International Journal of Database Theory and Application
Vol.8, No.3 (2015), pp.289-296
http://dx.doi.org/10.14257/ijdta.2015.8.3.25

A Novel Approach for Microblog Message Ranking Based on Trust Model and Content Similarity

Bei Li[1]* and Yanjie Liu[2]

[1]*Institute of Computer Science and Engineering, Henan University of Urban Construction*
[2]*Institute of International Education, Henan University of Urban Construction*
Pingdingshan, Henan Province, China 467000
lemonpepe1212@163.com, liuyanjie@hncj.edu.cn

Abstract

With the development of social network such as microblog, the number of microblog users increases rapidly. The problem of information overload caused by a large amount of data generated by users is becoming more and more serious. In order to mine the messages which specific users are interested in, we measure social relationship and interactive relationship of users respectively in this paper and propose the trust model based on the user's direct trust and indirect trust. By means of the trust model, we select the specific user's candidate user set from a large number of users. We measure the content similarity of messages in the candidate user set and propose a message ranking approach based on user trust model and content similarity. We analyze and compare the ranking results with users' real behavior in microblog platform. The experiment results show that the approach can accurately rank the microblog messages which the specific users are interested in.

Keywords*: Sina microblog, message ranking, trust model, content similarity*

1. Introduction

With the development of Twitter in the world, microblog obtains rapid development. With the advent of microblog, communication and the exchange of ideas in microblog have become a kind of fashion. Take Sina microblog in China as an example, the number of registered users in Sina microblog had been over 500 million and active users had been more than 46 million by the end of 2012. Microblog allows users to express their views using the ways of the text content and picture. Users can follow each other and repost, comment on other users' messages. Microblog has become an indispensable social media for people's life. Microblog not only plays the role of information dissemination, but also exerts a great influence on the whole society related to the political, economic, cultural fields. For the individual users of microblog, users need to obtain their interested information from massive amounts of information in microblog more accurately. For enterprise users, enterprise users need to pinpoint the target user group in microblog so that they can enhance the popularity of product and make it a well-known. For government users, the users need to discover and supervise the vicious information in microblog in order to monitor and control network public opinion. Therefore, mining interested messages for specific users and providing accurate ranking result from the massive microblog messages are very important.

However, microblog message ranking problem for specific users is faced with the following challenges:

(1) Since microblog takes the way of pushing the followings' microblog messages, it makes that most of their received messages are their followings'. The messages that users' followings posted may not be

ISSN: 2005-4270 IJDTA

messages that specific users are interested in. There are a lot of noisy messages in microblog which affect the user experience seriously.

(2) There are many users that specific users don't follow and specific users may be interested in the non-followings' messages. How to rank the non-followings' microblog messages for specific users is a challenge to be solved in message ranking problem.

Faced with these challenges, we present an approach to solve the problem that how to rank the microblog messages fast and efficiently based on trust relationship and the user's interests. We propose a message ranking approach based on user trust model and content similarity measure. In our approach, we select candidate user set of specific users in a large number of users based on user trust model and measure microblog content similarity to rank the messages.

The work of this paper is summarized as follows:

(1) Trust model is proposed based on the user direct trust and indirect trust which measures social relationship and interaction relationship for the specific users' followings and non-followings. User are added to the candidate user set when the user's trust exceeds the set threshold which provide the foundation for ranking approach based on content similarity measure and reduce the complexity of microblog message ranking.

(2) In this paper, we propose a microblog message ranking approach based on user trust model and content similarity and analyze the ranking results with users' real behavior in microblog platform. The experimental results show that our approach can rank microblog messages for specific users accurately.

The rest of the paper is organized as follows. We review the related work in Section 2. Section 3 presents the proposed microblog message ranking approach. The experimental results are reported in Section 4, and conclude this paper in Section 5.

