Kuang et al. EURASIP Journal on Wireless Communications and Networking (2016) 2016:46
DOI 10.1186/s13638-016-0532-5

EURASIP Journal on Wireless Communications and Networking
a SpringerOpen Journal

RESEARCH Open Access

A comprehensive ranking model for tweets big data in online social network

Li Kuang[1], Xiang Tang[2], MeiQi Yu[1], Yujian Huang[3] and Kehua Guo[3*]

Abstract

Online social network (OSN) is an important part of cyber physical system (CPS). In OSN, micro-blogging has grown rapidly to a popular online social network recently and provides a large number of real-time tweets for users. With the popularity of micro-blogging and the increase of active users, many users are faced with an information overload problem, especially for those with many followees and thousands of tweets arriving every day. In this paper, we aim to investigate the problem of recommending valuable tweets that users are really interested in personally, so as to reduce their efforts to find useful information. We consider three major aspects in our proposed ranking model, including the popularity of a tweet itself, the intimacy between the user and the tweet publisher, and the interest fields of the user. The detailed indicators for each aspect are introduced by analyzing users' behaviors and their meanings on micro-blogs. The experimental results show that the proposed model can help improve the ranking performance in precision and greatly outperform several baseline methods.

Keywords: Online social network, Micro-blogging, Collaborative ranking, Data analysis, Cyber physical system

1 Introduction

In cyber physical system, online social network is a critical part which is able to collect various data from real users. Micro-blogging is a social network-based platform where users can share, propagate, and acquire information. It allows users to share information with their friends or the public by posting text messages of up to 140 characters, which are called tweets, through SMS, instant messenger, email, web sites, or the third party applications [1, 2]. Micro-blogging has blossomed rapidly by the virtue of immediacy and high interaction. The most representative micro-blogging services include Twitter launched in 2006 with over 500 million users, and Sina micro-blog launched in 2009 which is the most popular and powerful local micro-blogging service in China with over 300 million users. So far, Sina micro-blog service has over 100 million monthly active users and over 60 million daily active users, including a large number of pop stars, government agencies, officials, enterprise, and individual certification account. The number of tweets published in Sina everyday exceeds 100 million. The open communication mechanism makes Sina micro-blog become a public online congress hall of China.

With the rise of social networks like micro-blogging, there has been a new contact way between people. People can follow anyone whom he is interested in, including his acquaintances or friends in real life, pop stars, official spokesman for government or enterprises, and even strangers, so that he can become a fan of them on micro-blog, and get to know their news through the tweets they published.

Online social relations provide a different way for individuals to communicate digitally and allow online users to share ideas and opinions with their connected users [3]. With more and more followees for personal users and rapidly increasing tweets generated every day, many users encounter a serious problem of information overload as a result, especially for those active users. The tweets which they are really interested in or care about may be flooded. Traditional ranking in chronological order where newly tweet is placed on the top cannot fully satisfy the reading requirement of micro-blogging users.

Therefore, some micro-blogging services have released new tweets ranking models, aiming to present users the

* Correspondence: guokehua@csu.edu.cn
[3]School of Information Science & Engineering, Central South University, Changsha, China
Full list of author information is available at the end of the article

tweets that they may be willing to see on top, for example, the tweets published by acquaintances or the ones he likes. Some ranking model can even merge the tweets with similar or the same contents, so as to avoid passive flooding, while some ranking model helps users filter tweets according to their followers, tags, and tweet contents.

The performance of the tweets ranking model becomes so important, since users have been accustomed to the timeline-based model where newly tweet is placed on the top. If the users cannot feel the obvious improvement of reading efficiency, they may then feel the uncomfortableness of usage obviously; moreover, they may feel the tweets that they are reading are totally controlled by the service provider. It can be said that the intelligence of the personalized tweets ranking model determines the success or failure of a tweet service.

Therefore, in this paper, we aim to further investigate the problem of recommending valuable tweets that users are really interested in personally, so as to reduce their efforts to find useful information. Many kinds of information can be available for help ranking and recommending, and we consider three major aspects, including the popularity of a tweet itself, the intimacy between the user and the tweet publisher, and the interest fields of the user. We look into the detailed indicators for each aspect by analyzing users' behaviors and their meanings on micro-blogs. And based on the indicators for all aspects, we propose a comprehensive ranking model to capture personal interests. A series of experiments are conducted on the dataset from Sina micro-blog compared with two baseline methods. The experimental results show the proposed model can help improve the ranking performance in precision and greatly outperform the baseline methods.

