Engineering Applications of Artificial Intelligence 45 (2015) 281–294

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Collaborative recommendation with user generated content

Yueshen Xu, Jianwei Yin *

College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310027, China

ARTICLE INFO

Article history:
Received 31 March 2015
Received in revised form
7 July 2015
Accepted 13 July 2015
Available online 31 July 2015

Keywords:
Recommender system
User generated content
Collaborative filtering
Topic modeling
Parameter estimation

ABSTRACT

In the age of Web 2.0, user generated content (UGC), such as user review and social tag, ubiquitously exists on the Internet. Although there exist different kinds of UGC in recommender systems, the existing works only studied a single kind of UGC in each of their papers. Thus, the previous works lose a chance to uncover the similar effects of different kinds of UGC in recommender systems. In this paper, we propose a unified way to utilize various types of UGC to enhance the recommendation accuracy. We build two novel statistical models, which are based on collaborative filtering and topic modeling. Incorporating UGC text, one model focuses on learning user preferences, and the other model aims to learn user preferences and item aspects jointly. With an effective parameter estimation algorithm, our models can not only acquire prediction values of missing ratings, but also produce interpretable topics. We conducted comprehensive experiments on three real-world datasets. The experimental results demonstrate that our proposed models can achieve large improvements compared to several well-known baseline models.

1. Introduction

The recommender system has been an indispensable component in a modern e-commerce system, which aims to gauge a user's preference. Collaborative filtering (denoted by *CF*)-based and content-based algorithms are two prevalent approaches to the recommendation problem. Among all the CF-based models, models based on matrix factorization (denoted by *MF*) are widely used in industry and academia (Koren et al., 2009). CF-based algorithms are based on the following assumption: users that had similar preferences before tend to give similar ratings to the same item in the future (Su and Khoshgoftaar, 2009). The key task of CF-based algorithms is to find similar users and infer a user's preference from the past rating records. The content-based recommender system also plays an important role in many practical scenarios (Pandora, 2015; Rovi, 2015). In such approaches, an item's and a user's profiles are collected to predict whether the user likes the item or not (Lops et al., 2011). Meanwhile, user generated content widely exists on the Internet, and has been utilized in recommender systems, such as tag-based and review-based recommendation algorithms (Liang et al., 2010; McAuley and Leskovec, 2013). Formally, user generated content (denoted by *UGC*) refers to the content created by users on the Internet (Moens et al., 2014), such as social tag, review, blog and tweet.

The utilization of side information to improve the recommendation accuracy is a popular trend in recent years, such as social relationship, implicit feedback and context (Ma et al., 2011; Yang et al., 2012; Liu and Aberer, 2013; Ostuni et al., 2013; Tang et al., 2013; Chen et al., 2014a). They assume that a user's preference can be inferred through his or her friends' preferences, or can be influenced by certain context. Although such side information can indeed help improve the recommendation accuracy, many e-commerce web sites do not have the mentioned side information. For example, in Amazon,[1] Ebay[2] and Expedia,[3] there are no social relationships. However, they do have user reviews. For building a model with high applicability, we have to infer a user's preference from the available information sources, such as rating records and UGC text.

In recent years, several works tried to utilize an item's description to learn the item's associated latent aspects, and make better recommendation (Agarwal and Chen, 2010; Wang and Blei, 2011; Zhang and Wang, 2014). More related works will be discussed in Section 2. The existing models are verified to achieve higher prediction accuracy than the traditional CF-based algorithms that only rely on the past rating records. However, there are still three problems in the existing models.

* Corresponding author. Tel.: +86 13396811588.
E-mail addresses: xyshzjucs@zju.edu.cn (Y. Xu), zjuyjw@zju.edu.cn (J. Yin).

http://dx.doi.org/10.1016/j.engappai.2015.07.012
0952-1976/

[1] http://www.amazon.com/
[2] http://www.ebay.com/
[3] http://www.expedia.com/

1. An item's description lacks distinctiveness since many words in the descriptions of different items tend to overlap. For example, the product descriptions of two cellphones like *Samsung Galaxy S5* and *iPhone* are overlapping largely with each other. The contained words are usually about product features, such as *battery*, *price* and *size*.
2. An item's description is static, and the description text can only be edited by website editors. Thus, the word space, which is constructed by the item descriptions, is usually very sparse. The high sparsity lowers the accuracy of the latent aspects.
3. An item's description is independent of a user's preference, so it is difficult to infer the user's preference from the description text.

