# Collaborative Model for Predicting Retweeting Behaviors on Twitter

Liang Guo(✉), Zhaoyun Ding, Sheng Zhang, Taowei Li,
Weiwei Jiang, and Hui Wang

College of Information Systems and Management,
National University of Defense Technology,
Changsha 410073, Hunan, People's Republic of China
{guoliang,zyding,zhangsheng,litaowei,jiangweiwei,huiwang}@nudt.edu.cn

**Abstract.** Nowadays, Twitter has become one of the most important ways for information sharing. Users can spread information they like by retweeting. However, with the growth of twitter, users are easily overwhelmed by large amount of data and it is very diffcult for users to dig out information that they are interested in. To address this problem, we predict tweets that users are really interested in and help them reduce the effort to find useful information. In this paper, we introduce the users' similarity and trust based on retweeting behaviors and propose a retweeting behaviors prediction model based on collaborative filtering. The experiments show that our model was applicable on the real-life data.

**Keywords:** Retweeting behaviors · Predict model · Collaborative filtering · Twitter

## 1 Introduction

Recently, twitter has become a popular microblogging platform that enables users to share information by sending short 140-character messages. Until now, it has nearly 302 million users, who generates 500 million tweets per day. Users can get timeline of specific persons' tweets by following them. However, with the rapidly increasing number of tweets, users especially those follow many persons' are facing a serious problem of information overload. The useful tweets for the user may be flooded by other tweets that the user does not care about at all. So predicting the tweets that user cared so much is a huge challenge [1]. Then what kind of tweets a user cared so much? Intuitively, it is determined by many factors, such as the users' personal interest, the retweeting times, the content of the tweet, etc. Among these factors, the personal interest is the most important. Unlike existing work, which always analyzes the content of tweets and uses social relationship, we introduced the similarity and the trust between the users based on the retweeting behaviors.

Twitter has a retweet mechanism that is to repost a message from another twitter user and share it with one's own followers. It can accelerate the spread

R. Cai et al. (Eds.): APWeb 2015 Workshops, LNCS 9461, pp. 101–112, 2015.
DOI: 10.1007/978-3-319-28121-6_10

of information and show the users' personal interests [2]. As tweets are used many different languages and users' personal interests may change with times, traditional methods like analyzing the content are difficult to discover the users' interests [3]. So we proposed a predicting retweeting behaviors model based on collaborative filtering. As the Fig. 1 shown, our model is based on data collection and preprocessing; and its main components,including (1) Computing user-tweet interest based on trust, (2) Computing user-tweet interest based on similarity, (3) Computing user-tweet interest based on weighted interest.

*Data collection and preprocessing:* To address the problem, we used Twitter API to crawl tweets and users' information based on several randomly selected