The remainder of the paper is organized as follows. Related work is discussed in Section 2. The ranking model including the three aspects and their detailed indicators is introduced in Section 3. The experiment preparation and results are shown in Section 4. And finally, we conclude the paper in Section 5.

2 Related work

With the popularity of micro-blogging and the increase of active users, many users are faced with an information overload problem. It becomes an important challenge to rank and recommend the tweets that users are really interested in on top.

Most micro-blogging services present tweets in reverse chronological order, which provides no guarantee that all of these tweets are interesting to users [4]. The micro-blogs with a short length pose a challenge to traditional content-based relevance ranking algorithms. Furthermore, there are only a few links in micro-blogs, which complicates the use of traditional link-based ranking algorithms such as PageRank.

Some researchers focus on analyzing the personal interests of users and then determine whether the contents of micro-blogs are accordant with users' interests [5–11]. Wu proposed a system to generate personalized tags for Twitter users to label their interests by extracting keywords from tweets [5]. Michelson and Macskassy [6] proposed an entity-based profiling approach to discover the topics of interest for Twitter users by examining the entities they mention in their Tweets. Naveed [7] used a learning approach based on pure content features to predict the probability of a message being retweeted. Ramage et al. investigated which topics users are interested in following a labeled-LDA approach [8]. Bernstein et al. [9] developed a novel algorithm for discovering topics in short status updates powered by linguistic syntactic transformation, and then the tweets can be grouped into topics mentioned explicitly or implicitly.

Some researchers focus on analyzing the retweeting behavior of users to discover users' interests [12–16]. For example, Uysal and Croft [12] proposed a model using a coordinate ascent algorithm to rank the incoming tweets based on the likelihood that the user will retweet them. Sheng Wang et al. [13] proposed a recommendation algorithm based on Bayesian personalized ranking (BPR) by modeling user's implicit feedbacks in micro-blogging services. The proposed algorithm collects implicit feedbacks in the form of micro-blogs pairs and uses them as training pairs to learn users' interest. The implicit feedbacks include not only the tweets that users retweeted but also those not retweeted.

Some researchers further consider not only the content of tweet and the authority of the publisher but also the personal social relations of users to improve the ranking and recommendation of tweets. Chen et al. [2] proposed to learn user preferences from tweet content, user social relations, and other explicit features like publisher authority based on collaborative ranking, in order to improve the personalized recommendation performance. Nagmoti et al. [4] proposed a ranking measure that takes the number of followers and followees of the author of the tweet into account, as well as the relative length and the presence or a URL in the tweet. Yan et al. [17] presented a co-ranking framework for a tweet recommendation system that takes popularity, personalization, and diversity into account. Tim Paek et al. [18] conducted a study in Facebook and learned classifiers of newsfeed and friend importance to identify predictive sets of features related to social media properties, the message text, and shared background information. J. Chen et al. [19] studied URL recommendation on Twitter as a means to better direct user attention and found that both topic relevance and the social voting process were helpful in providing recommendations. J. Wu et al. [20] developed a trust-

aware social media recommendation framework. A two-phase process that employs graph summarization and content-based clustering is developed to partition users into different interest groups. The interest group information is then used for recommendation purpose.

Compared with existing work, we propose a comprehensive ranking model by considering all the perspectives mentioned above. We try to divide the perspectives into three aspects, and the main contribution of our work is that we detail into each aspect for measuring indicators by analyzing users' behaviors and the meaning behind the behaviors on micro-blogs. The quantization for each indicator is defined based on a statistics of Sina micro-blog dataset. The experimental results show the listed contributions can help improve tweet ranking performance.

3 Tweets ranking model

3.1 Indicators for ranking

Intuitively, a tweet is valuable to a user, if the user is interested in or willing to read it. Whether a user is interested in a tweet is determined by many factors, and we try to partition the factors into three aspects, namely the popularity of the tweet, the intimacy between the user and the tweet publisher, and the interest fields of the user. We will illustrate detailed indicators of the three aspects in the following.

3.1.1 Popularity of a tweet

Crowd psychology is quite common in social life. Therefore, if a tweet is very hot, it may be also interesting to the current user. On the other hand, in most cases, a hot tweet means it is worth of reading. The reason why a tweet is hot may due to that the content of a tweet is about a big social event, a celebrity's affair, or a hot contest, film, and so on. In addition, the micro-blog has a strong celebrity effect, which means a tweet that is issued by a celebrity may get a high attention by its large number of fans. To evaluate whether a tweet is hot or not, we consider the following indicators: the number of retweets, the number of comments, and the number of attitudes.

