International Journal of Multimedia and Ubiquitous Engineering
Vol.9, No.3 (2014), pp.21-30
http://dx.doi.org/10.14257/ijmue.2014.9.3.03

Predicting Information Popularity Degree in Microblogging Diffusion Networks

Wang Jiang, Wang Li* and Wu Weili

College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan 030024, China

jokerjiang@163.com, wangli@tyut.edu.cn, wuweili@tyut.edu.cn
** Corresponding Author*

Abstract

Microblogs have rapidly become the most popular means by which people communicate with friends, pay close attention to celebrity at any time. Hence many studies on microblogging networks have been done recently, focusing on information diffusion, popularity prediction, topic detection and more. In this paper, we study the popularity of tweets in microblogging networks and introduce a novel concept "popularity degree" that help divide microblogging into four levels. Through the empirical analysis of different popularity degree, we find the retweeting information of a tweet at an earlier time can help predict its final popularity. Hence we propose a prediction model based on SVM with the retweeting information within one hour. Experimental results show our model has better ability of prediction.

Keywords: *popularity degree prediction; information diffusion; social networks*

1. Introduction

Microblog is one of the most popular kinds of social media services nowadays. Compared with traditional media, microblog has many unique characteristics such as a large amount of users, fast spread speed, brief messages and so on. The scale of microblogging network increases tremendously in recent years. Therefore, microblogging networks become fertile soil for information diffusion and have received widespread attention.

Microblogging services often support users to share their information publicly through Web or mobile applications [1]. And users can choose to retweet messages that they are interested in. In this way, the information carried by the message can be quickly spread in the social network [2]. From the perspective of transmission mode, retweeting behavior can greatly broaden the channels of information transmission and accelerate the replication and spread of information. In the process of retweeting, each user converts himself from the information "recipient" into the information "sender".

The popularity of microblogs is an important measurement for studying the influential of microblogs. Many enterprises and institutions regard it as a key propaganda media and marketing channel. Hence the prediction of popularity of microblogs is an essential problem for better understanding the retweeting behaviors in network. Furthermore, it can be used in applications such as online advertisement and recommendation.

The rest of the paper is organized as follows: in Section 2 we discuss related work. We then describe our empirical analysis for tweets with different popularity degree in Section 3. Followed in Section 4 we propose a method for modeling popularity degree. Section 5 de-

ISSN: 1975-0080 IJMUE

scribes our case study on Sina Weibo data, where we show that our model has better ability of prediction. In Section 6 we finish with a conclusion of our research.

2. Related work

Information diffusion on microblogging network is a hot topic which is receiving an increasing amount of attention of researchers. Several pioneering work have made efforts to study the popularity prediction on social networks [1, 2].

Zi Yang *et al.* [3] considered retweeting behavior is influenced by many factors: user, message, time and propose a semi-supervised framework of a factor graph model to predict users' retweet behaviors. Peng Bao *et al.* [4] regarded structural characteristics as effective clues for the popularity of short messages. They measured structural characteristics with link density and diffusion depth. The task of the popularity prediction was formulated as a classification problem by Hong *et al.* [5]. Haixin MA *et al.* [6] proposed mathematical models which could capture distributions of popularity and dynamics of retweeting behavior. They think different measurements should be used for better understanding of the popularity.

Therefore, existing methods for popularity prediction mainly focus on structural characteristics and content quality. But neither mining textual content that users generate nor analyzing the social network structure is always easy and reliable. In this paper, we just consider the users profile and retweeting information of a tweet at an earlier time to predict the tweet popularity.

3. Empirical Analysis

In this paper, we focus on Sina Microbloging Network which is the largest microblog service in China. In this section, we describe the dataset with basic statistics and empirical study for retweeting behaviors.

3.1. Dataset Description

The dataset we used is crawled from Sina Weibo.com, published by Jie Tang's group (http://arnetminer.org/Influencelocality). This dataset contains 1.7 million users and 0.3 billion following relationships among them. In order to analysis tweets popularity, we collect about 160 thousand tweets with complete information from dataset, including all its retweets, original user, retweeted time, retweeted users and so on. So we can trace a tweet's propagation path and observe information diffusion process. Table 1 lists statistics of the microblogging network.

Table 1. Data Ses

Dataset	#Users	#Following-relationships	#Original-mciroblogs	#Retweets
Weibo	1,776,950	308,489,739	300,000	23,755,810

3.2. Observation and Findings

In general, the spread of information in the Microblogging network is mainly by retweeting behavior. We often apply the number of retweets as a measurement of the popularity for a tweet. But for the common people, they have no necessary to care the accurate number of retweets, only need to know what is the degree of popularity of tweets .Meanwhile, number usually cannot give people the concept of intuitive.