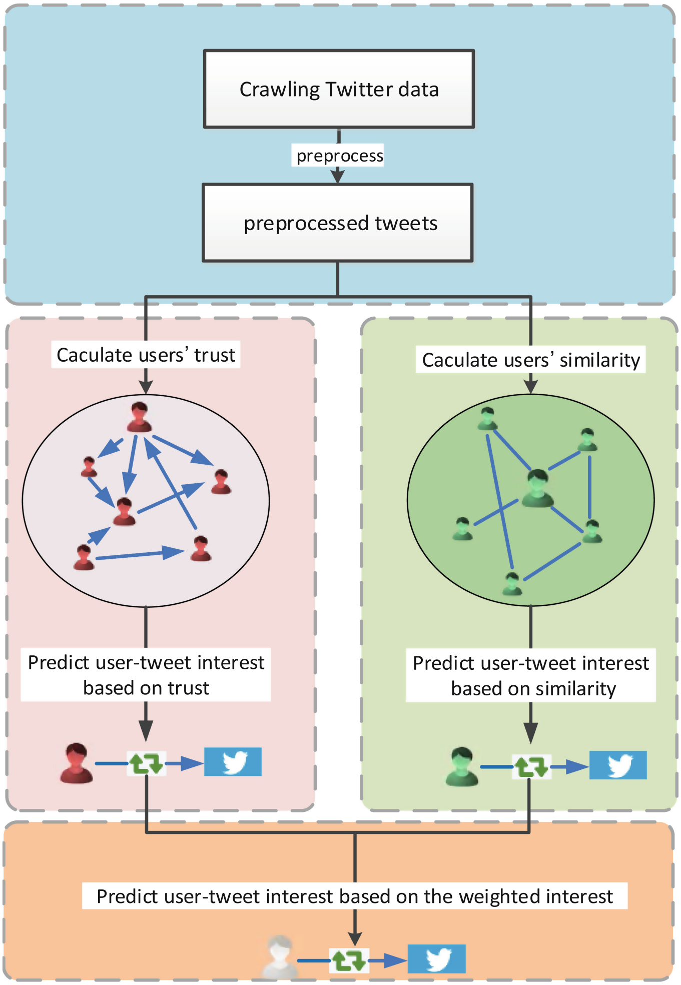

Crawling Twitter data
preprocess
preprocessed tweets
Caculate users' trust
Caculate users' similarity
Predict user-tweet interest based on trust
Predict user-tweet interest based on similarity
Predict user-tweet interest based on the weighted interest

**Fig. 1.** Predicting retweeting behaviors model

users and expand users by accessing their followers and followees links. In the preprocessing, we extracted retweeted users' screennames by matching $RT$@ and cut off the first 30 characters from the tweet text which removed retweeted users' screennames and mentioned users' screennames as the tag of the tweet.

*Computing user-tweet interest based on trust:* Intuitively, if a user always retweet tweets which the user $v$ post, there is a strong possibility that the user will retweet the tweet which the user $v$ post in the future. So we introduce the trust feature to predict the strength of the user's interest in a tweet.

*Computing user-tweet interest based on similarity:* Based on traditional collaborative filtering, we computed the similarity between users according to the retweeted relation and predicted the strength of the user's interest in a tweet.

*Computing user-tweet interest based on weighted interest:* After previous two steps, we propose a weighted formula to compute the user-tweet interest by integrating the strength of the user's interest based on trust and similarity.

Here, we make two assumption:

(1) If a user retweet a tweet, the user should be similar with users who also retweeted the tweet and be trust to the user who post the tweet.
(2) Users are likely to retweet the tweet which is retweeted by similar users or post by trustworthy users.

These assumptions make our model applicable to predict tweets. The remainder of this paper is organized as follows. Related work is discussed in Sect. 2. In Sect. 3, we describe the collaborative model for predicting retweeting behaviors. In Sect. 4, we present the results of our experiments and evaluate the effectiveness of our methodology compared with other baseline methods. Finally, conclusion and future works are given in Sect. 5.

## 2 Related Work

Collaborative filtering aims to do recommendation by finding users with similar preferences or items with similar properties based on a large number of users' ratings. In contrast to content-based Filtering, it doesn't consider content and totally based on user interactive information and historical items. In many scenarios, especially the lack of information about users' and items' description, collaborative filtering has great advantages. Nowadays, collective filtering has implemented in many large commercial systems, such as news recommendation in GroupLens movies, movies recommendation in MovieLens and music recommendation in Ringo , etc. Collaborative Filtering can be divided into two categories, neighborhood-based methods [4] and model-based methods [5].

Neighborhood-based methods have two basic assumptions below:

1. Users who rated similar in the past are likely to rate similar to the new item.
2. Items which got similar ratings in the past are likely to get similar ratings from the new user.

There have two methods called user-based collaborative filtering based on the first assumption and item-based collaborative filtering based on the second assumption respectively. Both methods essentially filter out irrelevant ratings information and do recommendation according to most similar user-item ratings information. Therefore, calculating similarity is important in neighborhood-based methods. Many works integrate a lot of features to calculate the similarity between users or items. Such as behaviors' time series, social relationships, etc. In this paper, we integrate the users' similarity and trust into our prediction model based on user-based collaborative filtering.

# 3 Methods

## 3.1 Computing Interest Based on Similarity

As the Fig. 2 shown, we regard the tweet as the item and calculate the users' similarity by the users-tweets ratings matrix based on the user-based collaborative filtering. The ratings matrix only contains 0 and 1. Retweeting a tweet corresponds to a 1 rating, not-retweeting to a 0 rating. Table 1 shows the users-tweets ratings matrix.