(a) Number of retweets

A representative character of a hot tweet is that there are many retweets of it. Retweeting is a common behavior in micro-blogs, which allows users to post the original tweet onto their own homepages in micro-blogs with comments. The retweeting behavior means the user is interested in the tweet to a certain extent. The 80/20 rule, or the Pareto Principal, states that for many phenomena, 80 % of consequences stem from 20 % of the causes [http://en.wikipedia.org/wiki/Pareto_principle]. The rule is also applicable to the micro-blog, where 80 % of the influence comes from 20 % of tweets, which are the most popular ones.

A survey is conducted on the Sina micro-blog, as shown in Fig. 1. From Fig. 1, we find that the number of retweets for up to 80 % tweets is below 200, and those tweets have a little influence on the micro-blog. Only 1 % of tweets have been retweeted over 10,000 times. According to the data distribution and the 80/20 rule, we design a three-phase formular to get the score on the part, namely $S_{\text{number of retweets}}$. For a tweet with retweets above 10,000 times, it will get 100 on $S_{\text{number of retweets}}$; for a tweet with retweets less than 200, it will get 0.1 times of retweets on $S_{\text{number of retweets}}$; and for a tweet with retweets between 200 and 10,000, there is a base score 20 and an addition part according to how much it exceeds 200. A detailed score on $S_{\text{number of retweets}}$ is defined as formula (1):

$$S_{\text{number of retweets}} = \begin{cases} 0.1 * \chi, & \chi \leq 200 \\ 20 + 0.008 * (\chi - 200), & 200 < \chi \leq 10,000 \\ 100, & \chi > 10,000 \end{cases} \quad (1)$$

(b) Number of comments

Users can also input comments below a tweet, including those with retweeting or not. A user is willing to write a comment to a tweet, can be interpreted as a kind of interest to the tweet. Therefore, another character of a hot tweet is the number of its comments.

But how many comments can be deemed as hot? To answer the question, we also make a survey on Sina micro-blog. From Fig. 2, we find that the number of comments for about 80 % of tweets is below 600, and only 1 % of tweets have over 20,000 comments. According to the data distribution and the 80/20 rule, we also design a three-phase formular to get the score on the part, namely $S_{\text{number of comments}}$. For a tweet with retweets above 20,000 times, it will get 100 on $S_{\text{number of comments}}$; for a tweet with retweets less than 600, it will get 0.033 times of retweets on

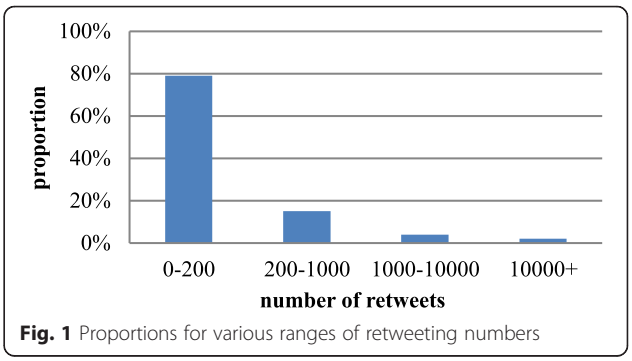

Fig. 1 Proportions for various ranges of retweeting numbers

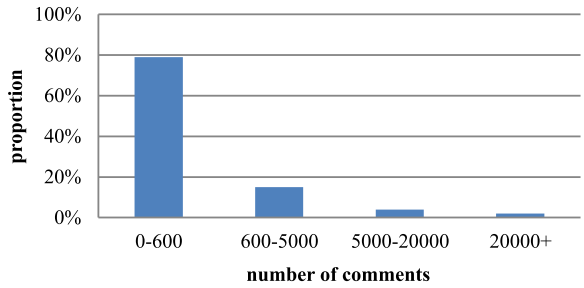

Fig. 2 Proportions for various ranges of comments numbers

$S_{\text{number of comments}}$; and for a tweet with retweets between 600 and 20,000, there is a base score 20 and an addition part according to how much it exceeds 600. A detailed score on $S_{\text{number of comments}}$ is defined as formula (2):

$$S_{\text{number of comments}} = \begin{cases} 0.033 * \chi, & \chi \leq 600 \\ 20 + 0.004 * (\chi - 600), & 600 < \chi \leq 20,000 \\ 100, & \chi > 20,000 \end{cases} \quad (2)$$

(c) Number of attitudes

Sina micro-blog also allows users to approve a tweet by clicking a raising hand. It reflects the user's attitude towards the tweet as well as a kind of interest. Therefore, if there are many users approve the tweet, it can be said that the tweet is hot.

