ARTICLE IN PRESS

IPL:5453

Information Processing Letters ••• (••••) •••–•••

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

Rating prediction using review texts with underlying sentiments

Dongjin Yu *, Yunlei Mu, Yike Jin

College of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China

A R T I C L E I N F O

Article history:
Received 7 November 2015
Received in revised form 28 June 2016
Accepted 2 August 2016
Available online xxxx
Communicated by Wen-Lian Hsu

Keywords:
Collaborative filtering
Latent factor model
Information retrieval
Cold-start
Sentiments

A B S T R A C T

Recommender systems typically produce a list of recommendations to precisely predict the user's preference for the items. For this purpose, latent factor models, such as matrix factorization, are usually employed to find latent factors that can characterize both users and items by observed rating scores. Recently, online user feedback accompanied with review texts has become increasingly common. The review texts contain not only users' attention to the situation of the different aspects of items, but also users' sentiment towards different aspects of specific items. However, traditional latent factor models often ignore such review texts, and therefore fail to characterize users and items precisely. Furthermore, although some current studies do employ review texts, many of them do not consider how sentiments in reviews influence the rating scores. In this paper, we propose an extended *Hidden Factors as Topics Model* (HFT) (a model combining the *Latent Factor model* and the *Latent Dirichlet Allocation*) based on *Aspect and Sentiments Unification Model* (ASUM) (an extended topic model), called *Ratings Are Sentiments* (RAS). By combining users' sentiments in review texts and their rating scores, our model can learn more precise latent factors of users and items compared with the baseline models. The extensive experiments on large, real-world datasets demonstrate that the RAS model performs better than both the latent factor model and the HFT model and alleviates the cold-start problem to some extent.

1. Introduction

Personalized recommender systems help users to find a wide variety of products online, and assist users in making decisions. At such highly rated Internet sites as Amazon.com, YouTube, Netflix, Yahoo, TripAdvisor, Last.fm, and IMDb, recommender systems play a very important role. Users could take advantage of these recommender systems to find a variety of products, videos, books, and news that they like from the massive available item set [1].

Recently, collaborative filtering (CF) techniques have achieved great success in the personalized recommender system. There are two types of approaches for CF recommender system: neighborhood-based approaches [2] and model-based approaches [3,4]. The neighborhood-based approaches focus on the relation between items or users. More specifically, by employing the neighborhood-based approaches, we can obtain the nearest-neighbors of items or users by calculating their similarity, and then take advantage of these neighbors to produce accurate and personalized recommendations. On the other hand, the model-based approach, such as the latent factor models (matrix factorization, aka SVD, as an example), transforms both items and users to the same latent factor space,

* Corresponding author.
E-mail addresses: yudj@hdu.edu.cn (D. Yu), muyunlei@gmail.com (Y. Mu), jyk0406@126.com (Y. Jin).

http://dx.doi.org/10.1016/j.ipl.2016.08.002
0020-0190/

which can be used to explain ratings by characterizing both items and users on factors automatically inferred from user feedback [3]. Generally, model-based approaches can produce more accurate prediction than neighborhood-based approaches do. However, both the neighborhood-based and the model-based approaches are sensitive to sparse data. They always suffer from the lack of available ratings, which is called the cold-start problem [5].

It is worth noting that users likely write a review when they vote on an item. Traditional CF approaches only use rating information but ignore the abundant information embedded in reviews. In fact, we can exploit the information embedded in reviews to solve the cold-start problem as mentioned above. McAuley et al. [6] proposed a model named *Hidden Factor as Topic model* (HFT), which combined the traditional latent factor model [3] and *Latent Dirichlet Allocation* (LDA) [7]. HFT maps the latent factors of the latent factor model to the topic distribution of LDA and makes the rating prediction more accurate compared with the traditional latent factor model. However, as a matter of fact, users not only talk about different aspects of one certain item but also the different sentiments towards these aspects. For instance, when Tom wrote a review for a laptop, he submitted some sentences to express satisfaction with the "shape" and "performance" of the laptop and some sentences to express his displeasure with the "battery" and "radiating". Unfortunately, HFT ignores the sentiments that the user expresses in review texts. However, these sentiments often express the extent of the user's satisfaction or displeasure.

