Structural-Interaction Link Prediction in Microblogs

Yantao Jia
Institute of Computing Technology, CAS
Beijing, P. R. China
jiayantao@ict.ac.cn

Yuanzhuo Wang
Institute of Computing Technology, CAS
Beijing, P. R. China
wangyuanzhuo@ict.ac.cn

Jingyuan Li
Institute of Computing Technology, CAS
Beijing, P. R. China
lijingyuan@ict.ac.cn

Kai Feng
Institute of Computing Technology, CAS
Beijing, P. R. China
fengkai@ict.ac.cn

Xueqi Cheng
Institute of Computing Technology, CAS
Beijing, P. R. China
cxq@ict.ac.cn

Jianchen Li
North China Electric Power University
Beijing, P. R. China
moning@gmail.com

ABSTRACT

Link prediction in Microblogs by using unsupervised methods aims to find an appropriate similarity measure between users in the network. However, the measures used by existing work lack a simple way to incorporate the structure of the network and the interactions between users. In this work, we define the retweet similarity to measure the interactions between users in Twitter, and propose a structural-interaction based matrix factorization model for following-link prediction. Experiments on the real world Twitter data show our model outperforms state-of-the-art methods.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms,Performance

Keywords

Link prediction, Microblogs, Structure-interaction

1. INTRODUCTION

The link prediction in Microblogs such as Twitter has been extensively studied during recent years. Although link prediction in Microblogs faces the challenge to build a unified framework to balance the social aspect and the information aspect of the Microblogs, the common methodology used in social networks is still instructive, which can be classified into two parts: the supervised methods and the unsupervised methods. Supervised methods treat the link prediction as a classification problem, but they often suffer from the imbalance and feature selection problem. In contrast, the unsupervised methods do not need to know the prior knowledge of the distribution of the data set and can avoid the drawbacks of the supervised methods. The unsupervised methods intend to define a statistics to measure the similarity between two users, such as common neighbors, Jaccard coefficients , Katz measure , etc. Very recently, Yin et al. [2] defined the structure similarity measures between two users with respect to another user, and proposed a structure based matrix factorization model (S-Model) for link prediction in Microblogs. They discovered that the model achieved higher F1-measure than that obtained by other seven measures such as the Jaccard coefficient and so on. For example, the F1-measure of S-model equals 0.197 in dynamic setting with an increase of 0.03 compared with the best method.

Although the S-Model gets a higher F1-measure, it does not consider the impact of the interaction between users on link formation. To this end, we propose an unsupervised method, the *structural-interaction model*(SI-Model), which integrates the structural information and the interaction information between users to predict future links. This idea comes from the observation that interaction between users correlates with the link formation in Twitter. More precisely, we define the retweet similarity to measure the similarity between two users. Then we establish an objective function consisting of the “interaction regulation” term in connection with the retweet similarity. Minimizing the function via the nonnegative matrix factorization leads to a method for link prediction. Experiments based on the real Twitter data show SI-Model outperforms state-of-the-art methods by reducing the rmse value by about 70% on average compared with that obtained by the best method.

2. THE SI-MODEL

In this section, we shall propose a nonnegative matrix factorization based model, called the SI-Model to unify the structure of the network and the interactions between users to predict new links of a given user. The problem can be formulated as follows: for a given source user v_u, we aim to find a list of target users v_i via a list of intermediate user v_k, and select the top N target users to recommend to v_u. Let $R_{n\times m} = (R_{ui})$ be the rating matrix, where n is the number of source users and m is the number of target users, $R_{ui} = 1$ if v_u follows v_i and $R_{ui} = 0$ otherwise. The matrix factorization method is to factorize the matrix R into two latent matrices $A_{n\times K}$ and $B_{K\times m}$ such that $R_{ui} = \sum_{k=1}^{K} a_{uk}b_{ki}$. We shall define the retweet similarity between two intermediate users v_k and $v_{k'}$ and the objective function $F(A, B)$ which SI-Model aims to minimize. Firstly, we define the retweet similarity between two intermediate users v_k and $v_{k'}$ based

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

on their interactions with one target user v_i respectively in the time interval $(t_0, t_1]$, denoted by $R_i(k, k')$. The interactions are referred to as the retweet behaviors. Suppose that v_i posted a list of s tweets $\{tw_1, tw_2, \ldots, tw_s\}$ in the time interval $(t_0, t_1]$. There are two ways to define $R_i(k, k')$. One is to compare the number of retweets of v_k and $v_{k'}$. Assume that v_k retweets n_k tweets of v_i and $v_{k'}$ retweets $n_{k'}$ tweets of v_i. Then $R_i(k, k')$ can be defined in a binary way: $R_i(k, k') = 1$ if $n_k = n_{k'}$ and $R_i(k, k') = 0$ otherwise. The other is to define a refined vector to record for v_k according to whether v_k retweet each of the s tweets of v_i as $r_k = [r_{k1}, r_{k2}, \ldots, r_{ks}]$, where $r_{ki} = 0$ if v_k does not retweet the i-th tweet, and $r_{ki} = 1$ otherwise. Similarly, we can get the refined vector for $v_{k'}$ as $r_{k'} = [r_{k'1}, r_{k'2}, \ldots, r_{k's}]$. Then $R_i(k, k')$ can be defined as the cosine similarity of the two vectors r_k and $r_{k'}$, that is, $R_i(k, k') = (r_k \cdot r_{k'})/(\|r_k\| \cdot \|r_{k'}\|)$. With the retweet similarity, we introduce the interaction regulation term $R(A)$ as $R(A) = \sum_{u=1}^{n} \sum_{k=1}^{l} \sum_{k'=1}^{l} R_i(k, k')(a_{uk} - a_{uk'})^2 / \sum_{u=1}^{n} \sum_{k=1}^{l} \sum_{k'=1}^{l} R_i(k, k')$, where n is the number of target users and l is the number of intermediate users. The SI-Model aims to minimize the objective function $F(A, B) = \frac{1}{2} \sum_{A,B} I_{u,i}(R_{ui} - \sum_{k=1}^{K} a_{uk} b_{ki})^2 + \frac{\lambda_1}{2} \|A\|^2_{\text{Fro}} + \frac{\lambda_1}{2} \|B\|^2_{\text{Fro}} + \lambda_2 S(A) + \lambda_2 S(B) + \lambda_3 R(A)$, where λ_3 is a nonnegative parameter called the interaction regulation parameter, $\| \cdot \|^2_{\text{Fro}}$ denotes the Frobenius norm and $S(\cdot)$ is the structural regulation function introduced by Yin [2]. To solve the model, we follow the multiplicative update rule by Lee and Seung [1].