Hence, we put forward a novel definition for popularity, called Popularity Degree. We divide microblog's popularity into four levels: Degree-0, represents the tweet is unpopular; Degree-1, represents the tweet with a low popularity; Degree-2, represents the tweet with a middle popularity; Degree-3, represents the tweet with a high popularity that maybe cause a hot topic. Each tweet is labeled with different popularity degree.

According to analysis of dataset, we find the average number of retweets for a popular microblog is about 80, so we give a reasonable definition for popularity degree as follows:

Table 2. Definition For Popularity Degree

Popularity Degree	Retweets Number	Implication
Degree-0	#Retweet <10	Unpopularity
Degree-1	10<#Retweet<=100	Low popularity
Degree-2	100<#Retweet<=1000	Middle popularity
Degree-3	#Retweet>1000	High popularity

First, we pay attention to basic data statistics of different popularity degree in our dataset. As shown in Figure 1, the percentage of low popularity degree is the largest than others with 41.6% and the percentage of high popularity degree is the lowest (2.7%). Meanwhile,we compare average speed of diffusion between different popularity degree and from Figure 2 ,we can see the result is the higher the popularity degree of tweet, the faster the speed of diffusion. This is common sense, so it is necessary to give further attention to temporal number of retweets and temporal speed at each period for different popularity degree.

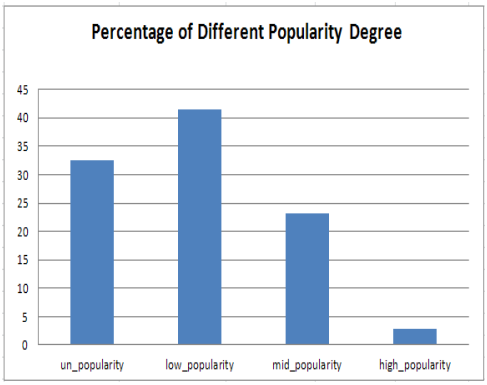


Figure 1. Percentage of Popularity Degree

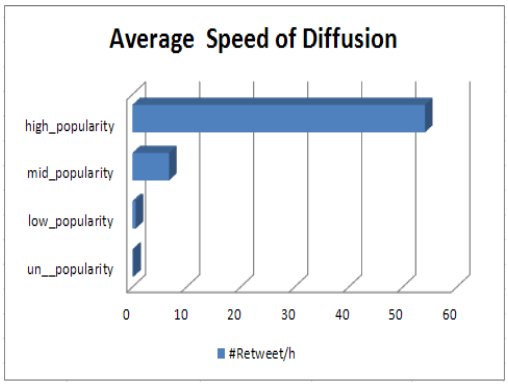


Figure 2. Average Speed of Diffusion

Figure 3 illustrates the temporal number of retweets at each period for different popularity degree. It is easy to find they have the similar curve although they have different popularity degree. At the period of six-hour, the temporal numbers all achieves the highest point. And eighty percent of retweeting behavior is finished within 48 hours which is in accordance with conclusion proposed by Ma H, Qian *et al.* [6].

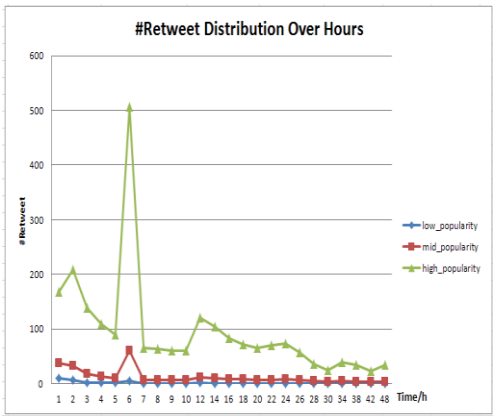


Figure 3. #Retweets at Each Period

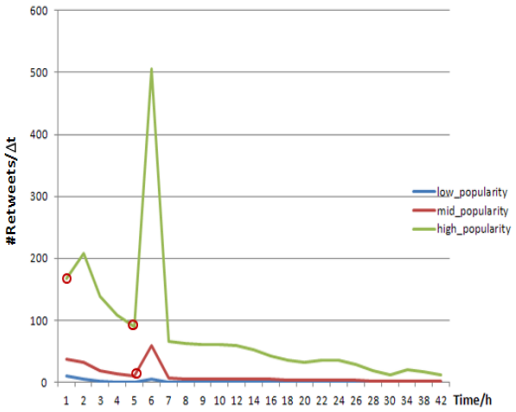


Figure 4. Speed of Retweets at Each Period

Next we research temporal speed at each period. As shown in Figure 4, we discover the speed of retweeting is not constant in the process of information diffusion. At some short period, the speed increases much faster than other periods. This stage in sociological is called "tipping point" [6]. Here we put forward the concept of acceleration point based on that, and the acceleration point hold the following conditions:

1. $TS(t+\varepsilon)-TS(t)>k$
2. $TS(t)-TS(t-\varepsilon)<k$
3. $TS(t+\varepsilon)-TS(t)>\mu*(TS(t)-TS(t-\varepsilon))$
4. $TS(t+\varepsilon)-TS(t)>N/\log(N)$

where TS(t) denotes temporal speed at period at t, ε is the small time window size for observing, N is the sum of retweets, k is average speed , μ is the threshold on change rate of slope.

Thus, we find there are two acceleration points in high popularity degree's curve while one in middle popularity degree's curve and low popularity degree's curve has no acceleration point. Figure 4 shows temporal speed of retweets at each period and the red circle stands for acceleration point.

Finally, we report the final retweeted number of a tweet with respect to the one-hour retweeted number. As shown in Figure 5, there exists a strong positive linear correlation between the final popularity and the one-hour popularity. This finding tells us that the retweeting information of a tweet at an earlier time can help predict its final popularity.

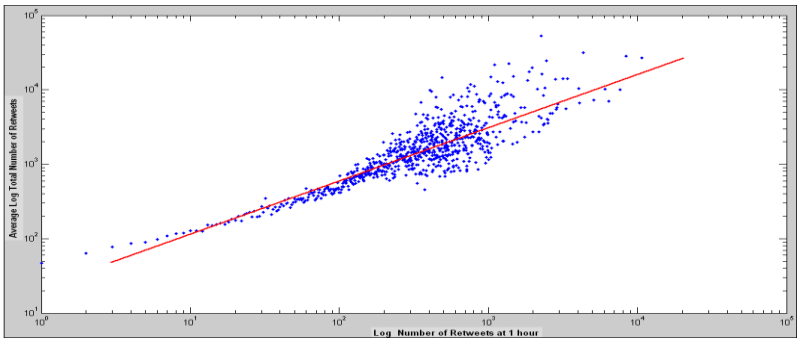


Figure 5. Correlation #Final Retweets and # Retweets-OneHour

4. Modeling Popularity Degree

In this section, we propose our model to predict popularity degree based on the above findings. We formalize our problem into a supervised machine learning model and describe the features of model.

4.1. Model

Our goal is to predict a tweet's final popularity degree, therefore it is a multi-classification problem. As we all known, Support vector machine (SVM) is a popular technique for data classification. So we propose a model based on SVM to predict the final popularity using statistical data of earlier popularity, called feature scoring in model.

Given a training dataset with feature-scoring and degree-label pairs $(f_i, d_i), i = 1,...,l$ where $f_i \in R^n$ denote values of features we selected and $d_i \in \{0,1,2,3\}$ denote four level of degree popularity. Thus, our model based on SVM is aimed to solve the following optimization problem:

$$\min_{w,b,\varepsilon} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \varepsilon_i \qquad (1)$$

$$\text{Subject to } d_i (w^T \phi(f_i) + b) \geq 1 - \varepsilon_i, \varepsilon_i \geq 0, i = 1,...,l$$

where $\phi(f_i)$ maps f_i into a higher-dimensional space, C>0 is the penalty parameter of the error term. w is the vector variable of possible high dimensionality. The parameter b determines the offset of the hyperplane from the origin along the normal vector w.

Here, we choose radial basis function (RBF) as kernel function:

$$K(f_i, f_j) = \exp(-\gamma \| f_i - f_j \|^2) \ , \ \gamma > 0 \qquad (2)$$

where γ is kernel parameters.

The reason for employment RBF as our model's kernel function is that the function nonlinearly maps samples into a higher dimensional space so it unlike the linear kernel. In addition, the number of RBF parameters is few, as a result it can reduce the complexity of the model effectively [19].

4.2. Feature Description and Scoring

Many researchers have demonstrated that the profile features of users who published microblog has the certain indicative function for prediction of retweets [10, 12]. In consequence, our model adopt the profile features of original users as our basic features as well. Furthermore, we add the retweeting information of a tweet in one hour into model.