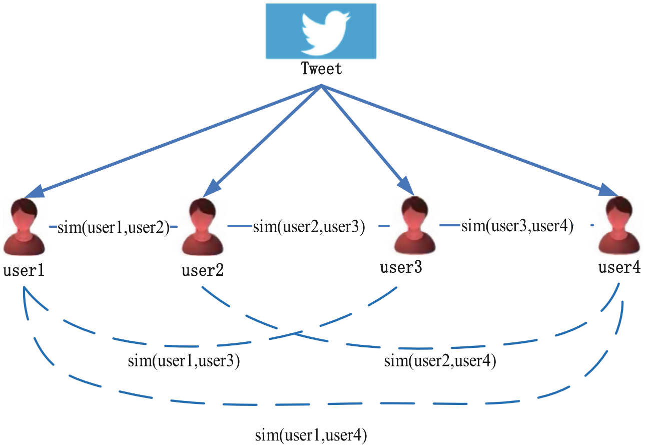


**Fig. 2.** Similarity between users who retweeted the same tweet

After getting the users-tweets ratings matrix, given a user $u$ and a tweet $t$, the predicted interest value that $u$ retweet $t$ is calculated as below:

**Table 1.** Users-tweets ratings matrix

| | Tweet 1 | ... | Tweet j | ... | Tweet n |
|---|---|---|---|---|---|
| User 1 | $r_{11}$ | ... | $r_{1j}$ | ... | $r_{1n}$ |
| ... | ... | ... | ... | ... | ... |
| User i | $r_{i1}$ | ... | $r_{ij}$ | ... | $r_{in}$ |
| ... | ... | ... | ... | ... | ... |
| User m | $r_{m1}$ | ... | $r_{mj}$ | ... | $r_{mn}$ |

$$p(u,t) = \sum_{v \in S(u,K) \cap N(t)} sim(u,v) r_{vt} \tag{1}$$

Here $S(u,K)$ contains most similar $K$ users with the user $u$, $N(t)$ indicates users who retweet the tweet $t$. $r_{vt}$ is whether the user $v$ retweet the tweet $t$ (1 for yes, 0 for no). $sim(u,v)$ is the similarity between the user $u$ and the user $v$,it is the key step in the user-based collaborative filtering and has an important impact on the algorithm results. The $sim(u,v)$ is caculated by using the IUF(Inverse User Frequence) [6]:

$$sim(u,v) = \frac{\sum_{i \in N(u) \cap N(v)} \frac{1}{\log(1+retweetCount(i))}}{\sqrt{|N(u)||N(v)|}} \tag{2}$$

Here, $N(u)$ is tweets that the user $u$ retweeted, $N(v)$ is tweets that the user $v$ retweeted. $retweetCount(i)$ is times which tweet $i$ is retweeted. $\frac{1}{\ln(1+retweetCount(i))}$ is the penalty factor [7]. The more the tweet is retweeted, the less its value is. The denominator is the regularization term, which scales the value from 0 to 1.

Then we can get the similarity matrix $W$ from the $sim(u,v)$:

$$W_{similarity} = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1m} \\ w_{21} & w_{22} & \dots & w_{2m} \\ \dots & \dots & \dots & \dots \\ w_{m1} & w_{m2} & \dots & w_{mm} \end{bmatrix} \tag{3}$$

### 3.2 Computing Interest Based on Trust

However, unlike the stuff purchasing, tweets' retweeting behaviors spread like a tree shown in Fig. 3. So we introduce the trust feature to build our collaborative model. The trust feature means that if the user $v$ retweet a tweet the user $u$ post, then we believe that the user $v$ is trust in the user $u$. For evaluation, we make an assumption below:

**Assumption.** Tweet retweeted time by the user $v$ closer the published time by the user $u$, the higher trust degree that the user $v$ to the user $u$.

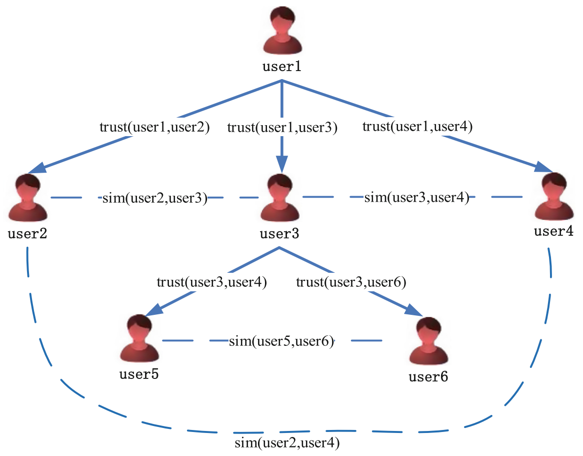


**Fig. 3.** Similarity between users who retweet same tweet

Intuitively, if the user $u$ post a tweet $k$ and the user $v$ retweets the tweet $k$, it means that the user $v$ like the tweet $k$ and will like to share this tweet with his followers. It also means the user $v$ trusted the user $u$. The more quickly retweeting time to the post time, the more attention the user $v$ pays to the user $u$. Then we introduce the negative exponential distribution to simulate the delay between retweeted time and post time:

$$trust_k(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{4}$$

Here $x$ is calculated as below:

$$x = \frac{createdAt_{user_u}(k) - createdAt_{user_v}(k)}{createdAt_{max}(k) - createdAt_{min}(k)} \text{(if u,v are adjacent)} \tag{5}$$

Where $createdAt_{user_u}(k) - createdAt_{user_v}(k)$ refers to the delay between the $user_u$ retweeted tweet and the time $user_v$ post the tweet. $createdAt_{max}(k) - createdAt_{min}(k)$ measures the max tweet's delay between latest retweeted time and original post time.

The maximum likelihood estimate for the rate parameter $\lambda$ is:

$$\widehat{\lambda} = \frac{1}{\bar{x}} \tag{6}$$

where: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean.

Meanwhile, trust can transfer. It means if the user $w$ retweets the tweet $k$ which the user $u$ retweet by the user $v$, the trust that the user $w$ to the user $v$ calculated as follow:

$$Path_k(user_w, user_v) = \{trust_1(w,u), trust_2(u,v)\} \tag{7}$$

$$trust_k(user_w, user_v) = \prod_{i \in S(user_w, user_v)} trust_i \tag{8}$$

Here, the Eq. (7) is the retweeting route from w to v, which contains the trust value between w and v. Eq. (8) is the product of trust value in $Path_k(user_w, user_v)$.

According to the one-tweet's trust, the trust that the user $w$ to the user $v$ calculated as follow:

$$trust(user_w, user_v) = \sum_{k=1}^{m} trust_k \tag{9}$$

Then we can get the trust matrix based on trust.

$$W_{trust} = \begin{bmatrix} w_{11} & w_{12} & ... & w_{1m} \\ w_{21} & w_{22} & ... & w_{2m} \\ ... & ... & ... & ... \\ w_{m1} & w_{m2} & ... & w_{mm} \end{bmatrix} \tag{10}$$

According to the trust feature, we can caculate the interest value if given a user and a tweet:

$$p(u,t) = \sum_{v \in S(u,K) \cap N(t)} trust(u,v) r_{vt} \tag{11}$$

The Eq. (11) is almost the same as the Eq. (1). It means the interest that the user $u$ to the tweet $t$ which the user $v$ post.

### 3.3 Computing Interest Based on Similarities and Trust

According to the previous description, we conclude two predict interest formulas according to the similarity and trust as follow:

$$p_{sim}(u,t) = \sum_{v \in S(u,K) \cap N(t)} sim(u,v) r_{vt} \tag{12}$$

$$p_{trust}(u,t) = \sum_{v \in S(u,K) \cap N(t)} trust(u,v) r_{vt} \tag{13}$$

The two equations both can be interpreted as the preference of the user $u$ to tweet $t$. To adapt the scenario of retweeting behaviors, we modify the user-based collaborative filtering model by integrating the trust feature. Then we propose a new predict interest formula based on the $p_{sim}(u,t)$ and the $p_{trust}(u,t)$ by normalizing respectively:

$$p_{interest}(u,t) = \alpha \| p_{trust}(u,t) \| + (1-\alpha) \| p_{sim}(u,t) \| \tag{14}$$

where $\alpha$ is the weighted parameter valued between 0 and 1. Then the algorithm updates the parameter $\alpha$ and loops over the test dataset. The effectiveness of algorithm is discussed in Sect. 4.

## 4 Experimental Studies

In this section, we describe our datasets and the preprocessing steps followed by the experimental results for each step in our model.

### 4.1 Dataset

Using Twitter's API, we crawled over 10 thousand users and 80 million tweets. we randomly selected users and expanded the user base by following their followers' and followees' links. After following several steps of links, we got over 10 thousands users and collected the tweets they had posted in a four months period from March 2014 to June 2014.

### 4.2 Preprocessing

In order to better close to real behaviors of users. Firstly we divided the crawled tweets into 4 parts by month and got the 4 parts of one-month dataset.