In this paper, we take the sentiments towards different aspects in reviews into consideration. These sentiments can explain the ratings naturally. Taking advantage of *Aspect and Sentiment Unification Model* (ASUM) [8], we propose an extended HFT model, named *Ratings Are Sentiments* (RAS). RAS maps the rating scores to sentiment distribution probability space, from which we obtain different aspects of different sentiments. In this way, we could obtain more precise latent factor vectors of users and items than the traditional latent factor model or HFT does. In other words, our approach improves the rating prediction performance and alleviates the cold-start problem to some extent.

The contributions of our work are as follows: 1) We structure a transformation that links the users' or items' average rating with sentiment probability. Furthermore, we propose a model that combines HFT with ASUM based on this transformation. 2) We extend the transformation of HFT that links latent factors with topic probabilities to model aspects of different sentiments. 3) We introduce an effective optimization algorithm that optimizes the parameters of RAS alternately. 4) We demonstrate our model's outstanding performance in prediction accuracy and convergence rate compared with the previous works on real-world datasets.

The remainder of this paper is organized as follows. We first present some related works in Section 2. After the motivation given in Section 3, we briefly introduce two related models: ASUM and HFT. Afterwards, we propose our RAS model in detail in Section 4. Then, in Section 5, we provide the optimization algorithm for the model learning of RAS and analyze its time and space complexity. Subsequently, we describe the experiment using a real-world dataset in Section 6. Following the discussion of the experiment provided in Section 6, the last section summarizes our work and outlines the future research directions.

2. Related work

Collaborative filtering (CF) techniques are one of the most important and widely used personalized recommender techniques. With the prospect of collective intelligence, CF is able to produce more accurate personalized recommendations. CF approaches can be divided into neighborhood-based ones [2] and model-based ones [3,4]. The latent factor model, such as the matrix factorization (MF) technique, is one of the most successful models in model-based approaches. In its basic form, the matrix factorization characterizes both items and users by vectors of factors inferred from item rating patterns. High correspondence between items' and users' factors leads to a recommendation. It is widely used in industrial applications because of its excellent performance, good scalability and flexibility [3]. However, when the ratings of users or items are sparse, the performance of CF approaches often becomes rather poor. Meanwhile, the problem of cold-start [5] also restricts personalized recommendation systems.

Recently, with the fast growth of user-generated online reviews, many techniques related to review analysis have been developed. The task of review analysis includes aspect discovery and sentiment analysis. A widely used aspect discovery method is extracting a set of frequently occurring noun phrases (NP) as aspect candidates and the subsequent retention of only the relevant ones by applying various filtering methods [9]. Some other approaches apply topic modeling to discover different aspects in reviews. For example, some works apply the Latent Dirichlet Allocation (LDA) [7] model by using sentences instead of documents [10,11]. Meanwhile, sentiment analysis, also known as opinion mining, mainly studies the opinion, evaluation, attitude, and emotion of individuals, problems, events, and topics of people [12,13]. Popular sentiment analysis approaches transform the sentiment analysis problem to the classification problem, and employ machine learning techniques to classify the sentiment [14]. Furthermore, many have proposed unified models of topics and sentiments [15–17]. For example, Y. Jo et al. present a model named Aspect and Sentiment Unification model (ASUM) that incorporates aspect and sentiment together to model sentiments with respect to different aspects, and is able to discover pairs of aspect, sentiment called senti-aspects [8].

A few works consider utilizing the review texts to improve the accuracy of rating prediction. McAuley et al. [6] introduce the Hidden Factors as Topics model (HFT), which combines the latent factor model [3] and LDA [7]. Specifically, they propose a transformation that links the stochastic vector obtained from LDA with the real valued vector in the latent factor model. HFT demonstrates the significant improvement over baseline methods that use ratings or reviews alone. It is also noteworthy that in [18] Q. Diao et al. introduce a probabilistic model based on collaborative filtering and topic modeling, whereas in [19] Yao Wu et al. propose a unified probabilistic model that