In the actual situation, the reason that many tweets become popular is not because user who published tweet has a powerful influence, but the retweeted users have more significant influence. For example, a normal user publishes a tweet and at the same time @ a famous user, then if this tweet attract his interesting and is retweeted by this famous user, it is possible that this tweet achieves a high popularity degree at short time. So we add the retweeting information of a tweet within one hour into model in order to make our model has more accurate predictive ability.

Table 3 shows the detail information of feature; each feature is adopted by our model as f_i:

Table 3. Detail Information of Features

Feature	Description	Scoring
Original__uFollowers	The number of original user's followers	#Followers
Original_uTweets	The number of original user has published Tweets	# Tweets
Original_uStatus	The verification status of original user	{0,1}
OneHour_retweets	The number of retweets within one hour	#Retweets_OneHour
OneHour_ uFollowers	The sum of retweeted user's followers	#Followers_OneHour
OneHour_ uTweets	The sum of retweeted user's tweets	# Tweets_OneHour
OneHour_ uStatus	The number of retweeted users whose verification status is 1	{0,n}

5. Experiment

5.1. Scale

Before applying data to train model, scaling is very important. The main advantage of scaling is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. Another advantage is to avoid numerical difficulties during the calculation [20].

Consequently, our dataset should be executed to preprocessing first. Here we deal with feature score with normalization method as follows:

$$y = (x - MinValue)/(MaxValue - MinValue) \quad (3)$$

where x, y are the value of before and after normalization respectively . MaxValue, MinValue are the maximums and minimums in sample respectively.

Thus, we scale each feature score to the range [0,1]. It makes easy to train model and speed up the convergence rate of the model.

5.2. Result

In our experiment, we take 85% of all the tweets as the training set and the rest 15% as the testing set. In addition, we use model without retweeting information within one hour as baseline. The results of two models as follows:

Table 4. Prediction Results of Our Model

Model	Parameter of Kernel Function	#Iteration	Accuracy
Our Model	8.5	6329	72.4%
Baseline Model	4.5	512	56.2%

The result of experiment show that our model has a better ability of prediction than baseline model, with the predicting accuracy improved 26%. Therefore, it also empirically demonstrates that early retweeting information affect the final popularity degree.

Table 5. Prediction Results for Different Popularity Degree

Model	Accuracy of Degree-0	Accuracy of Degree-1	Accuracy of Degree-2	Accuracy of Degree-3
Our Model	78.5%	83.2%	71.4%	52.8%
Baseline Model	62.4%	58.6%	45.3%	48.2%

But as shown in in Table 5, we discover our model has a high predicting accuracy when the popularity degree is 0, 1 and 2.However, the accuracy is reduced when to predict the tweet whose popularity degree is 3. This is because the high popularity has two or more acceleration points in its lifecycle as our findings in Section 3; therefore the retweeting information within one hour is not enough to let us detect the latter accelerated points. So it is deficiencies of our model and in need of improvement.

6. Conclusion

In this paper, we studied the popularity of tweets in microblogging network and in-troduce a novel concept “popularity degree” to divide popularity into four levels. Through the empirical analysis of different popularity degree, we find the retweeting information of a tweet at an earlier time can help predict its final popularity. Hence we propose a model based on SVM with the retweeting information within one hour. Experimental results show our model has better ability of prediction. In future work, we will continue to further study and improve its ability to predict the highly popular microblog.

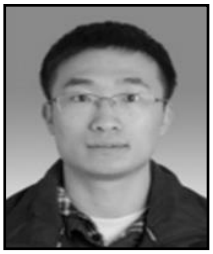
Acknowledgements

Partially supported by the Major State Basic Research Development Program of China (Grant No. 2013CB329602), the International Collaborative Project of Shanxi Province, China (Grant No.2011081034), the National Natural Science Foundation of China (Grant No. 61202215, 61100175, 61232010). China postdoctoral funding (Grant No. 2013M530738).