2. Related Work

Duan *et al*.[1] found a set of most effective features for tweet ranking including account authority, length of tweet and whether a tweet contains a URL which are used to produce a ranking strategy by applying learning to rank algorithm. Ibrahim et al.[2] trained a Coordinate Ascent learning to rank algorithm to rank the incoming tweets to help users interact with the tweets they are more likely to retweet using four groups of features. Ernesto el al.[3] proposed a personalized tweet ranking algorithm for epidemic intelligence(PTR4EI) that provides users a personalized, short listed of tweets based on user's context. Guo el al.[4] proposed a comprehensive, quantitative and personalized(user-oriented) tweet ranking mechanism called TweetRank based on AHP(Analytic Hierarchy Process) which can quantify the weight of each impact factor and model user preference precisely. Huang el al.[5] presented a novel propagation model that makes use of heterogeneous networks composed of tweets, users, and web documents to rank tweets which is the first work to integrating tweets with formal genres that improves tweet ranking quality. Jan el al.[6] designed a new set of link-based features, in addition to content-based features to filter and rank tweets according to their quality. Wei el al.[7] modeled retweet behavior as a graph made up of users, publishers and tweets. Based on the graph, a feature-aware factorization model is proposed to rank the tweets. Ahmad el al.[8] introduced several features to represent the tweets and proposed a weighted Borda-Count method to blend tweets ranked list using different features which can eliminate the noisy tweets in the ranking process and improve the accuracy. However, trustworthiness and quality of tweets often plays a critical role in social network and different trust models are proposed to model the compute the trustworthiness in social network[9-15]. Some researchers are trying to rank the tweets considering the

trustworthiness in their ranking approaches. Aditi el al.[15]analyzed the trustworthiness of information in tweets and adopted a supervised machine learning and relevance feedback approach to rank tweets according to their trustworthiness. Srijith el al.[14]modeled the Twitter as a three-layer graph and proposed a method of ranking of tweets considering trustworthiness and content based popularity. To the best of our knowledge, together with the recent study on ranking tweets, we are among the first to combine direct and indirect trust model with the problem of message ranking for specific users in Sina microblog.

3. Approach for Message Ranking

Our approach for microblog message ranking approach based on trust model and content similarity is presented in this section. The framework illustrated in Fig.1. It contains two functional modules, each of which has a specific task: trust modeling evaluation and microblog message ranking based on content similarity measure. The approach utilizes trust model to compute the trustworthiness of user based on direct trust and indirect trust, only the messages that posted by the users who are trusted are used to measure content similarity, which can decrease the computing complexity of microblog message ranking.

