Knowledge-Based Systems 000 (2018) 1–8

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Incorporating temporal dynamics into LDA for one-class collaborative filtering☆

Haijun Zhang[a], Xiaoming Zhang[b,*], Zhiyong Tian[a], Zhenping Li[a], Jianye Yu[a], Feng Li[a]

[a] School of Information, Beijing Wuzi University, Beijing 101149, China
[b] State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China

ARTICLE INFO

Article history:
Received 26 June 2017
Revised 20 February 2018
Accepted 24 February 2018
Available online xxx

Keywords:
One-class collaborative filtering
Latent dirichlet allocation
Temporal dynamics

ABSTRACT

In the one-class collaborative filtering (OCCF) scenario, the elements of the user-item rating matrix consist each take one of only two values: either "like" or unknown. Previous methods for solving the OCCF problem can be roughly categorized into content-based methods, pointwise methods, and pairwise methods. A fundamental assumption of these approaches is that all missing values in a rating matrix can be treated as "dislike". However, this assumption may not hold because the missing values are not always negative. Sometimes users do not give positive feedback on an item simply because they are not familiar with it rather than because they dislike it. In addition, content-based methods usually require textual information on the items. In many cases, however, sufficient textual information is not available; therefore, content-based methods are not applicable. Moreover, a user's preference for items usually drifts over time, but the previous methods cannot capture the temporal dynamics of this drift. In this paper, we propose to modify the latent Dirichlet allocation (LDA) model to address the above-mentioned problems. Our method uses only observed rating data to predict users' interests and effectively avoids the issue of data skew. Furthermore, to address the issue that users' preferences for items usually drift over time, we assign a different weight to each rating according to its timestamp when using Gibbs sampling to estimate the parameters of the LDA model. In this way, the temporal dynamics of the user preferences can be captured. We report experiments conducted to evaluate our model. The results show that the proposed model outperforms state-of-the-art approaches for the OCCF problem.

1. Introduction

Recommender systems have arisen from practical reqirements for personalized e-services in many application domains, such as e-government, e-business, e-commerce, e-libraries, e-learning, e-tourism, e-resource services and e-group activities [1]. Collaborative filtering is a typical and popular information filtering technology used in recommender systems [2]. There are two main focuses of research in the context of collaborative filtering recommender systems: one-class collaborative filtering (OCCF) [3–9] and multi-class collaborative filtering (MCCF) [10–21]. In the OCCF scenario, the values of the elements of the user-item rating matrix R can be only either 1 or unknown. A value of 1 associated with an element r_{ui} means that user u provided positive feedback on item i, e.g., "like" on Facebook, "bought" on Amazon, "collect" on Taobao or "follow" on Sina Weibo. In the MCCF scenario, the elements of the user-item rating matrix are multi-valued, and each represents the degree of a user's preference for an item. Collaborative filtering (CF) recommendation algorithms use the known values in the user-item rating matrix to predict unknown values. Based on this, items that users might wish to buy in the future can be recommended to those users. Machine-learning-based methods [10–13], such as matrix factorization [14–20], have achieved great success in solving the MCCF problem. However, such methods often suffer from a severe over-fitting problem when tackling the OCCF problem because the rating data in OCCF are highly imbalanced [5,22]. To alleviate this problem, content-based methods such as CTR [23] and CTR–SMF [24] have been proposed. However, when content information on the items cannot be obtained, these methods are infeasible. Wang et al. [21] proposed a novel diffusion-based recommendation algorithm which calculates the similarity between users and generates a recommendation list for the target user by considering both implicit and explicit feedback data. This algorithm achieves better performance than the baselines, but it

☆ A short version of this paper has appeared as a poster paper in the proceedings of CIKM 2014.

* Corresponding author.
E-mail addresses: zhanghaijun@bwu.edu.cn (H. Zhang), yolixs@buaa.edu.cn (X. Zhang), tianzhiyong@bwu.edu.cn (Z. Tian), lizhenping@bwu.edu.cn (Z. Li), jy.yu@siat.ac.cn (J. Yu), lifeng@bwu.edu.cn (F. Li).

https://doi.org/10.1016/j.knosys.2018.02.036
0950-7051/

requires explicit feedback data. Pointwise methods [4,7] and pairwise methods [3,5,9] use sampling techniques to alleviate the data skew problem in OCCF. Of the available pairwise methods, BPR [5,9] and GBPR [3] have achieved great success. However, in all of these methods, missing values are considered to be negative. Moreover, none of these methods models temporal dynamics to capture the evolution of a user's interests. Some time-dynamic topic models [25,26] have proposed in the text mining field; however, in that scenario, all the words in a document have the same time stamp, whereas in a recommender system, each rating on an item (treated as a word) recorded for a user (treated as a document) has a time stamp. Zhong et al. [27] proposed a novel LDA-based time-aware service recommendation method to recommend web services for users. Their method requires textual descriptions of web services and search keywords entered by users. Zheng et al. [28] split microblog data into N time intervals and applied the LDA model to each of these sub-datasets individually to capture the interest drift of users. However, in the OCCF scenario, the data are very sparse, and splitting the data into N time intervals will exacerbate this sparsity and cause the estimated parameters to be unreliable. Thus, these time-dynamic topic models cannot be applied to solve the OCCF problem. Recently, based on the FISM [29] approach, the TOCCF [30] method was proposed, in which a time-aware weight is introduced into the loss function to capture the interest drift of users. However, this method also requires the sampling of a considerable amount of negative data.