According to the statistics on the number of attitudes of a tweet, as shown in Fig. 3, the number for up to 80 % of tweets is less than 100, and only 1 % of tweets have over 5000 praises.

According to the data distribution and the 80/20 rule, we follow the similar design idea as above and a detailed score on this part, namely $S_{\text{number of attitudes}}$, is defined as formula (3):

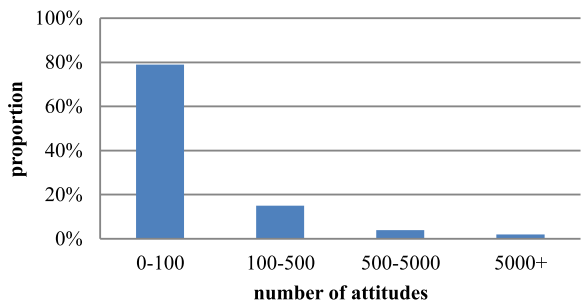

Fig. 3 Proportions for various ranges of attitudes numbers

$$S_{\text{number of attitudes}} = \begin{cases} 0.2 * \chi, & \chi \leq 100 \\ 20 + 0.016 * (\chi - 100), & 100 < \chi \leq 5000 \\ 100, & \chi > 5000 \end{cases} \quad (3)$$

According to our analysis, the popularity of a tweet can be reflected from its number of retweets, comments and attitudes, and we can calculate the popularity of a tweet by formula (4):

$$\begin{aligned} S_{\text{popularity}} = {} & \alpha * S_{\text{number of retweets}} + \beta \\ & * S_{\text{number of comments}} + \gamma \\ & * S_{\text{number of attitudes}}, \ \alpha, \beta, \gamma \in [0, 1], \text{ and } \alpha \\ & + \beta + \gamma \\ = {} & 1 \end{aligned} \quad (4)$$

Here, we simply think the three indicators are of the same influence, and we set α = β = γ = 0.33.

3.1.2 Intimacy between the user and the tweet publisher

The interactions between tweet users are based on a following and followed mechanism. The mechanism makes users subscribing information from their followees while spreading information to their followers. The users are connected by the following and followed mechanism, and a social network is formed.

In the list of followees, there are many kinds of social relations, such as friends, families, schoolmates, colleagues, and favorite pop stars in real life. Compared to some virtual public tweet users such as official representative for government and enterprises or some public users related to one's interest like funny stories, hairdressing, traveling, and cuisine, the social relations in real life have a closer relation than the virtual ones. Users tend to paying more attention to the tweets that are published or retweeted by the acquaintances in real life. Therefore, we think the interest of a tweet to a user is also related to the intimacy between the user and the tweet publisher. That is, if the user has a close relation and a high attention on the tweet publisher, there will be more probability that the user is interested in his published tweets.

Since there are different intimacy degrees between the user and his followees, we should further investigate the indicators that determine the intimacy degree. One thing need to note is, the relation here is single way from the user to his followees, and we just need to identify how the users cares about the followee, but not the reverse or both. According to our analysis on the interaction behaviors in micro-blogs, we think the indicators include

(a) Number of retweets, comments, attitudes, and mentions

There are four kinds of interaction behaviors between a user and his followees on micro-blogs:

retweeting a status posted by a followee, writing a comment of a tweet posted by a followee, stating an attitude of a tweet posted by a followee, and publishing a tweet with a mention of a followee.

If a user has many interaction behaviors with a followee, that is, retweeting many tweets of a followee, usually writing a comment or stating an attitude of a tweet posted by the followee, or usually mentioning a followee in the user's tweets, it means the user plays a high attention and interaction on the followee. Therefore, the total number of retweets, comments and attitudes of all tweets posted by a followee, and the number of mentions can reflect the intimacy of the users to the followee. The score on the total number of interactions is a function $f_{\text{number of interactions}}$ that is related to the number of interactions. According to a statistics on the number of various interactions between users on Sina micro-blog as shown in Fig. 4, the number of interactions between over 80 % user pairs is below 50, while only 1 % is above 500. Therefore, the function $f_{\text{number of interactions}}$ is defined as formula (5):

$$f_{\text{number of interactions}} = \begin{cases} 0.4 * \chi, & \chi \le 50 \\ 20 + 0.17 * (\chi - 50), & 50 \le \chi \le 500 \\ 100, & \chi > 500 \end{cases} \tag{5}$$

(b) Average response time of retweets, comments, and attitudes

In addition to the total number of the interaction behaviors, we find that the extent of attention is also related to the average response time of the interaction behaviors. Specifically speaking, if the user retweets, comments, or states an attitude about the tweet after a long time when the followee published it, it means the user reviewed the unread tweets of the followee purposely, and it can be concluded that the user pays a close attention to the followee. Another case is, if the user often makes a quick response about the followee's tweets, the user may also pay close attention to the followee in real time. The average response time can also reflect the intimacy of the users to the followee.

