Knowledge-Based Systems 83 (2015) 42–50

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Addressing cold-start: Scalable recommendation with tags and keywords

Ke Ji [a], Hong Shen [b,c,*]

[a] School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
[b] School of Information Science and Technology, Sun Yat-sen University, Guangzhou, China
[c] School of Computer Science, University of Adelaide, Australia

ARTICLE INFO

Article history:
Received 29 October 2014
Received in revised form 19 January 2015
Accepted 6 March 2015
Available online 13 March 2015

Keywords:
Recommender systems
Matrix factorization
Tag-keyword
Cold start
Scalability

ABSTRACT

Cold start problem for new users and new items is a major challenge facing most collaborative filtering systems. Existing methods to collaborative filtering (CF) emphasize to scale well up to large and sparse dataset, lacking of scalable approach to dealing with new data. In this paper, we consider a novel method for alleviating the problem by incorporating content-based information about users and items, i.e., tags and keywords. The user-item ratings imply the relevance of users' tags to items' keywords, so we convert the direct prediction on the user-item rating matrix into the indirect prediction on the tag-keyword relation matrix that adopts to the emergence of new data. We first propose a novel neighborhood approach for building the tag-keyword relation matrix based on the statistics of tag-keyword pairs in the ratings. Then, with the relation matrix, we propose a 3-factor matrix factorization model over the rating matrix, for learning every user's interest vector for selected tags and every item's correlation vector for extracted keywords. Finally, we integrate the relation matrix with the two kinds of vectors to make recommendations. Experiments on real dataset demonstrate that our method not only outperforms other state-of-the-art CF algorithms for historical data, but also has good scalability for new data.

1. Introduction

Collaborative filtering (CF) is a common technique in recommender systems, which only uses past review or rating data to make good recommendations without need for exogenous information. The classic CF algorithms, such as memory-based [1–4] and mode-based [5–7], has been widely applied to many areas, like ecommerce (e.g., Amazon, Netflix), social networks (e.g., Twitter, Facebook), and review sites (e.g., Movielens, Douban). Despite the great success in earlier period, recommender systems based on pure CF approach suffer from several thorny problems, such as data sparsity, scalability, cold start and poor prediction. Two recent attempts have been done to alleviate the above problems, one with contextual information [8–11] and one with social information [12–18]. The present CF technique is far more advanced now than when it was proposed. However, most existing methods are concerned chiefly with how to improve recommendation accuracy on large and sparse dataset, careless of cold start problem for new data, i.e., new users and new items.

In real application, recommender systems would continue to collect data on uptime. New users and new items that appear as time goes on cause cold start problem – the bad performance on new users and new items because there is few or no rating for them. A robust and scalable recommender system need not only to scale well up to large and sparse dataset, but also to have incremental processing for new data. Unfortunately, there is little effective research [14,19] in solving cold start satisfactorily, and especially the direct CF approaches on the user-item rating matrix are hardly extended to meet these needs. Thus, this requires additional content-based information on user's interest and item's character that are adequate to identify new users and items.

The simplest way to describe content-based information about users and items is to maintain an explicit list of features (also often called tag or keyword). Tags selected by users represent their specific interest and keywords extracted from the corresponding profile of items quite reasonably indicate the characters of items. For example, if a user likes mountain climbing and swimming, he may select "mountain climbing" or "swimming" to be his tag. If an item's content is about Apple and phone, "iphone" or "iphone 6" may be extracted as the item's keyword. Note that tag and keyword respectively belong to the two corpus with different context, and there is no direct relation between them. But as shown in

* Corresponding author at: School of Computer and Information Technology, Beijing Jiaotong University, China.
E-mail addresses: 12120425@bjtu.edu.cn (K. Ji), hshen@bjtu.edu.cn (H. Shen).

http://dx.doi.org/10.1016/j.knosys.2015.03.008
0950-7051/

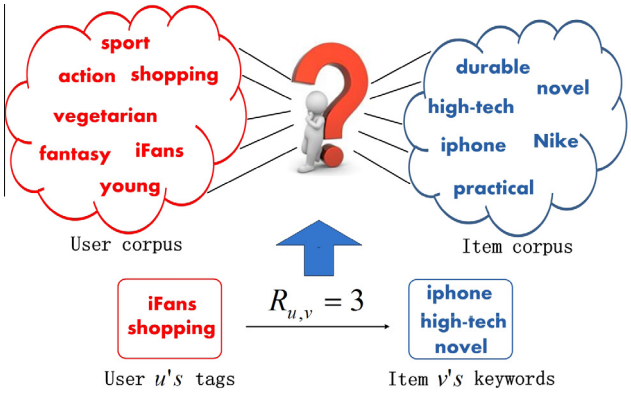

Fig. 1. It is difficult to build the relationship between tags and keywords because they belong to different corpus. But, the user-item ratings imply the relevance of users' tags to items' keywords.

Fig. 1, the user-item ratings imply the relevance of users' tags to items' keywords, so a very intuitive way (Fig. 2) is that we can convert the direct prediction on the user-item rating matrix into the indirect prediction on the tag-keyword relation matrix. More importantly, once any new user's tags or item's keywords are known, it is very easy to provide cold start recommendations for them using the known tag-keyword relations.