In contrast, UGC text can emphasize an item's outstanding features. For example, the song *You belong with me* receives many following tags in Last.fm,[4] including *love*, *country*, *pop* and *Taylor Swift*, which precisely highlight this song's aspects. UGC can reflect a user's interested topics as well. For example, if *adventure*, *space* and *science* are frequently used by Bob in his reviews, we can infer that Bob may like the movies *Star Trek* and *Interstellar* since the two movies have the related themes. Also, the UGC text opens a gate for us to learn the reason why a user likes or dislikes an item. For example, after purchasing a computer, if Baron writes a review "I like its color and its light weight. It's easy to carry", it gives us a hint that we can recommend Baron other electronic devices with the same color and light weight. Besides, the UGC text increases along with time, so the amount of text information is abundant. Also, some kinds of UGC, such as social tag, provide a direct way to acquire meaningful phrases without chunking (for more details about chunking in NLP, refer to Manning and Schütze, 1999). For example, in Last.fm, the song *Freezing Moon* receives *black metal* (a tag) for many times, which precisely depicts this song's genre. Clearly, *black metal* should be regarded as an integrated phrase. If not, this tag will be mistakenly regarded as a simple combination of a color term (*black*) and a material term (*metal*).

Although in some previous works (Liang et al., 2010; Zanardi and Capra, 2011; McAuley and Leskovec, 2013; Ling et al., 2014), UGC has been verified to be helpful to improve the prediction accuracy in recommender systems, there are still three challenging problems:

1. Not every web site allows a user to give tags to products, and some web sites even do not give the chance to write reviews. For example, in Bestbuy,[5] and Newegg,[6] users are not allowed to tag. In Movielens,[7] users are not allowed to write reviews.
2. The existing works only study a single kind of UGC in each paper, and leverage different kinds of UGC in very different ways. This makes it difficult to uncover the similar effects which are shared among diverse kinds of UGC in recommender systems.
3. There exist many semantic problems in UGC text, like synonym and acronym. However, some traditional methods ignore these problems without considering the semantic correlation among words.

To deal with the first and second problems, it is necessary to propose an approach that is capable of leveraging different types of UGC in the same way. To deal with the third problem, we need to employ effective techniques, such as topic modeling, to relieve the semantic problems in UGC text.

To tackle the issues mentioned above, in this paper, we study different types of UGC, and incorporate them into the recommender system in a unified manner. Because reviews and tags are two kinds of widely existing UGC, in this paper, we focus on user reviews and social tags. We propose our first model to learn a user's interested topics from his or her UGC text. Most of the previous review-based models only focus on learning item aspects. We want to investigate the function of UGC in learning a user's preference. Then, we propose the second model which learns a user's preference and an item's aspects jointly. Different from the item description-based algorithms, our proposed model learns the aspects from the collection of the UGC text that an item receives. Both of the proposed models are statistical models and have well-designed generative processes. We design an effective iterative algorithm to learn the model parameters. We conduct comprehensive experiments on three real-world datasets to attest the effectiveness of our models. We also investigate our models' performance in the "cold-start" scenario, which means a user has limited or even very few ratings. The "cold-start" problem is a very challenging issue since the traditional CF-based models cannot make satisfactory recommendation with insufficient data.

The main contributions of this paper are summarized as follows:

1. It studies the effect of UGC in learning a user's preference and an item's aspects. It finds that reviews and social tags are two valuable information sources to learn user preferences and item aspects.
2. It proposes two novel statistical models, both of which can take UGC text and rating records together as input. The first model focuses on learning users' interested topics. The second model extends the first model, and takes users' and items' topics/aspects learning task into consideration jointly.
3. It proposes a parameter estimation algorithm for the proposed models, which is verified to be effective by the experiments.
4. It conducts sufficient experiments on three public real-world datasets, which demonstrate that our proposed models outperform all baseline models. It also expands one of the datasets by crawling more contents. We will release the expanded dataset publicly to facilitate related research in the community.

The rest of this paper is organized as follows: Section 2 summarizes the related work. Section 3 gives a concise explanation of two base models. Section 4 explicates the details of our proposed models. Section 5 shows the experimental results along with a thorough analysis. Lastly, Section 6 concludes the paper and discusses the future work.

2. Related work

In recent years, in recommender systems, it has been a main stream to introduce side information to enhance the performance of traditional collaborative filtering models. The typical side information includes social relationship, social tag, item description, etc. Our work is closely related to tag-based recommendation, review-based recommendation and item description-based recommendation (Agarwal and Chen, 2010; Liang et al., 2010; Gemmell et al., 2011; Ling et al., 2014).

Tag-based recommendation. There have been some works that employ social tags to build the connection between users and items (Liang et al., 2010; Gemmell et al., 2011; Zanardi and Capra, 2011). Liang et al. (2010) proposed a tag-based recommendation algorithm. Each item was attached with a tag vector as this item's profile, which was used to calculate the similarity between each pair of items. Each user also had a profile, which was constructed from his or her tagging record and used to calculate the similarity between each pair of users. A score for each item in the candidate list was computed for top-*N* recommendation. Since not every web site has tags, this model's

[4] http://www.last.fm
[5] http://www.bestbuy.com/
[6] http://www.newegg.com/
[7] https://movielens.org/