The Journal of Systems and Software 95 (2014) 231–241

Contents lists available at ScienceDirect

The Journal of Systems and Software

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

# Distributed collaborative filtering with singular ratings for large scale recommendation

Ruzhi Xu[a,b], Shuaiqiang Wang[a,b,*], Xuwei Zheng[b], Yinong Chen[c]

[a] Department of Information Engineering, Qilu University of Technology, 58 Sangyuan Road, Jinan 250100, China
[b] School of Computer Science and Technology, Shandong University of Finance and Economics, 7366 2nd East Ring Road, Jinan 250014, China
[c] School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, P.O. Box 878809, Tempe, AZ 85287-8809, USA

ARTICLE INFO

*Article history:*
Received 7 January 2014
Received in revised form 24 April 2014
Accepted 25 April 2014
Available online 6 May 2014

*Keywords:*
Collaborative filtering
Recommender systems
Distributed framework

ABSTRACT

Collaborative filtering (CF) is an effective technique addressing the information overloading problem, where each user is associated with a set of rating scores on a set of items. For a chosen target user, conventional CF algorithms measure similarity between this user and other users by utilizing pairs of rating scores on common rated items, but discarding scores rated by one of them only. We call these comparative scores as *dual ratings*, while the non-comparative scores as *singular ratings*. Our experiments show that only about 10% ratings are dual ones that can be used for similarity evaluation, while the other 90% are singular ones. In this paper, we propose SingCF approach, which attempts to incorporate multiple singular ratings, in addition to dual ratings, to implement collaborative filtering, aiming at improving the recommendation accuracy. We first estimate the unrated scores for singular ratings and transform them into dual ones. Then we perform a CF process to discover neighborhood users and make predictions for each target user. Furthermore, we provide a MapReduce-based distributed framework on Hadoop for significant improvement in efficiency. Experiments in comparison with the state-of-the-art methods demonstrate the performance gains of our approaches.

## 1. Introduction

The Web has been flooded with an overwhelming amount of information, which represents one of today's major challenges of Web-based computing and information mining (Chen and Tsai, 2014). As an effective technique addressing the problem, recommender systems attempt to make predictions and give recommendations based on a large collection of users' historical behaviors and ratings. They have become a de facto standard and must-owned tools for e-commerce to promote business and to help customers discovering new products (Sarwar et al., 2000). Prominent examples include those systems used in Amazon (www.amazon.com), eBay (www.ebay.com/), Facebook (www.facebook.com/), LinkedIn (www.linkedin.com), and Netflix (www.netflix.com/).

Collaborative filtering (CF) (Chen et al., 2013; Herlocker et al., 2004; Su and Khoshgoftaar, 2009) is one of the most successful methods of building recommender systems. CF algorithms are based on the assumption that users will rate and act on other items similarly if they have rated items similarly or had similar behaviors (Goldberg et al., 1992; Resnick et al., 1994). CF utilizes the user–item rating matrix to make predictions and recommendations, avoiding the need of collecting extensive information about items and users. In addition, CF can be easily adopted in different recommender systems without requiring any domain knowledge (Liu and Yang, 2008).

Given the effectiveness and convenience, many CF methods have been proposed, which fall into two categories: memory-based (Deshpande and Karypis, 2004; Herlocker et al., 1999; Liu and Yang, 2008; Resnick et al., 1994; Sarwar et al., 2001; Wang et al., 2012) and model-based (Liu et al., 2009; Rendle et al., 2009; Shani et al., 2005; Si and Jin, 2003; Sun et al., 2012; Weimer et al., 2007). Memory-based methods make predictions based on similarities between users or items, while model-based methods make predictions based on a mathematical model.

In this study, we focus on the memory-based CF. Memory-based approach has been shown to be easy to implement, have strong robustness, and are effective (Hofmann, 2004). They take

* Corresponding author at: School of Computer Science and Technology, Shandong University of Finance and Economics, 7366 2nd East Ring Road, Jinan 250014, China. Tel.: +86 18253179913.
*E-mail addresses:* xrzpuma@gmail.com (R. Xu), shqiang.wang@gmail.com (S. Wang), xuwei_zheng@gmail.com (X. Zheng), yinong@asu.edu (Y. Chen).

http://dx.doi.org/10.1016/j.jss.2014.04.045
0164-1212/© 2014 Elsevier Inc. All rights reserved.

the advantage of big data analysis with instance-based learning to make predictions. The impact of the noise data can be significantly reduced in the averaging process of large amount of data. This approach is particularly promising in commercial environments where large amount of data are available, and thus many commercial systems, such as Amazon.com, are memory-based (Linden et al., 2003).

In the existing memory-based CF algorithms, each user is associated with a set of scores on rated objects, either based on rated items or based on user preferences. These algorithms start with measuring the similarity between users by utilizing pairs of rating scores on commonly rated objects, but discarding scores rated by one of them only. In this paper, we called these comparative rating scores as *dual ratings*, while the non-comparative scores as *singular ratings*.

Our experimental results on two movie rating datasets show that only about 10% of ratings are dual ones that can be used for similarity evaluation in CF algorithms, while the other 90% are singular ones and are discarded in these approaches. Indeed, the singular rating data are less relevant to a target user. However, the amount matter in big data analysis. The more data we have, the more relevancies can be discovered. We believe that it is an important issue to explore a practical way of making full use of the majority of data resources, by adding the singular ratings into dual rating studies.

