users and employing the most similar k users is the best approach to form neighborhood. After this fundamental study, researchers have continued working on improving NN selection in CF process in order to boost accuracy. Kim and Yang [19] present an effective threshold-based neighborhood selection method for CF. Their method utilizes substitution of neighbors for a customer having unusual preferences. In another work, Baltrunas and Ricci [4] consider dynamic changes in users' profiles and propose an adaptive neighborhood selection scheme relying on users' tastes for target item. According to their experimental results, the proposed approach improves accuracy of predictions. Liang et al. [21] discuss calculating user correlations from a subset of items instead of the whole set. Koren [20] brings a different perspective for determining appropriate NN of a customer. In their work, neighborhood is formed by optimizing a global cost function which leads to improve prediction accuracy. Cleger-Tamayo et al. [13] show that it is possible to introduce a new user selection criterion based on distance between real ratings. Anand and Bharadwaj [3] introduce two types of user similarities, namely local and global, and they combine them to compute a customer's NN. Their method is useful for increasing accuracy of CF systems.

Besides studies aiming at improving recommendation quality, different neighbor selection methods are proposed in order to relieve scalability issues in CF. Boumaza and Brun [10] come up with a global neighbors suggestion in which global users are shared among all active users. Since clustering is a helpful technique to determine distribution of members of any data, it is employed in CF systems in order to form NN according to previously clustered groups of customers [16,27,31,7]. Although these methods do not improve accuracy of predictions, they enhance online performance of CF.

In addition to proposed neighborhood selection methods, researchers have been motivated to study on presenting new similarity measures which enhance neighbor selection capability of CF algorithms. Ahn [2] develops a measure employing domain specific interpretation of user ratings and it overwhelms weaknesses of conventional similarity measures against cold start problem. Bobadilla et al. [9] develop a new similarity metric which combines rating information with contextual information obtained from customers and they utilize that kind of information to compute singularity of items. A novel similarity measure, which is called Jaccard Uniform Operator Distance (JacUOD), relying on comparison of vectors in different multidimensional vector spaces is proposed by Sun et al. [30]. According to experimental results of the study, JacUOD boosts accuracy of CF systems. Finally, Choi and Suh [12] take target item similarities into account in order to compute user correlations and they present a new similarity function.