Then considering the effectiveness of algorithm, we preprocessed the one-month dataset according to the following criteria:

- Users should have retweet at least 10 tweets in order to ensure that they are relatively active.
- In order to get the retweeting information, tweets should have been retweeted.

As there is no API to directly get followees' tweets for each user without authorization. The only way to get users' scanned tweets is to simulate the timeline of a user, we collected users who have over fifteen followees in our one-month dataset and regarded tweets posted by followees in the one-month dataset as scanned tweets of users. Then we sorted scanned tweets of each user in chronological order. Finally we obtained simulated scanned tweets set of each user. The first three-fourths of simulated scanned tweets set of each user was put in the training set and others in the test set. Table 2 shows the number of tweets and users in the training set and test set of each month.

We used the retweeted tweets of each user in the training set to build the similarity matrix and trust matrix and predicted the retweeted tweets of each user in the test set based on similarity matrix and trust matrix.

### 4.3 Evaluation Metrics

We compute the mean average precision ($MAP$) to evaluate the proposed approach. For a given user, average precision ($AP$) is defined as:

$$AP = \frac{\sum_{r=1}^{N} p(r) \times rel(r)}{|R|} \tag{15}$$

**Table 2.** Training dataset and test dataset

| Time | Training set | Test set |
|---|---|---|
| 2014.03 | users:6279 tweets:52746 | users:6279 tweets:17581 |
| 2014.04 | users:8327 tweets:121630 | users:8327 tweets:40543 |
| 2014.05 | users:5732 tweets:64367 | users:5732 tweets:21455 |
| 2014.06 | users:5437 tweets:40620 | users:5437 tweets:13540 |

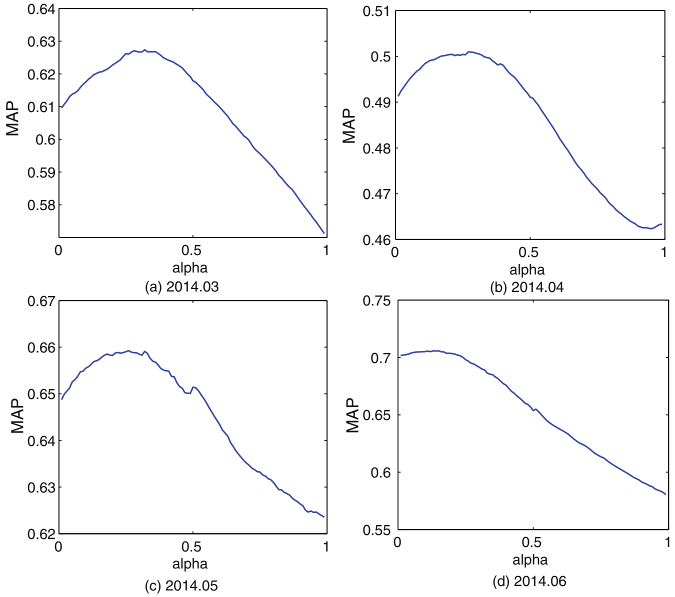
Where $r$ is the number of scanned tweets for a given user; $|R|$ is the number of retweeted tweets for a given user; $rel(r)$ is a binary function to describe whether the user has retweeted the tweet in the scanned tweet list.

Then the $MAP$ can be obtained by averaging the $AP$ values of all the users:

$$MAP = \frac{\sum_{j=1}^{m} AP_j}{m} \tag{16}$$

## 4.4 Method Comparison and Result Discussion

The Fig. 4 presents our proposed approach results of MAP by adjust the model parameter *alpha*. We find that our proposed method can get the best result



(a) 2014.03

(b) 2014.04

(c) 2014.05

(d) 2014.06

**Fig. 4.** Method comparison

**Table 3.** Comparsion of the two methods based on map (%)

| Time | 2014.03 | 2014.04 | 2014.05 | 2014.06 |
|---|---|---|---|---|
| SBM | 61.91 | 48.93 | 64.78 | 69.87 |
| TBM | 57.46 | 43.41 | 57.42 | 53.99 |
| Our method | 62.74 | 50.10 | 65.93 | 70.57 |

