Exploit latent Dirichlet allocation for collaborative filtering

Zhoujun LI1, Haijun ZHANG (✉)1,2, Senzhang WANG3,4, Feiran HUANG1, Zhenping LI2, Jianshe ZHOU5

1 State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China
2 School of Information, Beijing Wuzi University, Beijing 101149, China
3 College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
4 Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 211106, China
5 Beijing Advanced Innovation Center for Imaging Technology, Capital Normal University, Beijing 100048, China

© Higher Education Press and Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract Previous work on the one-class collaborative filtering (OCCF) problem can be roughly categorized into pointwise methods, pairwise methods, and content-based methods. A fundamental assumption of these approaches is that all missing values in the user-item rating matrix are considered negative. However, this assumption may not hold because the missing values may contain negative and positive examples. For example, a user who fails to give positive feedback about an item may not necessarily dislike it; he may simply be unfamiliar with it. Meanwhile, content-based methods, e.g. collaborative topic regression (CTR), usually require textual content information of the items, and thus their applicability is largely limited when the text information is not available. In this paper, we propose to apply the latent Dirichlet allocation (LDA) model on OCCF to address the above-mentioned problems. The basic idea of this approach is that items are regarded as words, users are considered as documents, and the user-item feedback matrix constitutes the corpus. Our model drops the strong assumption that missing values are all negative and only utilizes the observed data to predict a user’s interest. Additionally, the proposed model does not need content information of the items. Experimental results indicate that the proposed method outperforms previous methods on various ranking-oriented evaluation metrics. We further combine this method with a matrix factorization-based method to tackle the multi-class collaborative filtering (MCCF) problem, which also achieves better performance on predicting user ratings.

Keywords latent Dirichlet allocation, one-class collaborative filtering, multi-class collaborative filtering

1 Introduction

There are two major research focuses in collaborative filtering recommendation systems: one-class collaborative filtering (OCCF) [1–6], and multi-class collaborative filtering (MCCF). In the OCCF scenario, the values of the elements in the user-item rating matrix R can only be either 1 or unknown. Element $r_{ui}$ with a value 1 means user $u$ provided positive feedback on the item $i$, e.g., “like” in Facebook, “bought” in Amazon, “collect” in Taobao and “follow” in Sina weibo. For the MCCF problem, the element in the user-item rating matrix is multi-valued, which represents the degree of a user’s preference to an item. Collaborative filtering (CF) recommendation algorithms utilize the observed values in the user-item rating matrix to predict users’ interests regarding specific items. Based on this, the algorithm then recommends some new items that users would potentially buy in the future. Machine learning based methods [7–13], such as matrix factorization methods [14–19], have achieved great success...
solving the MCCF problem. However, these methods usually suffer from severe overfitting when they are applied to the OCCF problem due to the extremely skewed distribution of the rating data [20]. Pointwise methods [2,3], pairwise methods [1,3], and content-based methods [21,22] are proposed to solve the data skew issue in the OCCF problem. However, a major limitation of these methods is that they assume that all the unknown data are negative. This can lead to new data skew problems. Meanwhile, content based methods need the textual contents of the items, but the content information is not always available. Although matrix factorization methods such as RSVD [14] usually achieve better performance in addressing the MCCF problem, they have too many parameters and also exhibit overfitting problems.

The latent Dirichlet allocation (LDA) model is widely used to capture the latent topics from textual corpus [21–25]. In that scenario, the corpus is a collection of documents, and a document is a collection of words. The corpus can be modeled as a matrix with each row of the matrix denoting a document, each column of the matrix denoting a word, and each entry of the matrix denoting the times the word appears in the document. In a recommendation system, usually only the user-item rating matrix is available while the item’s textual contents are not easy to obtain. If we consider a user as a document, an item as a word, and a rating score as the occurrence frequency of the word in the document, the LDA model can be applied without textual contents of the items to capture the latent topics from the user-item rating matrix. The parameters learned from LDA are probability distributions, and these parameters are all positive numbers whose summation is 1. These characteristics make the LDA model more robust to overfitting. Based on this idea, we propose two novel methods, LDA-OCCF and LDA-RSVD, to tackle the OCCF and the MCCF problems, respectively.

This study makes the following two major contributions: 1) Different from conventional LDA-based CF methods [21–25] for the OCCF problem, our method, LDA-OCCF, only uses the observed rating scores of the items to predict a user’s interests. It neither needs the content information of items nor assumes that users prefer the items on which they gave positive feedback over other items. 2) We further extend the method to deal with the MCCF problem, and propose a novel method, LDA-RSVD. In this method the parameters learned from the LDA model are used to restrict the RSVD model to alleviate overfitting. Experiments show that our methods outperform current state-of-the-art methods in both OCCF and MCCF.

Compared with our previous work [6], the new contributions of this paper are as follows. 1) In our previous work [6], an LDA-based collaborative filtering algorithm was proposed, which is presented in Section 3.1 of this paper. Compared with the seminal work of Blei et al. [23], our method was designed for both OCCF and MCCF problems, and we used the Gibbs sampling method to estimate the parameters in the LDA model. The Gibbs sampling method provided a simple and effective implementation for the LDA model. In contrast to other conventional LDA-based CF algorithms [21,22], the proposed method uses only the observed rating scores for items to predict a user’s interest. It neither needs the content information of the items nor assumes that users prefer the items to which they previously gave positive feedback over other items. 2) In Section 3.2 of this paper, we extend the previous model [6] to cope with the MCCF problem. 3) In Section 4, three real-world datasets are used to evaluate the method. Experiments show that our methods make better recommendations than current state-of-the-art methods.

The rest of the paper is organized as follows. In the next section, we review previous works related to the problems of OCCF and MCCF. In Section 3, an LDA based approach is proposed for the OCCF and MCCF problems. In Section 4, we empirically compare our method to current state-of-the-art methods, and show experimental results. Finally, we conclude the paper and discuss future work in Section 5.

2 Related work

In this section, we review the literature of state-of-the-art approaches proposed to address OCCF and MCCF. Overall, there are mainly three types of approaches for the OCCF problem: 1) pointwise methods, 2) pairwise methods, and 3) content-based methods. There are many methods to solve MCCF problems, of which the matrix factorization based method showed an outstanding performance.

2.1 Pointwise methods

Pointwise methods take an implicit feedback as the absolute preference score. For example, an observed user-item pair is interpreted as positive feedback and is assigned a high absolute rating score, e.g., 1. An unobserved user-item pair is considered as negative feedback and is assigned a rating score of 0. Then, machine learning based methods, e.g., weighted low-rank approximations (WLRA) [2], are designed to fit the rating score matrix. Because most of the data are negative, two types of techniques are usually adopted to correct the negative bias. 1) The weight of the negative data is assigned with a small value relative to the positive data, and 2) the