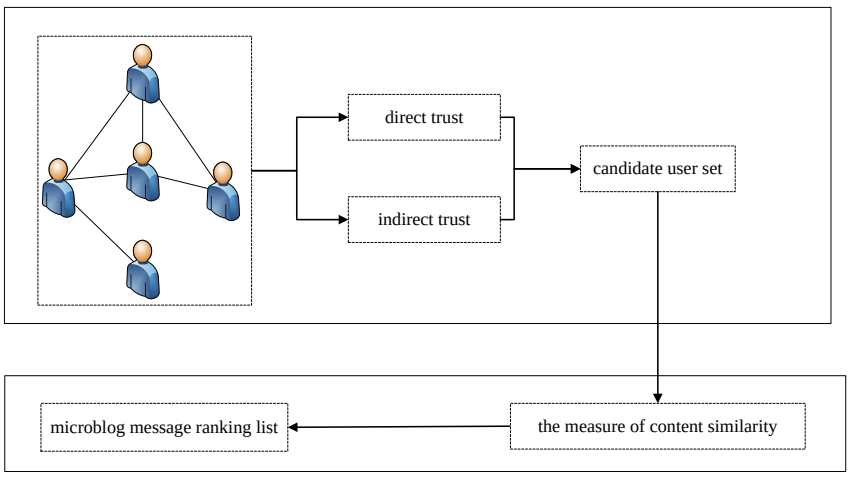

Figure 1. Overview of the Microblog Message Ranking Approach

3.1. Trust Model based on Direct Trust and Indirect Trust

The current mainstream social media are mainly divided into two categories: one is the strong relationship social media represented by Facebook and Renren, the other is the weak relationship social media represented by Twitter and Sina microblog. The establishment of user social relationship must require users to add each other in the strong relationship social networks such as Facebook. After the social relationship is established, the messages updated can be seen in their own message list. The interactive behavior and the trust relationship of user in this strong relationship network can be regarded equivalent. Users' social relationship is one-way in the weak relationship social networks such as Sina microblog, users can follow other users without agreement, which makes social and trust relationship between users are also one-way. Since the transmission of trust relationship is one-way, this feature makes that users with no direct relationship can build a kind of uncertain trust relationship. In this section we will describe user trust model from two perspectives of the direct trust and indirect trust and select candidate user set based on the user trust model.

3.1.1 Direct trust: In microblog network, the user's trust relationship depends on not only the user's follow relationship but also the user's interaction relationship. Users in microblog can increase their trust by interaction with each other. Direct trust relationship between users can be measured by social relation and interactive behavior of user. The way of interaction in microblog can be classified into four categories: repost behavior, comment behavior, reply behavior, mention behavior.

Each type of user behavior produces different effects because of the one-way relationship in microblog. One-way interaction behavior represents that users have the intention of interaction such as comment behavior, reply behavior and mention behavior. User reply behavior belongs to a response to one-way interaction behavior which shows they have strong trust relationship. The interaction degree based on one-way interaction behavior and two-way interaction behavior can be defined as follows:

$$I(u,v) = N_o(u,v) + 3N_t(u,v)$$

Where $N_o(u,v)$ is the number of one-way interaction between user u and user v, $N_t(u,v)$ is the number of two-way interaction between user u and user v.

The direct trust between user u and v can be defined as follows:

$$T_d(u,v) = \lambda_1 \bullet \frac{F(u,v)}{F(u)} + (1-\lambda_1) \bullet \frac{I(u,v)}{I(u)}$$

Where user v is the following of user u, $F(u,v)$ is the common following number of user u and user v, $F(u)$ is the following number of user u, $I(u,v)$ is the interaction number of user u, the $\lambda_1(0 \le \lambda_1 \le 1)$ can adjust the weight of the social relationship and the interaction relationship.

3.1.2 Indirect Trust: The trust relationship between users not only reflected in the following but also non-following. How to measure the trust between any two non-following users through the user's social relationship is a problem need to be solved. The way of indirect trust includes single path indirect trust and multipath indirect trust based on social relationship.

(1) Single Path Indirect Trust

As shown in Figure 2, each node represents a user in microblog, where the arrow represents the direction of the following behavior.

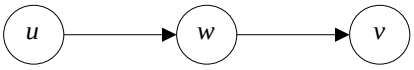

Figure 2. Single Path Indirect Trust

Since there is no direct trust relationship between user u and user v, the indirect turst can be calculated by user w. The indirect turst between user u and user v in a single path can be defined as follows:

$$T_s(u,v) = \min(T_d(u,w), T_d(w,v)) + T_d(u,w) \bullet T_d(w,v)$$

Where $T_d(u,w)$ is the indirect trust between user u and user v, $T_d(w,v)$ is the indirect trust between user w and user v.

(2) Multipath Indirect Trust

As shown in Figure 3, the indirect trust between users is the results of multipath transfer in general. User w_1 and w_2 are intermediate nodes that establish the trust relationship between user u and user v.

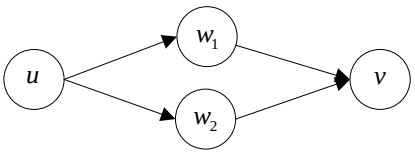

Figure 3. Multipath Indirect Trust

Multipath indirect trust can be defined as follows:

$$T_m(u,v) = \sum T_s(u,v)_k \Big/ K$$

Where $T_S(u,v)_k$ is the indirect trust in $k-th$ path, K is the number of path.