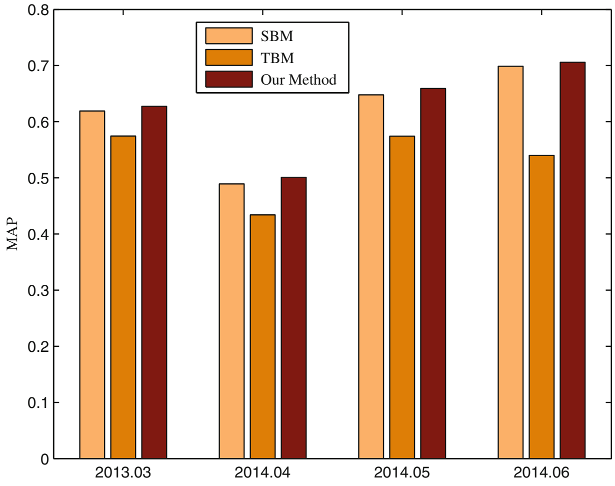


**Fig. 5.** The MAP value of our proposed method based on different alpha

when parameter is between the 0.2 and 0.3. That means the similarity feature is more userful than the trust feature to present the users' personal interest in our dataset.

In order to discover more result details based on similarity and trust, We compared with following methods:

- Similarity-Based Method (SBM): The traditional collaborative filtering based on users' similarity.
- Trust-Based Method (TBM): Only based on the users' trust to predict retweeting behaviors.

By using our proposed method which select the best weighted parameter and compared methods to process our dataset, we obtain following results, as shown in Fig. 5 and Table 3.

According to the Fig. 5 and the Table 3, our proposed method averagely exhibits higher MAP by 9.26 % and 0.96 % than the Trust-Based Method and the Similarity-Based Method on the each month of dataset, respectively. The results show Similarity-Based Method has better performance than Trust-Based Method on our dataset. However, according to *AP* results of each month's test set, we found the Trust-Based Method have higher $AP$ value than the Similarity-Based Method on some users' data. Analyzing their retweeted tweets which our method predicted, we found among these users, most users are fans of musical stars or movie stars in the real world and they followed these stars on twitter. They retweet tweets of their favorite stars frequently. So the value of weighted parameter in our model depend on the proportion of stars' fans in our dataset. The higher the ratio, the bigger the weighted parameter.

For the limitations of our method, which doesn't analyze the text of tweets, caculating the weighted parameter and users' ratio is a diffcult part in our research. To improve our method's accuracy, this will be the key point of our next research.

## 5 Conclusion and Future Work

In this paper, we propose a new collaborative model to predict retweeting behaviors. Our approach takes advantage of user-based collaborative filtering by collecting retweeting behaviors. Moreover, we incorporate the users' trust into our model to improve the effectiveness of the algorithm. Our experiments with real-life dataset have demonstrated the effectiveness in retweeting behaviors prediction.

Some future research directions are as follows. First, through analyzing of our dataset, we find that spreading information by retweeting is not complete and it can weak the effectiveness of the algorithm by calculating the users' trust and similarity. This happens due to the lack of data collection, so we must upgrade twitter crawler. Then the cold-start is another important issue to deal with. Our model doesn't consider the new users and tweets. To address this problem, we consider to detect the users' interest by using some text mining algorithms on tweets and incorporating the users' social relationship.

**Acknowledgments.** The research is supported by National Natural Science Foundation of China (No. 71331008).

## References

1. Chen, K., Chen, T., Zheng, G., Jin, O., Yao, E., Yu, Y.: Collaborative personalized tweet recommendation. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Rretrieval, pp. 661–670. ACM (2012)
2. Yang, Z., Guo, J., Cai, K., Tang, J., Li, J., Zhang, L., Su, Z.: Understanding retweeting behaviors in social networks. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 1633–1636. ACM (2010)
3. Chen, J., Nairn, R., Nelson, L., Bernstein, M., Chi, E.: Short and tweet: experiments on recommending content from information streams. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1185–1194. ACM (2010)

4. Volkovs, M., Zemel, R. S.: Collaborative ranking with 17 parameters. In: Advances in Neural Information Processing Systems, pp. 2294–2302 (2012)
5. Yuan, Q., Chen, L., Zhao, S.: Factorization vs. regularization: fusing heterogeneous social relationships in top-n recommendation. In: Proceedings of the Fifth ACM Conference on Recommender Systems, pp. 245–252. ACM (2011)
6. Uysal, I., Croft, W. B.: User oriented tweet ranking: a filtering approach to microblogs. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 2261–2264. ACM (2011)
7. Breese, J.S., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, pp. 43–52. Morgan Kaufmann Publishers Inc., San Francisco (1998)