The average response time is worked as a coefficient, and if the average response time of all interactions between the users is very small or very large, the coefficient is set to a larger number. According to the statistics on the average response time of interaction, as shown in Fig. 5, about 80 % interactions are taken from 3 to 24 h after the original tweet is issued, and only 2 % interactions are taken within 0.5 h or after over 72 h. Therefore, $\Phi_{\text{average response time}}$ is defined as follows:

$$\Phi_{\text{average response time}} = \begin{cases} 0.4 & x\langle 0.5 || x\rangle 72 \\ 0.1 & x \in (0.5,\ 3) || x \in (24, 72) \\ 0 & x \in (3,\ 24) \end{cases} \tag{6}$$

(c) Binary follow and mutual follow

If the user A follows another user B, and meanwhile user B follows A, in most cases, they may know each other and be willing to know each other's news, and their relation is closer than unilateral following. Binary follow is worked as a coefficient, and $\Delta_{\text{binary follow}}$ is defined as:

$$\Delta_{\text{binary follow}} = \begin{cases} 0.1 & \text{binary follow is true} \\ 0 & \text{binary follow is false} \end{cases} \tag{7}$$

If there are many mutual followees between the two users A and B, it represents the mutual social circle and friendship value orientation between them, which can reflect the intimacy of the users to an extent. Mutual follow is also worked as a coefficient, and $\Psi_{\text{mutual follow}(A,B)}$ is defined as:

$$\Psi_{\text{mutual follow}(A,B)} = \frac{\text{followee}_A \cap \text{followee}_B}{\text{followee}_A} \tag{8}$$

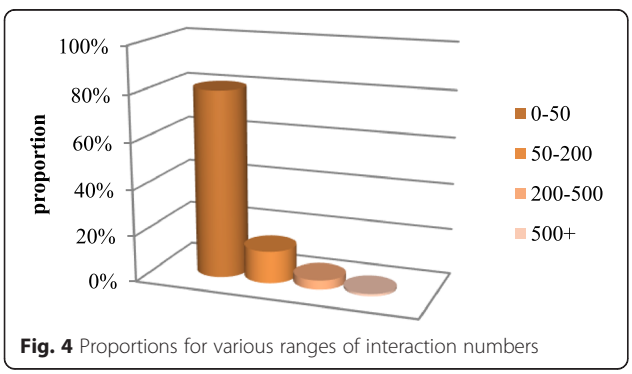

Fig. 4 Proportions for various ranges of interaction numbers

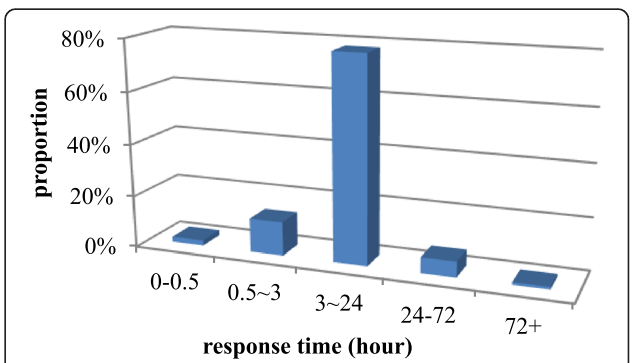

Fig. 5 Proportions for various ranges of response time

In summary, there are four indicators in calculating the intimacy between the user and a followee. The number of interaction behaviors carried by the user to the followee is worked as a base score, while the latter three indicators are worked as an adding coefficient. Specifically, the intimacy between the user A and a followee B can be calculated by formula (9):

$$S_{\text{intimacy}(A,\,B)} = f_{\text{number of interactions}} * \begin{pmatrix} 1 + \Phi_{\text{average response time}} \\ + \Delta_{\text{binary follow}} \\ + \Psi_{\text{mutual follow}(A,B)} \end{pmatrix} \tag{9}$$

3.1.3 Interest fields of the users

In addition to the hot tweets that appeal to most people and the tweets that are posted by close friends or loved ones, users may also be interested in the tweets that are matched with his interest fields. For example, a user loves traveling, he will be apt to reading the tweets related to discount flight tickets, accommodation, or tourism strategy.

The interest fields can be discovered from the tags marked by the user himself, such as music, travel, and delicacy. The interest fields can also be discovered from the celebrities that the user follows, as well as the tweets posted by the user. Text mining and clustering can be applied here to get the interest fields implicitly revealed by the user.