3.1.3 Selecting Candidate Users: Direct trust and indirect trust can be used to quantitatively describe the trust relationship between users. The message set of trusted users serve as the input for ranking. The set of candidate users can be selected by:

$$\mathrm{U(u)}=\left\{\mathrm{v}\middle|\left(T(u,v)>\varphi\right)\right\}$$

Where $T(u,v)$ is the trust between user u and user v according to corresponding trust type, φ is the trust threshold. User v is the element of $\mathrm{U(u)}$ when the trust between user u and user v is larger than trust threshold.

3.2 Ranking Approach based on Trust Model and Content Similarity

In order to rank the interested messages on top, a content similarity measure between candidate users' message and original users' messages is proposed.

3.2.1 Content Similarity Measure: LDA model is used to get the topic of document in the field of natural processing. As shown in Figure 4, $\alpha \to \theta \to z$ represents topic distribution as dirichlet prior distribution. $\beta \to \varphi \to w$ represents words distribution of each topic. α and β are prior parameters. Content similarity between user u and message m can be defined as follows:

$$\mathrm{S(u,m)=D(m)}\cdot\frac{\sum_{n\in M(\mathrm{u})}\mathrm{D(u,n)}}{|\mathrm{M(u)}|}$$

Where m is the messages of candidate users, $\mathrm{D(m)}$ is topic distribution vector of message m, $\mathrm{D(u,n)}$ is topic distribution vector of message n that posted by user u, $\mathrm{M(u)}$ is message set of user u, $|\mathrm{M(u)}|$ is the message number of $\mathrm{M(u)}$.

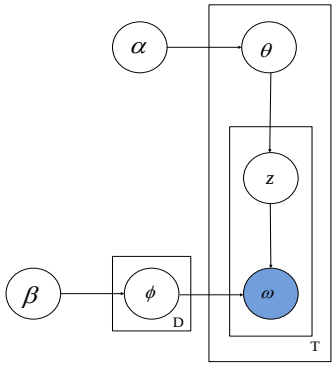

Figure 4. LDA Model

3.2.2 Ranking Interest Measure: The interest measure between user u and candidate users' messages m can be defined as follows:

$$W(u,m) = \lambda_2 \bullet T(u,v) + (1-\lambda_2) \bullet S(u,m)$$

Where $\lambda_2 (0 \le \lambda_2 \le 1)$ can adjust the weight of the trust relationship and the content similarity. We can get the ranking list by sorting the interest measure.

4. Experiment

In Section 3, we present our approach for microblog message ranking approach based on trust model and content similarity measure. We conduct extensive experiments to evaluate the effectiveness of our approach for microblog message ranking, explain empirical parameter setting, and perform a validation by comparing it with state-of-the-art method. We describe the dataset in Section 4.1. In Section 4.2, we introduce the evaluation metric and empirical parameter setting. In Section 4.3, we demonstrate the effectiveness of our method by comparing it against state-of-the-art method.

4.1. Dataset

To demonstrate the effectiveness and practicability of our method, we collected dataset used in our experiments from Sina microblog. We randomly selected 1000 users as original users and crawled their latest 100 messages, profile and following list. In order to measure the trustworthiness of user, we crawled the relevant information of users in original users and their following list with a breadth-first strategy. Finally, the statistics description of our dataset is shown in table 1. For evaluation purposes, 48 volunteers were interested in our work and assigned each message of original users a grade according to the degree of real user interest. Messages with grade 3 are the most informative, while messages with label 1 are the least informative.

Table 1. The Statistics Description of Dataset

Users	Authenticated user	Messages
1021465	63489	100398023

4.2. Evaluation Metric

To evaluate message ranking, we rely on three-fold cross validation using $nDCG$ as a measure, which considers both the informativeness and the position of a message:

$$nDCG(U,k) = \frac{1}{|U|} \sum_{i=1}^{|U|} \frac{DCG_{ik}}{IDCG_{ik}}$$

$$DCG_{ik} = \sum_{j=1}^{k} \frac{2^{label_{ij}} - 1}{\log(1+j)}$$

where U is the set of original users, $label_{ij}$ is the annotated label for the message j of user i, and $IDCG_{ik}$ is the DCG score for the ideal ranking. The average $nDCG$ score for top k messages is:

$$Avg@k = \sum_{i=1}^{k} nDCG(U,i) \Big/ k$$

4.3. Comparison with Other Approaches

In this section, we compare our approach with other ranking methods[2; 14] used in microblog which are similar to our work. The parameters of our approach are determined by determined by experiences in the experiments and we set $\lambda_1 = 0.4, \lambda_2 = 0.6, \varphi = 0.2$ in our approach the best performance. To better describe the results, we label the work[2] as UOTR(user oriented tweet ranking), the work[14] as RTCTR (ranking tweets considering trust and relevance) and our work as MRTMCS(message ranking based on trust model and content similarity). The experimental results are shown in Figure 5.

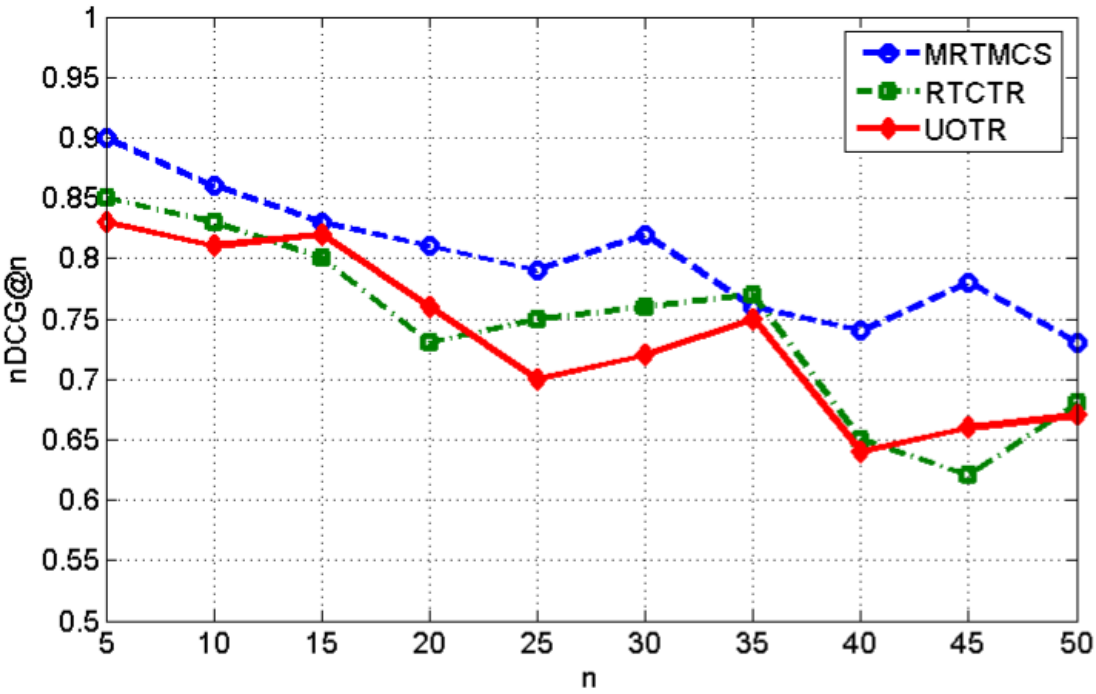

Figure 5. The Result of this Experiment

Figure 5 shows the performance of MRTMCS is enhanced considering trust model and content similarity. The experimental data show that our approach's ranking result is more accordance with the practical user behavior in microblog. Using the MRTMCS in microblog for ranking messages can be useful in increasing the effectiveness of ranking. According to the analysis based on the results of the experiments, we can find that some messages are not in ranking list of our approach. Most of the producers of message are not in candidate users because of the producers are opinion leaders and have no relationship with the original users.

5. Conclusion

In this paper, we have addressed the problem of microblog message ranking in microblog. An effective and efficient message ranking approach based on trust model and content similarity is presented. Trust model is proposed based on the user direct trust and indirect trust which measures social relationship and interaction relationship for the particular user's followings and non-followings. User are added to the candidate user set when the user's trust exceeds the set threshold which provide the foundation for ranking approach based on content similarity measure. Experiments demonstrate our approach can rank microblog messages for specific users accurately.