In this paper, we present an effort to enhance the scalability of CF systems on incremental processing for new users and new items with the aid of the tag-keyword relation. We first propose a novel neighborhood approach, which predicts the missing relation based on the known relations between tags and keywords with the highest co-occurrence frequency, for building the full tag-keyword relation matrix. Then, we propose a novel matrix factorization model over the rating matrix, which constructs a 3-factor factorization through the sub-relation matrices corresponding to every user-item pair, for learning every user's interest vector for selected tags and every item's correlation vector for extracted keywords. Finally, the tag-keyword relation matrix and two kinds of vectors are integrated to make recommendations. We have conducted experiments on real dataset published by KDD Cup 2012. The result and analysis demonstrate that our method not only improves recommendation accuracy for historical data, but also has good scalability for new users and new items.

The rest of this paper is organized as follows: In Section 2, we review the related work. In Section 3, we give the problem formulation. In Section 4, we present our method in detail. In Section 5, we report the experimental settings and results. In Section 6, we conclude this paper.

2. Related work

Collaborative filtering is one of the most popular and successful techniques in recommender systems, which can be divided into two major classes: memory-based and model-based. Memory-based approaches [1–4] use statistical techniques to build neighborhood relationship between users or items, and then a predict missing rating based a weighted sum of ratings from similar users or items. Model-based approaches [5–7], in contrast, use the user-item ratings to train a model first, and then make recommendation via the model instead of the similarity comparison (e.g., *VSS* or *PCC* [2]) on the original database. Many experimental results and conclusions argue that model-based approaches lead to somewhat more accurate results, while memory-based approaches have some practical advantages.

Matrix factorization is an effective model-based CF approach, which decomposes the $m \times n$ rating matrix into two low-rank matrices: $U \in R^{m\times d}$ and $V \in R^{n\times d} (d < min(m,n))$, the product of which can be used to construct a matrix approximation $R \approx UV^T$. One of the most popular methods by minimizing the sum-of-squared-errors objective function with quadratic regularization terms is

$$\mathcal{L} = \sum_{u=1}^{m}\sum_{v=1}^{n} I_{u,v}\left(R_{u,v} - U_u V_v^T\right)^2 + \lambda_1 \|U\|_F^2 + \lambda_2 \|V\|_F^2 \quad (1)$$

where λ_1 and λ_2 is the extent of regularization and $\|\cdot\|_{Fro}$ is the Frobenius norm. $I_{u,v}$ is the indicator function that is equal to 1 if user u rated item v and equal to 0 otherwise. The optimization problem in Eq. (1) is generally solved by performing gradient decent on U_u and V_v. Then, the missing rating can be predicted by $U_u V_v^T$.

The rapid development of the Internet poses some problems for traditional CF systems, such as data sparsity, scalability, cold start and poor prediction. In recent years, two recent attempts have been done to alleviate the above problems, one with contextual information [10,11] and one with social information [16–18]. Context-aware recommendation models [8,9] have been proposed

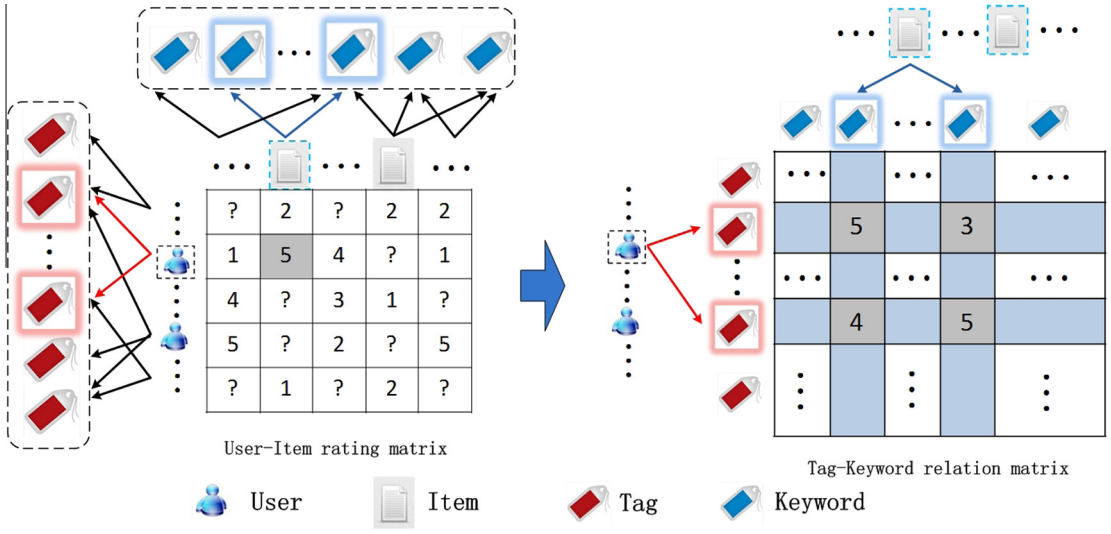

Fig. 2. Each user is labeled with some tags and each item is labeled with some keywords. We convert the direct prediction on the user-item rating matrix into the indirect prediction on the tag-keyword relation.