In this paper, we exploit the latent Dirichlet allocation (LDA) model to address the OCCF problem. Compared with our previous work [8], the differences and new contributions of the present paper are as follows. (1) In our previous work [8], an LDA-based collaborative filtering algorithm was proposed, which is presented in Sections 3.1 and 3.2 of this paper. In Section 3.3 of this paper, we extend the previously proposed model [8] to incorporate the temporal dynamics of a user's preferences. Each observed datum is assigned a specific weight according to its rating timestamp for use when estimating the parameters of the LDA model. Because a user's preference for items can drift over time, the modeling of the temporal dynamics of user preferences should be a key concern when designing recommender systems or general customer preference models. (2) In Section 4, four real-world datasets, including MovieLens 100K[1], MovieLens 1M[1], MovieLens 10M[1] and Netflix[2], are used to evaluate the method. Experiments show that our methods, including our previous method [8], generate better recommendations than state-of-the-art methods. (3) In contrast to the seminal work of Blei et al. [31], Our method is specially designed for solving the OCCF problem and uses Gibbs sampling to estimate the parameters of the LDA model. Gibbs sampling provides a simple and extensible means of incorporating more information into the LDA model. In contrast to other conventional LDA-based CF algorithms [23,24,32–34], the proposed method uses only the observed rating scores for items to predict a user's interests. It neither requires content information on the items nor assumes that users prefer items to which they have previously given positive feedback over other items.

2. Related works

In this section, we review the literature on the state-of-the-art approaches proposed to address the OCCF problem. Overall, there are four main types of approaches for the OCCF problem: (1) content-based methods, (2) pointwise methods, (3) pairwise methods, and (4) time-aware methods.

2.1. Content-based methods

When textual content information on items is available, content-based methods can be applied. In such a method, a graph model is used to model the user-item rating scores by treating each item as a document and all items as a corpus. The values of all unknown data are assumed to be negative and are assigned a small value such as 0, whereas positive feedback is assigned a value of 1. Collaborative topic regression (CTR) [23] and collaborative topic regression with social matrix factorization (CTR–SMF) [24] are two representative examples of content-based methods. The CTR–SMF model is an extension of the CTR model and can be used to provide users with better recommendations when the item contents, rating records and social network information are available. However, the necessary textual contents are usually difficult to obtain, and these two models also treat missing data as negative, which is an unreasonable assumption.

2.2. Pointwise methods

In pointwise methods, unknown user-item pairs are considered equivalent to negative feedback and are assigned a rating score of 0. Then, machine learning methods, e.g., weighted low-rank approximations (WLRA) [4], are designed to fit the rating score matrix. Because most of the ratings are negative, two techniques are commonly adopted to alleviate the data skew issue: (1) assigning a low weight to negative data and a higher weight to positive data or (2) sampling the negative data at a low probability.

2.3. Pairwise methods

Pairwise methods can usually achieve better performance on the OCCF problem than pointwise methods. Some representative pairwise methods include the Bayesian personalized ranking-based matrix factorization (BPR) method [5,9] and the group Bayesian personalized ranking (GBPR) method [3]. The BPR method assumes that user u prefers item i to item j if the user-item pair (u, i) is observed and (u, j) is not. Then, it treats pairs of items as the basic units for analysis and maximizes the likelihood of the pairwise preferences over the observed and unobserved items. In comparison with BPR, GBPR uses a stronger constraint, which can be expressed as shown in Eqs. (1) and (2):

$$if \;\; u \in G_i \; and \;\; u \notin G_j \;\; then \;\; r_{G_i,i} > r_{u,j} \tag{1}$$

$$r_{G_i,i} = \frac{\sum_{u \in G_i} r_{u,i}}{|G_i|} \tag{2}$$

where G_i is the group of users who have given positive feedback on item i, $r_{u,j}$ denotes the preference value of user u for item j, and $r_{G_i,i}$ denotes the group preference value of group G_i for item i, which is defined as the average preference value of all users in group G_i for item i, as expressed in Eq. (2).

In GBPR, the sigmoid function $f\left(r_{G_i,i} - r_{u,j}\right) = \frac{1}{1+\exp\left(-r_{G_i,i}+r_{u,j}\right)}$ is used to approximate the probability $\Pr\left(r_{G_i,i} > r_{u,j}\right)$, and $r_{u,j}$ is factorized as a product of two vectors, as expressed in Eq. (3):

$$r_{u,j} = q_u \cdot s_j^T \tag{3}$$

where q_u is the latent vector of user u and s_j is the latent vector of item j. These latent vectors are learned by maximizing the likelihood of the pairwise preferences.

2.4. Time-aware methods

Time-aware One-Class Collaborative Filtering (TOCCF) [30] is a recently proposed time-aware method for solving the OCCF prob-

[1] http://grouplens.org/datasets/movielens/.
[2] http://www.netflixprize.com/.

Please cite this article as: H. Zhang et al., Incorporating temporal dynamics into LDA for one-class collaborative filtering, Knowledge-Based Systems (2018), https://doi.org/10.1016/j.knosys.2018.02.036