3. EXPERIMENT RESULTS

In this section, we describe the prediction result. The data is crawled by Twitter API by randomly selecting 10000 Twitter users, update their immediate neighbors per day from the period of Oct. 1st 2012 and Nov. 19th. This leads to the user networks. Meanwhile, we extracts the tweets of these users per day and use them to construct the retweeting network, where user A have relations with user B if A's tweet contains the syntax @B or RT@B, or equivalently, A retweets B or mentioned B in his tweets. In total, there are 140,000 users and 400,000,000 tweets. To conduct our experiment, we randomly select 1000 pairs of snapshots of the data set, and use the first snapshot to predict the following links in the second snapshot. The interval between these two snapshots is chosen as one week. Note that the interval can be chosen differently, for instance, two weeks and so on. Our model runs on the matrix R with 10000 rows and 10000 columns. Two evaluation criterions of the predicting result are used, the Root Mean Square Error (RMSE) and the F1-measure based on the breakeven point. To evaluate the performance of the SI-Model, we tune the parameters λ_1, λ_2 and λ_3 in the full grid, where three parameters range from 0 to infinity. After a full search, we find that when $\lambda_1 = 0.01$, $\lambda_2 = 0.01$, the optimal RMSE value of the S-Model is 0.102. Similarly, when $\lambda_1 = 0.01$ and $\lambda_2 = 0.001$, and set $\lambda_3 = 0.005$, the optimal RMSE value of the SI-Model is 0.033. The following table lists the comparison of the result by SI-model with those by other three methods, the S-Model, the Jaccard coefficient (JC) and the common neighbors (CN). Note that for the JC and CN methods, there are no RMSE by definition.

Table 1: The comparison for different methods

Methods	RMSE	F1-measure
SI-Model	0.033	0.278
S-Model	0.102	0.252
Jaccard coefficent	\	0.125
Common neighbors	\	0.091

From Table 1 we see that SI-Model achieves smaller RMSE value and bigger F1-measure than any of the other three model. Especially, the rmse value is reduced by about 0.07 compared with that obtained by S-Model. Note that the S-Model obtained the RMSE value 0.102. If we aims to reduce it, the maximal reduction is 0.102 (corresponding to the value 0). In other words, we get 70% reduction by using the SI-Model. On the other hand, our SI-Model get the F1-measure 0.278, with the increase 0.026 compared to the S-Model. Notice that Table 1 lists the average performance. For detailed comparison, we also conduct the experiment. For instance, as for the F1-measure, we illustrate the F1-measure of the S-Model and the SI-Model for 50 different snapshot pairs.

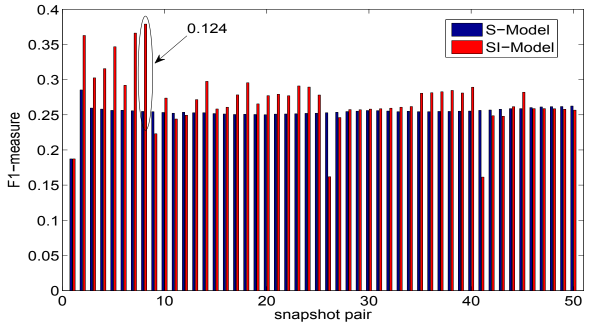

Figure 1: The F1-measure of S-Model and SI-Model

From Figure 1, we see that the SI-Model performs better than the S-Model for 72% snapshot pairs in which the second column is higher than the first. Especially, when $t = 8$, SI-Model get 0.124 increase. For the rest 28% snapshot pairs, we find SI-Model is not better because in these pairs, the retweet behavior between users does not correlate so much with the link formation process. To conclude, the average performance of the SI-Model is better than S-Model, and it is also competitive in most snapshots.

4. ACKNOWLEDGEMENT

This work is supported by National Grand Fundamental Research 973 Program of China (No. 2013CB329602), National Natural Science Foundation of China (No. 61173008, 61232010,61202213) and National Science Supported Planning (No. 2012BAH39B02).

5. REFERENCES

[1] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Proc. NIPS*, pages 556–562. ACM, December 2001.

[2] D. Yin, L. Hong, and B. D. Davison. Structural link analysis and prediction in microblogs. In *Proc. CIKM*, pages 1163–1168. ACM, October 2011.