Once the set of interest fields is achieved from the three parts, we can simply determine whether a tweet falls into one of the user's interest fields by the mature text matchmaking technology. Hence, the score for the tweet t on the user A's interest fields $S_{\text{interest}(A,t)}$ can be defined as the max matching degree between the text of tweet and A's interest fields. In formula (10), Sim (m, n) is the semantic similarity between term m and n, t represents the keywords in the tweet t, Interest_A is the set of interest fields of user A, and $\text{Interest}_{A,\,j}$ is one of the interest field of user A.

$$S_{\text{interest}(A,t)} = \max_j \text{Sim}\left(t,\ \text{Interest}_{A,\,j}\right) \tag{10}$$

3.2 Personal tweet recommendation

Every user has his own attention preference on tweets. For example, some users who like reading news may prefer to paying close attention to hot topics and hot tweets, so that they can get a quick glance at what happened recently and what the others are paying attention to. Some users are inclined to interacting with their friends or fans, and some other ones may focus on the parts that are conformed to their interest field or personality. Therefore, for different types of users, there should be different recommendation formula, in which the three aspects are fixed while their coefficients are different. A general partition is shown as follows according to 80/20 rule and Maslow's demand theory.

3.2.1 Users of type 1 Social demand is one kind of demands in Maslow's demand theory. For the users who care most about social demand, they may pay extra attention to their followees and friends. They wish to extend their friendship by social network. Therefore, for these users, the weight of tweets which are published by their followees with close relation may take a primary place, and the formula is set as (11). Please note, $S_{\text{popularity}}$, S_{intimacy}, and S_{interest} are normalized before taking into formula (11). And it is similar in formulas (12) and (13).

$$V_{\text{total}} = S_{\text{popularity}} * 0.1 + S_{\text{intimacy}} * 0.8 + S_{\text{interest}} * 0.1 \tag{11}$$

3.2.2 Users of type 2 Respect demand is another kind of psychology demands. People wish to have a stable social position and wish to be recognized by their personal ability and achievements. Such kind of demand can be realized by following hot tweets or topics. So they can spread their humanistic concern to earn the respect, believe, and high appraisal from others and to prove their social value. For this kind of users, the weight of popularity of tweet takes a primary place, and the formula is set as follows:

$$V_{\text{total}} = S_{\text{popularity}} * 0.8 + S_{\text{intimacy}} * 0.1 + S_{\text{interest}} * 0.1 \tag{12}$$

3.2.3 Users of type 3 Self realization is the highest level in Maslow's demand theory. It means elaborating personal ability to the best and realizing personal dream and aspiration. Users can improve their learning ability by following the experts in their research area, and meanwhile, they may wish to improve their impact through their words. For this kind of users, the weight of their own interest fields may take a primary place, and hence, the formula is set as follows:

$$V_{\text{total}} = S_{\text{popularity}} * 0.1 + S_{\text{intimacy}} * 0.1 + S_{\text{interest}} * 0.8 \tag{13}$$

4 Experiments

4.1 Dataset

We obtained micro-blog dataset for experiments by invoking APIs provided by Sina. Most API accesses such

as publishing a tweet, acquiring a private message, and adding a follow require user identification and authorization. Hence, we need to take OAuth2.0 authorization before catching data.

According to the analysis of ranking indicators above, we found the corresponding API interfaces to catch the needed data. Table 1 shows the URL and function of API interfaces we used in the experiments.

The interfaces are mainly used in the data catching in the following three aspects of indicators: popularity of a tweet, intimacy between the user and the publisher of the tweet, and the interest fields of the user.

(a) Popularity of a tweet
The user ID is passed into the API invocation of No. 15 in Table 1 (statuses/friends_timeline), in order to get the latest tweets published by all the followees of current user, including the number of retweets and comments of all the tweets, and if a tweet is the retweeting of another tweet, the number of retweets and comments of the original tweet is also returned.
Meanwhile, the API interfaces of No.10 (trends/hourly), No.11 (trends/daily) and No.12 (trends/weekly) are also invoked, in order to get the hot topics in the latest 1 h, 1 day, and 1 week. The obtained tweets are matched against these hot topics through an approximate matchmaking. The matched ones are marked as 1, while those without hot topic tags are marked as 0.

(b) Intimacy between the user and the tweet publisher
The user ID is passed into the API invocation of No. 6 (friendships/friends) and No.7 (friendships/friends/bilateral) in Table 1, in order to get all the followees and double follow list of current user and meanwhile get the number of comments and mentions that the users made to a followee through API of No. 3 (statuses/mentions) and No. 4 (comments/by_me), so that we can calculate the intimacy degree between the user and each followee.