For this reason, we propose SingCF, which takes the advantage of the unused singular ratings to improve the accuracy of CF algorithms. We first process the unrated scores of singular ratings, find the correlations, and transform them into dual ones. Then we perform a CF process to discover neighborhood users and make predictions for each target user. In addition, we study the equivalence of the similarity measure and the prediction formula between rating-oriented and ranking-oriented CF algorithms. We prove that they are equivalent in principle by showing that the ranking-oriented algorithms can be obtained from the rating-oriented techniques.

In this paper, we implement two versions of SingCF for validation purpose, a rating-oriented and a ranking-oriented. Experimental results in comparison demonstrate that our approach outperform the existing methods.

All CF algorithms have a time complexity $n^2$, where $n$ is the number of users, as the similarities between each pair of users need to be computed to discover potential neighborhood users for each target user to make predictions (Dean and Ghemawat, 2004). Furthermore, the users and data amount of the recommender systems increase rapidly, and most of them are stored in a distributed storage system for the scalability consideration (Lee and Chang, 2013). The solutions to the big data problem are the utilization of the distributed framework for CF algorithms.

In light of this requirement, in this paper, we also investigate the distributed framework for SingCF based on MapReduce in Hadoop, aiming to significantly improve the efficiency of SingCF.

Now, we discuss why SingCF can achieve better recommendation accuracy than conventional CF. The main difference between the two approaches is that SingCF uses additional singular ratings data. In fact, these data are valuable and useful, because 50% of the scores of these additional singular data are ground truth scores rated by users, and the other 50% of data are estimated scores. The mean errors of these data are quite low. Our study shows that the mean average errors of these estimated scores are only about 20% higher than that of the final predictions. As the result, SingCF achieves more accurate prediction than the dual CF algorithms.

The effectiveness of SingCF can be understood from another perspective. The unrated ratings of the singular ones are the missing values. Data cleaning is an essential technique in data mining, and one possible step is to attempt to fill in the missing values automatically with a measure of central tendency (Han et al., 2011). SingCF proposes an effective method to fill in the missing values using singular ratings, instead of pure speculation. It improves the prediction accuracy, which in turn improves the recommendation accuracy.

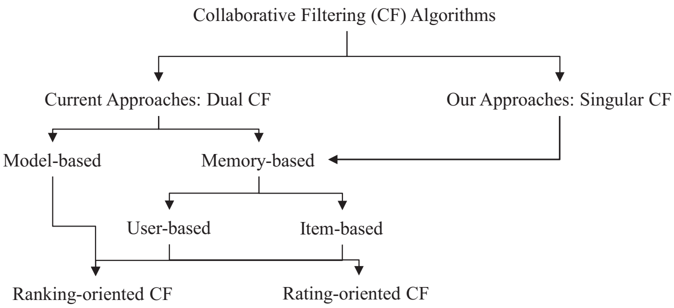
Collaborative Filtering (CF) Algorithms
Current Approaches: Dual CF
Our Approaches: Singular CF
Model-based
Memory-based
User-based
Item-based
Ranking-oriented CF
Rating-oriented CF

**Fig. 1.** Spectrum of different CF algorithms.

Our main contributions in this research include:

(1) We proposed SingCF, a collaborative filtering algorithm that incorporates singular ratings for improvement in recommendation accuracy, where we firstly estimated the unrated scores of singular ratings and transform them into dual ones as additional data for training and prediction.
(2) We proved that they were equivalent by showing that the ranking-oriented algorithms could be obtained from the rating-oriented techniques. Based on the same framework, we implemented two versions of SingCF for validation, a rating-oriented and a ranking-oriented for verification purpose.
(3) We provided DSingCF, a MapReduce-based distributed SingCF algorithm on Hadoop, aiming at significantly improving the efficiency of SingCF. Experiments in comparison with the state-of-the-art methods demonstrated the performance gains of our approaches.

The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 presents the preliminaries on memory-based collaborative filtering. Section 4 proposes the SingCF algorithm. Section 5 provides the distributed SingCF on Hadoop. Section 6 reports the experimental results. Section 7 concludes the paper.

## 2. Related work

### 2.1. Collaborative filtering

Fig. 1 gives the spectrum of different CF algorithms, as well as the position of the proposed SingCF algorithms in the spectrum.

Model-based CF algorithms use a mathematical model, a statistical model, or a learning model to analyze data and to make predictions on what a target user may purchase. Memory-based algorithms make predictions based on similarities of scores given by neighboring users and given to the related items. Many commercial systems, such as Amazon.com, use memory-based CF, as the approach is relatively easy to implement and results in good performance in recommendation.

The memory-based CF algorithms can be further categorized into two types: user-based and item-based (Wang et al., 2012). The user-based CF algorithms estimate the unknown ratings of a target user based on the ratings given by a set of neighboring users who tend to rate terms similarly, and thus, what they purchased may apply to the target user. On the other hand, in the item-based CF algorithms, item–item similarities are used to select a set of neighboring items that have been rated by the target user. In particular, given a target user/item, each user/item can be determined as her