Information Systems 42 (2014) 59–77

Contents lists available at ScienceDirect

Information Systems

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

TWILITE: A recommendation system for Twitter using a probabilistic model based on latent Dirichlet allocation

Younghoon Kim, Kyuseok Shim*

Department of Electrical and Computer Engineering, Seoul National University, Kwanak P.O. Box 34, Seoul 151-600, Republic of Korea

ARTICLE INFO

Article history:
Received 7 November 2013
Received in revised form
24 November 2013
Accepted 26 November 2013
Recommended by: D. Shasha
Available online 4 December 2013

Keywords:
Twitter
Recommendation system
Probabilistic model
Collaborative filtering
LDA
Matrix factorization

ABSTRACT

Twitter provides search services to help people find users to follow by recommending popular users or the friends of their friends. However, these services neither offer the most relevant users to follow nor provide a way to find the most interesting tweet messages for each user. Recently, collaborative filtering techniques for recommendations based on friend relationships in social networks have been widely investigated. However, since such techniques do not work well when friend relationships are not sufficient, we need to take advantage of as much other information as possible to improve the performance of recommendations.

In this paper, we propose *TWILITE*, a recommendation system for Twitter using probabilistic modeling based on latent Dirichlet allocation which recommends top-K users to follow and top-K tweets to read for a user. Our model can capture the realistic process of posting tweet messages by generalizing an LDA model as well as the process of connecting to friends by utilizing matrix factorization. We next develop an inference algorithm based on the variational EM algorithm for learning model parameters. Based on the estimated model parameters, we also present effective personalized recommendation algorithms to find the users to follow as well as the interesting tweet messages to read. The performance study with real-life data sets confirms the effectiveness of the proposed model and the accuracy of our personalized recommendations.

1. Introduction

Twitter has emerged recently as a new medium in spotlight for communication. Twitter offers a unique mechanism of information diffusion by allowing each user to receive all messages (called tweets) from those whom he follows. We refer to those who follow a user as *followers* and refer to those whom a user follows as *followees*.

Generally, users would like to follow other users who post interesting messages. However, assisting users to find new people to follow is not a simple task. Twitter itself provides a service to help people find the users to follow by recommending popular users or the friends of their friends. However, these services do not offer the most relevant users to follow for a user. Furthermore, recommending the most interesting tweet messages for a user will be very useful, but Twitter does not provide this functionality.

There has been much work done on developing new approaches for recommendation systems over the last decade. The interest in the area still remains high because personalized recommendations have many practical applications (e.g., shopping cart recommendations of Amazon, MovieLens and AdaptiveInfo.com). Moreover, with the recent fast growth of Social Network Services (SNSs) such as Twitter and Facebook, the problem of recommending other users or other users' posted messages to a user with common interests has received a lot of attention.

* Corresponding author. Tel.: +82 2 880 7269.
E-mail address: shim@kdd.snu.ac.kr (K. Shim).

0306-4379/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.is.2013.11.003

Many recommendation systems such as [7,20,24] take the approach of collaborative filtering which recommends friends or articles based on the pattern of the other users who have connected to a similar set of friends in the past. However, if a user is new to the system and thus has little friends, we need to take advantage of not only the friend relationships but also the content of their messages for a better recommendation.

Due to the distinctive feature of Twitter which allows users to post their Twitter messages and to follow others, we have more opportunities for better recommendations than the other traditional applications. For example, even for the users with few Twitter messages, we can provide better recommendations using the Twitter messages of their followees. To take advantage of such a distinctive feature of Twitter, we previously proposed *TWITOBI*, a recommendation system for Twitter using probabilistic modeling for collaborative filtering in [22].

The probabilistic model used in *TWITOBI* was a generalization of the probabilistic latent semantic indexing (PLSI) in [16]. The model assumes that topics are selected by following not only the user's preference to the topics but also the preference of the user's followees to the topics. An Expectation–Maximization (EM) algorithm was next developed to learn the parameters of our model by maximizing the log-likelihood of expectation. The performance study with Twitter data showed the effectiveness and scalability of our algorithms for Twitter. However, our previous PLSI model does not capture the generative process of establishing friend relationships in Twitter. Furthermore, it is known in [6,29] that PLSI suffers from the overfitting problem compared to the latent Dirichlet allocation [5].

In this paper, we propose a new probabilistic model which can capture the generative process of posting tweet messages as well as establishing friend relationships. It is reported that users in social network have a strong tendency to follow other users who have many common friends [2]. Since the matrix factorization is one of the most successful models for describing the connections between users by assuming that a user has a high chance to be a friend with the ones who have many common friends [27,40,42], we exploit the friend relationships between users using matrix factorization to enhance the accuracy of recommendations for Twitter. In our model, we seamlessly combine the probabilistic model based on LDA to model the process of posting tweet messages and the collaborative filtering based on matrix factorization to capture the process of establishing friend relationships. We next develop an inference algorithm that utilizes the variational EM algorithm and the matrix factorization.

To learn model parameters for LDA models, Gibbs sampling [13] or variational EM algorithms has been widely used. Since it is simpler to draw an inference algorithm using Gibbs sampling rather than a variational EM algorithm, many probabilistic models based on LDA utilize Gibbs sampling. However, Gibbs sampling methods usually require many iterations to find a solution resulting in too slow convergence. Furthermore, since Gibbs sampling does not guarantee the convergence of log-likelihood, it is hard to determine when to stop the iterations [19]. On the other hand, deriving a variational EM algorithm is sometimes difficult depending on models but it converges faster and guarantees the convergence. In this paper, we thus develop a variational EM algorithm for our LDA model.

The contributions of this paper are as follows:

- We propose a novel probabilistic generative model which is suitable for representing the activities in Twitter. Our model represents a realistic process of posting tweet messages by exploiting the LDA model as well as the process of connecting to friends by utilizing matrix factorization.
- We develop an inference algorithm which utilizes a variational EM algorithm with matrix factorization to learn the posterior distributions and model parameters of our LDA model.
- We also provide the ranking algorithms for recommending top-K followees or top-K tweets to a user using the estimated posterior distributions and model parameters.
- By performance study with Twitter data, we show the effectiveness of our model as well as the accuracy of the top-K followee and tweet recommendation algorithms for Twitter.

The rest of this paper is organized as follows. After discussing related work in Section 2, we provide our problem formulations as well as the preliminary works in Section 3. We next propose a probabilistic generative model in Section 4 and develop an inference algorithm using a variational EM algorithm with matrix factorization for our model in Section 5. Section 6 presents how to utilize our model parameters for recommendations. Finally, the performance study is given in Section 7 and we summarize the paper in Section 8.

2. Related work

We first discuss the model-based collaborative filtering algorithms [15,17,28,31,32,36,41] and next describe the recent works on recommendations for social networks [27,40,42].

Model-based algorithms build their models to describe the behaviors of users using training data and utilize the trained models to predict the users' preference on the items unseen in the training data. Examples of model-based approaches include the probabilistic models [5,22], random walk models [11,41], latent factor models [18,30] and combined ones [36] of probabilistic and latent factor models.

A naive recommendation method which can be used for Twitter is to recommend users or contents based on similarity. In [9,14], TF-IDF weighting is used for recommending users to follow and tweet messages to read based on the cosine similarity between their friends and tweet messages. Even if those algorithms consider the tweet messages as well as followers and followees, a simple use of TF-IDF weighting makes the algorithms to suffer the quality of recommendations.