References

[1] Y. Duan, L. Jiang, T. Qin, M. Zhou, H.-Y. Shum, "An empirical study on learning to rank of tweets", Proceedings of the 23rd International Conference on Computational Linguistics, **(2010)** August 23-27; Beijing, China
[2] I. Uysal, W. B. Croft, "User oriented tweet ranking: a filtering approach to microblogs", Proceedings of the 20th ACM international conference on Information and knowledge management. **(2011)** October 24-28; Glasgow, United Kingdom.
[3] E. Diaz-Aviles, A. Stewart, E. Velasco, K. Denecke, W. Nejdl, "Towards personalized learning to rank for epidemic intelligence based on social media streams", Proceedings of the 21st international conference companion on World Wide Web. **(2012)** April 16-20; Lyon, France
[4] Y. Guo, L. Kang, T. Shi, "Personalized Tweet Ranking Based on AHP: A Case Study of Micro-blogging Message Ranking in T. Sina. International Conferences on Web Intelligence and Intelligent Agent Technology. **(2012)** December 4-7; Macau, China
[5] H. Huang, A. Zubiaga, H. Ji, H. Deng, D. Wang, H. K. Le, T. F. Abdelzaher, J. Han, A. Leung, J. P. Hancock, "Tweet Ranking Based on Heterogeneous Networks", Proceedings of the 24th International Conference on Computational Linguistics, **(2012)** December 8-15; Mumbai, India
[6] J. Vosecky, K. W.-T. Leung, Wilfred Ng, "Searching for quality microblog posts: Filtering and ranking based on content analysis and implicit links", Proceedings of the 17th international conference on Database Systems for Advanced Applications, **(2012)** April 15-19; Busan, South Korea
[7] W. Feng, J. Wang, "Retweet or not?: personalized tweet re-ranking", Proceedings of the sixth ACM international conference on Web search and data mining. **(2013)** February 6-8; Rome, Italy
[8] A.-H. Ahmad, V. P. Quoc, Y. Xu, "Utilizing voting systems for ranking user tweets", Proceedings of the 2014 Recommender Systems Challenge, **(2014)**
[9] A. A. Almansour, L. Brankovic, C. S. Iliopoulos, "A Model for Recalibrating Credibility in Different Contexts and Languages-A Twitter Case Study", International Journal of Digital Information and Wireless Communications, vol. 1, no. 4, **(2014)**.
[10] M. Agarwal, B Zhou, "Using Trust Model for Detecting Malicious Activities in Twitter", Social Computing, Behavioral-Cultural Modeling and Prediction.Springer, **(2014)**, pp.207-214.
[11] S. Nepal, C. Paris, S. Ku. Bista, W. Sherchan, "A trust model–based analysis of social networks", International Journal of Trust Management in Computing and Communications, vol. 1, no. 1, **(2013)**.
[12] G.-K. Sonja, S. Bitter, "Trust in online social networks: A multifaceted perspective", Forum for Social Economics, **(2013)**.
[13] G. Yin, F. Jiang, S. Cheng, X. Li, X. He, "AUTrust: A Practical Trust Measurement for Adjacent Users in Social Networks", Second International Conference on Cloud and Green Computing, **(2012)** November 1-3; Xiangtan, China.
[14] S. Ravikumar, R. Balakrishnan, S. Kambhampati, "Ranking tweets considering trust and relevance", Proceedings of the Ninth International Workshop on Information Integration on the Web, **(2012)** May 20; Arizona, USA.
[15] A. Gupta, P. Kumaraguru, "Credibility ranking of tweets during high impact events", Proceedings of the 1st Workshop on Privacy and Security in Online Social Media, **(2012)** April 17; Lyon, France.

Authors

Bei Li, she received a Master's degree from Central China Normal University in 2006, is currently working for Henan University of Urban Construction as a teacher. She is currently researching on information security, data mining.

