Knowledge-Based Systems 97 (2016) 111–122

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

# Collaborative Topic Regression with social trust ensemble for recommendation in social media systems

Hao Wu*, Kun Yue, Yijian Pei, Bo Li, Yiji Zhao, Fan Dong

*School of Information Science and Engineering, Yunnan University, No. 2 North Green Lake Road, Kunming 650091, China*

ARTICLE INFO

*Article history:*
Received 9 May 2015
Revised 8 January 2016
Accepted 9 January 2016
Available online 16 January 2016

*Keywords:*
Social media systems
Item recommendation
Trust ensemble
Matrix factorization
Recommender systems

ABSTRACT

Social media systems provide ever-growing huge volumes of information for dissemination and communication among communities of users, while recommender systems aim to mitigate information overload by filtering and providing users the most attractive and relevant items from information-sea. This paper aims at providing compound recommendation engine for social media systems, and focuses on exploiting multi-sourced information (e.g. social networks, item contents and user feedbacks) to predict the ratings of users to items and make recommendations. For this, we suppose the users' decisions on adopting item are affected both by their tastes and the favors of trusted friends, and extend Collaborative Topic Regression to jointly incorporates social trust ensemble, topic modeling and probabilistic matrix factorization. We propose corresponding approaches to learning the latent factors both of users and items, as well as additional parameters to be estimated. Empirical experiments on Lastfm and Delicious datasets show that our model is better and more robust than the state-of-the-art methods on making recommendations in term of accuracy. Experiments results also reveal some useful findings to enlighten the development of recommender systems in social media.



## 1. Introduction

Social media is a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0 [14], and allow content generation, dissemination and communication among communities of users [34]. Social media systems have a wide range of potential impacts on both application and research perspectives, for example, promoting sales for profit business, facilitating creativity, interaction and learning of individual users for information consumption. However, the abundance and popularity of social media sites flood users with huge volumes of information and hence present a significant challenge in terms of information overload. Too much information may make users helpless in their process of finding useful contents.

Recommender Systems (RS) aim to mitigate the negative impact of information overload by filtering and providing users the most attractive and relevant items (such as photos, videos, music, articles, news, comments, tags, people, etc.) from information-sea. RS often use personalization techniques tailored to the needs and interests of the individual user, or the collective intelligence. And various techniques, such as neighborhood-based collaborative filtering [25], matrix factorization [16], network-based approaches [12], and fuzzy-based collaborative filtering [36] have been made to automatically predict the interests of a particular user [3,17].

Social media and recommender systems can mutually benefit for each other [8]. On one hand, social media introduces ever-growing rich information, such as tags, ratings, comments, and friendship of users, which can be used to strengthen recommendations; On the other hand, recommender technologies can increase adoption, engagement, and participation of new and existing users, resulting in the success of social media applications. Consequently, how to utilize rich information in social media to enhance recommendation models, has become a hot issue of great interest to both academia and industries [7,8].

Existing works can be divided into two categories. One focuses exploiting item-specific contents, such as tags, comments, link-relations, which can effectively deal with the cold-start recommendation for items; The other emphasizes the usage of user-specific information (mainly the trust relationship among users), which can effectively alleviate the cold-start recommendation for users. However, these works have utilized either content of items or social information of users, and few have considered them jointly. Naturally, a question will be asked whether and how the contents of items and social networks of users can be combined together to produce compound recommendation engines, and thus

* Corresponding author. Tel.: +86 18687001878.
*E-mail addresses:* haowu@ynu.edu.cn (H. Wu), kyue@ynu.edu.cn (K. Yue), pei3p@ynu.edu.cn (Y. Pei), libo@ynu.edu.cn (B. Li), yijizhaoneo@gmail.com (Y. Zhao), donnfan117@gmail.com (F. Dong).

http://dx.doi.org/10.1016/j.knosys.2016.01.011

have the best of both worlds. In this regard, there have been several exploratory works. Ye et al. [33] proposed a PLSA-like [10] probabilistic generative model, called unified model, which naturally unifies the ideas of social influence, collaborative filtering and content-based methods for item recommendation. Expectation-maximization algorithm is devised to learn the model parameters. However, the model can only exploit observed ratings, zero or unobserved ratings which may potentially reflect users' interests cannot be exploited under this algorithmic framework, therefore may dramatically impact the predictive performance. CTRSMF [22] integrates Collaborative Topic Regression (CTR) with social matrix factorization models (SMF), to build a hybrid recommender system. Leveraging on matrix factorization techniques, CTR can deal with both explicit and unobserved ratings, by putting different confident numbers for them, and make CTR-based recommendation systems more powerful. Owing to the lack of physical explanation, directly factorizing social trust matrix in CTRSMF does not reveal the underlying relations among the users. Different from CTRSMF which utilizes homophile effect in social media to smooth the similarity of users' interests to their friends, LACTR [13] directly learns how much attention users allocate to other users, and leverages these learned influence to make recommendations. LACTR is a sophisticated model, but it implicitly presets a strong condition that users' social interactions usually follow topically similar contents. Although such condition may be satisfied easily in discussion-thread-oriented systems (e.g. Digg), it is not always accurate, resulting in that LACTR may be sensitive to different datasets.