(c) Interest fields of the user
The user ID is passed into the API invocation of No. 13 (favorites) and No.14 (favorites/tags) in Table 1, in order to get all the tweets that the user marked as favorite, as well as their tags. Meanwhile, the celebrities that the user followed as interest can also be analyzed to mine the interest fields of the user. If the content or the tag of a tweet falls into the interest fields of the user, it is marked as 1, otherwise 0.

Table 1 Common API interfaces

Number	API	Function
1	statuses/public_timeline	Get the latest public tweets
2	statuses/user_timeline	Get the tweets that the user published
3	statuses/mentions	Get the latest tweet that @ current user
4	comments/by_me	Get the comments that I issued
5	comments/to_me	Get the comments that I received
6	friendships/friends	Get the follow list of current user
7	friendships/friends/bilateral	Get double follow list
8	friendships/followers	Get the fan list of current user
9	friendships/followers/active	Get fans with high quality of current user
10	trends/hourly	Return the hot topics in the latest 1 h
11	trends/daily	Return the hot topics in the latest 1 day
12	trends/weekly	Return the hot topics in the latest 1 week
13	favorites	Get the favorite tweets of current user
14	favorites/tags	Get the favorite tags of current user
15	statuses/friends_timeline	Get the latest tweets published by all followees

4.2 Results

In order to test the proposed ranking model, we asked 1048 volunteers who are relatively active in Sina microblog to participate in our experiments. The number of their followees ranges from 10 to 1000, and the average number of newly arrived tweets from their followees per hour is 56. They are asked to give their feedbacks to the chronological and intelligent ranking model by Sina and our proposed ranking model. If a tweet is useful to them, the tweet is marked as true, otherwise false. Their feedbacks are collected five times per day and the time interval between each time is over 2 h. We keep track of their opinions for 1 month.

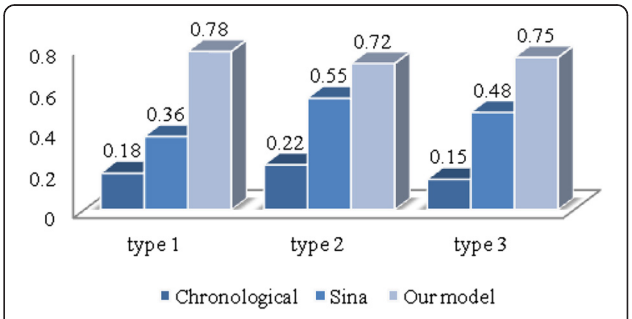

Fig. 6 Evaluation result of compared methods

We use Mean Average Precision (MAP), a popular rank evaluation method to evaluate the three models. For a single user, the precision for one time is defined as the number of tweets that are marked as true divided by the number of tweets that are generated after the last login. And then, the average precision for one user is defined as the sum of precision values for 150 times divided by 150.

The users are asked to choose one type of recommendation model that is appropriate to them. Therefore, the users are divided into three groups, and the number of users of type 1, type 2, and type 3 is 557, 348, and 143, respectively. Figure 6 shows the results of MAP on the three groups of dataset.

Not surprisingly, the performance of the chronological ranking is close to a random strategy, and it is decided by the proportion of positive samples that happen to be posted just now. Also, ranking by the intelligent model of Sina performs poorly with $(0.36 + 0.55 + 0.48)/3 = 0.46$ MAP. On the other hand, our proposed ranking model has $(0.78 + 0.72 + 0.75)/3 = 0.75$ MAP. The difference between the latter two models is especially large for the users of type 1. This means that there is still a wide gap between personal interests and the focus of public attention, which indicates that personalization is very important on micro-blog.

From the above results, we conclude that our proposed method gives a great improvement in ranking performance. The result can be explained by the fact that the model includes more indicators to describe the personal interests, the attributes of tweets, and user social relations, and this helps detect the detailed preferences of users.

5 Conclusions

In this paper, we propose a comprehensive ranking model for recommending valuable tweets to users. Our approach considers three important aspects, namely the popularity of a tweet, the intimacy between the user and the tweet publisher, and the interest fields of the user, to make a comprehensive decision on ranking. Experiments on real world data show all the information used can help improve the recommendation performance, and our final method outperforms several baseline methods.

Our future work includes

1. The coefficients in the personalized ranking formula for three types of users are predefined and fixed, and we plan to learn the coefficients as well as his preference from the user's historical behaviors.
2. Our proposed model is generic, and we can incorporate more information and indicators for ranking by analyzing users' behaviors.
3. We also plan to conduct a series of experiments on Twitter to verify our proposed model.