References

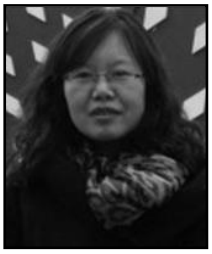
[1] A. Java, X. Song, T. Finin and B. L. Tseng, "Why we twitter: An analysis of amicroblogging community", Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM, **(2007)** August 11-15, San Jose, USA.
[2] Z. Yang, J. Guo, K. Cai, *et al.*, "Understanding retweeting behaviors in social networks", Proceedings of the 19th ACM international conference on Information and knowledge management, ACM, **(2010)** October 26-30, Toronto, Canada.
[3] P. Bao, H. W. Shen, J. Huang, *et al.*, "Popularity prediction in microblogging network: a case study on sina weibo", Proceedings of the 22nd international conference on World Wide Web companion, International World Wide Web Conferences Steering Committee, **(2013)** May17-21, Rio de Janeiro, Brazil.
[4] L. Hong, O. Dan and B. D. Davison, "Predicting popular messages in twitter", In Proc. Of WWW'11, **(2011)** March 14–18, Byderabad, India.
[5] K. Lerman and T. Hogg, "Using a model of social dynamics to predict popularity of news", In Proc. of WWW '10, **(2010)** April, pp. 621–630, Raleigh, USA.
[6] H. Ma, W. Qian, F. Xia, *et al.*, "Towards modeling popularity of microblogs", Frontiers of Computer Science, vol. 7, no. 2, **(2013)**, pp. 118-123.
[7] J. Yang and S. Counts, "Predicting the speed, scale, and range of information diffusion in twitter", In ICWSM, **(2010)** May 23-26, Washington, USA.
[8] Z. Liu, L. Liu and H. Li, "Determinants of information retweeting in microblogging", Internet Research, vol. 22, no. 4, **(2012)**, pp. 28-37.
[9] S. A. Macskassy and M. Michelson, "Why do people retweet? anti-homophily wins the day!", ICWSM, **(2011)** July 17-21, Barcelona, Spain.
[10] D. Gruhl, R. Guha, D. Liben-Nowell, *et al.*, "Information diffusion through blogspace", Proceedings of the 13th international conference on World Wide Web, ACM, **(2004)** March 14 -17, Nicosia, Cyprus.
[11] D. Kempe, J. Kleinberg and É. Tardos, "Maximizing the spread of influence through a social network", Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, **(2003)** June 9-12, California, USA.
[12] J. Ratkiewicz, M. Conover and M. Meiss, "Detecting and tracking the spread of astroturf memes in microblog streams", arXiv preprint arXiv:1011.3768, **(2010)**.
[13] S. Asur and B. A. Huberman, "Predicting the future with social media", 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE, **(2010)** July 25 – 30, Hawaii, USA.
[14] J. Zhang, B. Liu and J. Tang, "Social influence locality for modeling retweeting behaviors", IJCAI'13, **(2013)** August 3-9, Beijing, China.
[15] J. Yang and S. Counts, "Predicting the Speed, Scale, and Range of Information Diffusion in Twitter", ICWSM, **(2010)** May 23-26, Washington, USA.
[16] C. Lagnier, L. Denoyer, E. Gaussier, *et al.*, "Predicting Information Diffusion in Social Networks Using Content and User's Profiles", Advances in Information Retrieval, Springer Berlin Heidelberg, vol. 74, no. 85, **(2013)**.
[17] T. Sakaki, M. Okazaki and Y. Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors", Proceedings of the 19th international conference on World Wide Web, ACM, **(2010)** October 26-30, Toronto, Canada.
[18] B. E. Boser, I. M. Guyon and V. N. Vapnik, "A training algorithm for optimal margin classifiers", Proceedings of the fifth annual workshop on Computational learning theory, ACM, **(1992)** July 12-15, San Francisco, California.
[19] H. T. Lin and C. J. Lin, "A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods", Neural Computation, vol. 1, no. 32, **(2003)**.
[20] W. Jiyi, Z. Jianlin, W. Tong and S. Qianli, "Study on Redundant Strategies in Peer to Peer Cloud Storage Systems", Applied Mathematics & Information Sciences, vol. 5, no. 2, **(2011)**, pp. 235S-242S.

Authors



Wang Jiang

He is a master of College of Computer Science and Technology in Taiyuan University of Technology. His research focuses on mining and modeling large social networks, community detection and diffusion of information.



Wang Li

She received her MS and PhD degrees in computer science both from TaiYuan University of Technology, in 1999 and 2010 respectively. She is currently a full professor of TaiYuan University of Technology. Her research focuses on social network analysis, mobile networks communication and data mining.

Wu WeiLi

She received her MS and PhD degrees in computer science both from University of Minnesota, in 1998 and 2002 respectively. She is currently a full professor of the University of Texas at Dallas. Her research interest is mainly in database systems, social networks analysis and wireless networks.