To deal with existing problems and exploit both social factors and content information for enhancing item recommendation, in this paper, we present a novel generative model CTRSTE by extending CTR which naturally incorporates content information via Latent Dirichlet Allocation [2] into collaborative filtering framework [27]. Different from previous studies, we suppose the users' ratings on items are affected both by the personal tastes and their trusted friends' favors, and naturally integrate this principle into the CTR model. We propose parameter learning method to infer latent factors both for users and items in the new model. Our experiments on large-scale datasets from Lastfm and Delicious show that our CTRSTE model significantly outperforms the state-of-the-art variants of CTR in term of accuracy.

The main contribution of this paper consists of two folds. First, we extend CTR with trust ensemble principle for item recommendation in social media and show its effectiveness on two large-scale datasets. It sets up a new algorithmic framework to seamlessly exploit user–item feedback, item content, and social network for building powerful recommender engines. Second, we compare social-enhanced variants of CTR from separate aspects of recommendation quality, such as the accuracy, diversity, novelty and then show their strengths and weaknesses on item recommendation in social media systems. All these can contribute to the practices of recommender systems.

The remainder of this paper is arranged as follows: in Section 2, we provide an overview of related works on recommendation systems and Collaborative Topic Regression models. In Section 3, we propose our CTRSTE model and discuss how to learn parameters and do inference. The experimental results and discussion are presented in Section 4, followed by conclusions and future works in Section 5.

## 2. Related works

### 2.1. Social-based collaborative filtering

Collaborative filtering (CF) has grown up to be a hot research topic due to its successful application in the recommendation systems. In traditional CF methods, only the feedback matrix, which contains either explicit(e.g., ratings) or implicit feedback(e.g., tagging, clicks, purchases) on the items given by users, is used for training and prediction. Typically, the feedback matrix is sparse, which means that most users come into contact with a few items. Resulting from this sparsity problem, traditional CF with only feedback information will suffer from unsatisfactory performance.

Recently, many researchers have proposed utilizing auxiliary information to alleviate the data sparsity problem in CF. Among them, the most popular approaches focus on exploiting social information as various social media systems are booming in Internet [18–20,32,35]. These methods increase the accuracy of traditional CF by taking collective interests and social trusts between users in an online social network as additional inputs. Social trust between a pair of friends $(u, v)$ may be established based on explicit feedback of user $u$ concerning user $v$ (e.g., by voting or following), or it may be inferred from implicit feedback (e.g., the mutually shared resources/items between $u$ and $v$). However, different algorithms explore social networks and the embedded social information differently. Ma et al. [18–20] have proposed three ways to integrate social information with matrix factorization process, namely SoRec, Social Trust Ensemble(STE) and social regularization. In SoRec [19]. The user–item feedback matrix and the user–user social matrix are simultaneously factorized by using shared user latent factors. In STE [18], the predicted rating for item $i$ by user $u$ is a linear combination of three terms: a global offset of ratings, prediction based on $u$'s and $i$'s latent factors, and a weighted sum of the predicted ratings for item $i$ from all of user $u$'s friends. Social regularization model addresses the transitivity of trust in social networks, and exploits the social circles and users' latent factors to create a term to regularize the matrix factorization process [20,35]. All of three models achieve better prediction accuracy than the original matrix factorization. Shambour and Lu [24] proposed a trust-semantic fusion-based recommendation approach for B2B e-service, where user-based trust-enhanced CF and item-based semantic-enhanced CF are fused to utilize trust intuitive property. By this it can alleviate the data-sparsity problem associated with users and items. Besides, some algorithms predict a user's rating for an item by traversing the user's neighborhood and querying the item ratings of her/his direct and indirect friends [32], e.g. Filmtrust [5], MoleTrust [21], TrustWalker [12].

### 2.2. Collaborative Topic Regression models

Beyond exploiting social information, other methods often utilize item content information to enhance collaborative filtering. Collaborative Topic Regression (CTR) is one of these methods which have achieved promising performance by successfully integrating both feedback information and item content information [27]. The CTR model combines the merits of both probabilistic matrix factorization and topic modeling approaches. Here, we gradually restate the background approaches constructing the CTR model.

*2.2.0.1. Probabilistic matrix factorization.* In matrix factorization, users and items are both represented as latent vectors in a shared latent $K$-dimensional space, $\mathbb{R}^K$, where user $i$ is represented as a latent vector $u_i \in \mathbb{R}^K$ and item $j$ is represented as a latent vector $v_j \in \mathbb{R}^K$. The prediction of whether user $i$ will like item $j$ is given by the inner product between their latent representations, $r_{ij} = u_i^T v_j$. To use matrix factorization for collaborative filtering, the latent representations of the users and items must be learned given an observed matrix of ratings. The common approach is to minimize the regularized squared error loss with respect to user