Physica A 396 (2014) 195–203

Contents lists available at ScienceDirect

Physica A

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

Information filtering via collaborative user clustering modeling

Chu-Xu Zhang [a,b,c], Zi-Ke Zhang [a,b,*], Lu Yu [b], Chuang Liu [a,b], Hao Liu [a], Xiao-Yong Yan [d,*]

[a] Alibaba Research Center for Complexity Sciences, Hangzhou Normal University, Hangzhou 311121, PR China
[b] Institute of Information Economy, Hangzhou Normal University, Hangzhou 310036, PR China
[c] Web Sciences Center, University of Electronic Science and Technology of China, Chengdu 610054, PR China
[d] School of Systems Science, Beijing Normal University, Beijing 100875, PR China

HIGHLIGHTS

- Propose a user behavior model to cluster users and improve recommendation.
- Optimize the results by integrating the user clustering regularization term based on genres.
- Experimental results show that our method performs better than two other baseline methods.

ARTICLE INFO

Article history:
Received 3 August 2013
Received in revised form 12 November 2013
Available online 21 November 2013

Keywords:
Recommender systems
Collaborative filtering
Matrix factorization
User clustering regularization

ABSTRACT

The past few years have witnessed the great success of recommender systems, which can significantly help users to find out personalized items for them from the information era. One of the widest applied recommendation methods is the Matrix Factorization (MF). However, most of the researches on this topic have focused on mining the direct relationships between users and items. In this paper, we optimize the standard MF by integrating the user clustering regularization term. Our model considers not only the user-item rating information but also the user information. In addition, we compared the proposed model with three typical other methods: User-Mean (UM), Item-Mean (IM) and standard MF. Experimental results on two real-world datasets, *MovieLens* 1M and *MovieLens* 100k, show that our method performs better than other three methods in the accuracy of recommendation.

1. Introduction

Facing the explosive growth of the web, people would become lost in the web's info-thickets and often waste much time to search useful and personalized information, namely *Information Overload*. Confronting with such a problem, researchers from different areas have invented various tools, among which the search engine is the most outstanding. However, compared with recommender systems [1,2], which automatically match the user's taste based on the historical behaviors, the search engine is not personalized enough because it produces the same results for all users. Among various kinds of recommender systems, Collaborative Filtering (CF) [3,4] is the most widely used in different fields due to its advantages of requiring no domain knowledge, implementing easily and detecting the complex pattern that is hard to be exploited with

* Corresponding author at: Institute of Information Economy, Hangzhou Normal University, Hangzhou 310036, PR China. Tel.: +86 18657192267.
E-mail addresses: zhangzike@gmail.com (Z.-K. Zhang), yanxy@mail.bnu.edu.cn (X.-Y. Yan).

0378-4371/$ – see front matter
http://dx.doi.org/10.1016/j.physa.2013.11.024

the known data. As a result, CF has attracted much attention from both academic and industry fields in the past decade. In particular, the competition of Netflix Prize (NP) [5], has inspired different fields of researchers to propose various solutions to build corresponding recommender systems.

The basic idea of CF is that recommendation for the target user is made by predicting the preference of the uncollected items based on the neighbors. Neighbor is a group of persons with similar tastes when they rate the same items. Generally, there are two main types of CF: neighborhood and model based approaches [6]. In the early time, neighborhood based methods, including user- and item-based approaches, were the most widely applied in the industry, such as Amazon [7], and Google [8]. In recent years, experts from both academia and industry have witnessed the excellent performance of model-based approaches, especially the Latent Factor Model (LFM) [9]. As the typical representative technique of LFM based methods, Matrix Factorization (MF) provides an alternative method to represent the relationship between users and items. In the LFM, users and items are both represented in the same Latent Factor Space (LFS), hence the prediction is accomplished by directly evaluating the preferences of users for the uncollected items. Some MF methods [10–13] have been proposed in CF because of the high efficiency in dealing with large-scale datasets. Those approaches tend to fit the user-item rating matrix with low-rank matrix factorization and apply it to make rating predictions. MF is efficient in training since it assumes that only a few factors influence preferences in user-item ratings. The objective for minimizing the sum-squared errors can be easily solved by Singular Value Decomposition (SVD), and Expectation Maximization (EM) algorithms for solving weighted low-rank approximation was proposed in Ref. [13].

Since the success of MF in the Netflix Prize competition, a great many of variants are proposed. In Ref. [14], a matrix factorization framework with social network regularization was described. It provided a general method for improving recommender system by incorporating social network information. Karatzoglou [15] presented two simple models that take advantage of the temporal order of choices and ratings. These two models not only exploited the collaborative effects in the data, but also took into account the order in which items could be viewed by the users. Koren et al. [16] introduced a Markov Chain model which considered the collaborative effects using Tensor Factorization [17].

In addition, besides the traditional CF methods in the recommender system, there also emerged many variant methods based on statistical physics with the development of network science, see Refs. [18–24,24–28]. Most of these methods are based on bipartite networks [29]. Some of these methods are innovative and effective in improving not only accuracy but also diversity and novelty. Zhang et al. [20] proposed a recommendation algorithm based on an integrated diffusion on user-item-tag tripartite graphs and significantly improved accuracy, diversification and novelty of recommendations. Yang et al. [30] argued that the anchoring bias of users' online voting pattern can help to improve recommendation. In Ref. [21], a new algorithm that specifically addressed the challenge of diversity in recommender system is proposed. Lü et al. [23] introduced a recommendation algorithm based on the preferential diffusion process on user-object bipartite network.

In this paper, inspired by Social Network Matrix Factorization (SNMF) [14], we consider the neighbors' impact on the interest of each user in the same LFS and propose a recommendation model based on clustering [31,32] users (UCMF). Firstly, we represent the interest of each user with the statistical information of her behaviors on different item genres which are static tags given by system. Secondly, we classify all users into several groups by the K-Means clustering algorithm. Finally, we expand the standard MF by integrating a user clustering regularization term which describes that the users in the same group have similar interest. The results on *MovieLens 1M* and *MovieLens 100k* show that our model outperforms the standard MF method and other two baseline methods in the accuracy of recommendation.

2. User clustering model

2.1. Low-rank matrix factorization

CF techniques based on the MF method assume that users' ratings on items can be represented by a $N \times M$ matrix (N is the number of users and M the number of items). A low-rank matrix factorization approach tends to approximate R by multiplying L-rank factors,

$$R \approx U^T V, \tag{1}$$

where $U \in \mathbb{R}^{L\times N}$, $V \in \mathbb{R}^{L\times M}$ with $L < min(N, M)$, and the matrix R is usually sparse.

Traditionally, the SVD method is employed to approximate the rating matrix R by minimizing

$$\frac{1}{2}\|R - U^T V\|_F^2, \tag{2}$$

where $\|\cdot\|$ denotes the Frobenius form. We only need to factorize the observed ratings in matrix R because of large missing values. So Eq. (2) can be transformed to

$$\min_{U,V} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}\left(R_{ij} - U_i^T V_j\right)^2, \tag{3}$$

where I_{ij} equals 1 if user u_i rates item v_j and 0 otherwise.