Competing interests
The authors declare that they have no competing interests.

Acknowledgement
This work was supported in part by a grant from National Natural Science Foundation of China (61202095, 61202341), the scientific research project of Central South University (7608010001), the Major Science and Technology Research Program for Strategic Emerging Industry of Hunan (2012GK4106), International Science and Technology Cooperation Special Projects of China (2013DFB10070), Hunan Science and Technology Plan (2012RS4054), Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education Innovation Fund (JYB201502), and Project of Innovationdriven Plan in Central South University (2015CXS010).

Author details
[1]School of Software, Central South University, Changsha, China. [2]Institute of Services Engineering, Hangzhou Normal University, Hangzhou, China. [3]School of Information Science & Engineering, Central South University, Changsha, China.

Received: 19 October 2015 Accepted: 18 January 2016
Published online: 10 February 2016

References

1. J Zhang, Y Qu, J Cody et al., *A case study of micro-blogging in the enterprise: use, value, and related issues*, 2010, pp. 123–132. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM
2. K Chen, T Chen, G Zheng, O Jin, E Yao, Y Yu, Collaborative personalized tweet recommendation, in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, 2012, pp. 661–670
3. J Tang, X Hu, H Liu, Social recommendation: a review. Soc. Netw. Anal. Min. **3**(4), 1113–1133 (2013)
4. R Nagmoti, A Teredesai, M De Cock, *Ranking approaches for microblog search*, 2010, pp. 153–157. IEEE/WIC/ACM International Conference on Web Intelligence and In-telligent Agent Technology (WI-IAT)
5. W Wu, B Zhang, M Ostendorf, Automatic generation of personalized annotation tags for Twitter users, in *Human language technologies: the annual conference of the North American chapter of the association for computational linguistics*, 2010, pp. 689–692
6. M Michelson, S Macskassy, Discovering users' topics of interest on Twitter: a first look, in *Proceedings of the fourth workshop on Analytics for noisy unstructured text data, ACM*, 2010, pp. 73–80
7. N Naveed, T Gottron, J Kunegis et al., *Bad news travel fast: a content-based analysis of interestingness on twitter*, 2011, pp. 1–7. Proceedings of the 3rd International Web Science Conference. ACM
8. R Daniel, D Susan, L Dan, Characterizing microblogs with topic models, in *International AAAI Conference on Weblogs and Social Media*, 2010, pp. 130–137
9. MS Bernstein, B Suh, L Hong, J Chen, S Kairam, EH Chi, *Eddi: interactive topic-based browsing of social status streams*, 2010, pp. 303–312. Proceedings of the 23nd annual ACM symposium on User interface software and technology
10. W Geyer, C Dugan, DR Millen et al., *Recommending topics for self-descriptions in online user profiles*, 2008, pp. 59–66. [C]//Proceedings of the 2008 ACM conference on Recommender systems. ACM
11. D Ramage, S Dumais, D Liebling, *Characterizing microblogs with topic models*, 2010
12. U Ibrahim, W Bruce Croft, User oriented tweet ranking: a filtering approach to microblogs, in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, 2011, pp. 2261–2264
13. S Wang, Z Wang, M Zhang, Personalized recommendation algorithm on microblogs. Journal of Frontiers of Computer Science and Technology **6**(10), 895–902 (2012)
14. S Adalı, J Golbeck, Predicting personality with social behavior: a comparative study. Soc. Netw. Anal. Min. **4**(1), 1–20 (2014)
15. H Wang, L Yiping, F Zhuonan, F Ling, *Retweeting analysis and prediction in microblogs: an epidemic inspired approach*, 2013, pp. 13–24. China Communications
16. F Xin, Y Shen, *Study of collective user behaviour in Twitter: a fuzzy approach*, 2014, pp. 1–12. Neural Computing and Applications

17. R Yan, M Lapata, X Li, *Tweet recommendation with graph co-ranking*, 2012, pp. 516–525. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics
18. T Paek, M Gamon, S Counts et al., Predicting the importance of newsfeed posts and social network friends. AAAI **10**, 1419–1424 (2010)
19. J Chen, R Nairn, L Nelson et al., *Short and tweet: experiments on recommending content from information streams*, 2010, pp. 1185–1194. [C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM
20. J Wu, C Liang, Q Yu, P Han, W Zhaohui, *Trust-aware media recommendation in heterogeneous social networks*, 2013, pp. 1–19. World Wide Web