Trend Analysis of News Topics on Twitter

Rong Lu and Qing Yang

***Abstract*—This paper focus on trend analysis of news topics on Twitter, which include trend prediction and reasons analysis for the changes of trend. We first present a novel method to predict the trends of topics on Twitter based on MACD (Moving Average Convergence-Divergence) indicator. It is one of the simplest and most effective tendency indicators in technique analysis of stocks. We improve the original MACD indicator according to the characteristics of news topics on Twitter, define a new concept as trend momentum and use it to predict the trend of news topics. Then, we propose some reasons for the variation of trends. Experimental results show that the trend prediction is simple and effective. And, the reasons for variation of trends are also verified.**

***Index Terms*—Trend analysis, trend prediction, twitter.**

I. Introduction

Twitter is a micro-blogging and social-networking service, which is very popular today. More than 160 million users around the world are using it to remain socially connected to their friends, family members and co-workers [1]. As a micro-blogging service, it allows users to use a short text within a limit of 140 characters as their posts (also called **tweets**). The topics of these tweets range widely from the simple, such as "what I'm doing right now," to the thematic, such as "football games". There are also many different ways for users to update their tweets, including the mobile phone, Web and text messaging tools [2] and so on. As a social-networking service, Twitter employs a social model called "following" [3], in which the user is allowed to choose any other users that she wants to follow without any permission or reciprocating by following her back. The one she follows is her friend, and she is the follower. Being a follower on Twitter means that she receives all the updates of her friends [4].

There are three types of tweets: reply, retweet and normal tweet. Reply is a message to any user with the sign "@" followed by the target user's name at the beginning of the tweet. Retweet is a message which user shares from friends to her followers which begins with the sign "RT". Normal tweet are the tweets except replies and retweets. Users can also add tags for the tweets. On Twitter the tag is called "hashtag", and begins with sign "#". Different types of tweets and hashtags make Twitter data stream more readable.

Twitter is very suitable for news information publishing and propagation. The text of tweet is short, so it is easy for users to post their tweets with mobile phones or other tools. As a result, users can publish what they see in real-time, especially for the news information. The "following" function makes it very proper for news information diffusion. When a user participates in a news topic discussion, the related tweets will be pushed to its followers. If its followers are also interested in the topic, they may take part in the discussion, too. Thus, the news topic diffuses very fast.

There are many news events and topics discussed on Twitter. Some topics may get a lot attention and some may not. So, some time, we just want to know what topics will become hot on Twitter, and why it becomes hot. Therefore, we need to predict the trend of topics and give some explanations for the important variation of trend.

For the purpose of predicting trends of topics on Twitter, we borrow a widely used indicator in stock technique analysis: MACD, which was developed by Gerald Appel in the 1960s [5]. The MACD turns two trend-following indicators, exponential moving averages, into a momentum oscillator by subtracting the longer moving average from the shorter moving average. As a result, the MACD offers the best of both worlds: trend following and momentum.

So, we first monitor some key words of news topics on twitter and compute two different time-span moving averages in real-time. Then we define the concept of trend momentum of a news topic by subtracting the longer period moving average from the shorter one. Finally, we use the trend momentum to predict the trends of news topics in the future. Experimental results show that, our method is simple and very effective.

Next, we propose some reasons for some important changes of trend. We suppose there are two main situations. When a topic involves with some new events or some influential users, its trend may go up. When a the trend of a news topic goes down, it is because some other topics is going hot, and attract many users' attention. Experiments also verify our conjectures.

II. Related Work

Traditionally, researches about trends analysis of topics mainly focus on emerging trends detection (ETD). An emerging trend is a topic area that is growing in interest and utility over time. For example, Extensible Markup Language (XML) emerged as a trend in 1990s [6]. Porter et. al. proposed an emerging trends detection system called Technology Opportunities Analysis System(TOA) in 1995, which illustrates the range of information profiles possible by examining research and development publications and patents. There are also other emerging trends detection systems, like Timeminner [7], Patent Minner [8], HDDI[9] and so on. These methods all tried to find the emerging trends in the text database and visualize the evolution of topics.

Recently, LDA-style Topic Model becomes one of the hottest research spot in machine learning and information

Manuscript received April 9, 2012; revised May 7, 2012.

Authors are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China (e-mail: rlu@ nlpr.ia.ac.cn; qyang@ nlpr.ia.ac.cn)..

retrieval. It was also introduced in emerging trends detection. However, the original LDA [10] is independent with time. So, several topic models with temporal information were proposed, such as DDTM(Discrete-time Dynamic Topic Model) [11], CDTM(Continuous Time Dynamic Topic Model) [12], TOT(Topics over Time) [13] and TAM(Trend Analysis Model) [14]. DDTM requires that time be discredited, and the complexity of variational inference for the DDTM grows quickly as time granularity increases. The CDTM improve DDTM by using Brownian motion to model the latent topics through a sequential collection of documents. TOT considers the mixture distribution over topics is influenced by both word co-occurrences and the document's timestamp. TAM focuses on the difference between temporal words and other words in each document to detect topic evolution over time.

All the researches about ETD mainly pay attention to the evolution of topics or long-term changes of trends. However, our works focus on the variation of trends. We want to predict whether a topic on Twitter will be popular in the next few hours or it is dying. The prediction is real-time.

The studies of trend predicting are very mature in price tendency prediction of stocks and futures trading. Many indicators are used to help precisely predict the price trend, such as EMA (Exponential Moving Average) [15], MACD (Moving Average Convergence-Divergence) [5], OBV (On Balance Volume) [16], DMI (Directional Movement Index), RSI (Relative Strength Index) [17], BOLL (Bolinger Bands) [18], ROC(Rate of Change) [19] and so on. Among them, the EMA and MACD indicators are simple and widely used.

III. Trend Prediction

In this section, we give the formal definition of trends momentum. As described above, the trends momentum is related to two moving averages. So, we first give the definition of moving average. We divide the continuous time into several successive equal-sized time slices. In time slice t_i, the occurrence of a key word is denoted as $f(t_i)$, and the time-window size of the moving average is *k*. At the *n*th time slice, the moving average $MA(n,k)$ is:

$$MA(n,k)=\frac{\sum_{i=n-k+1}^{n} f(t_i)}{k} \tag{1}$$

If $n<k$, the moving average is defined as:

$$MA(n,k)=\frac{\sum_{i=1}^{n} f(t_i)}{n} \tag{2}$$

The moving average can track the trends of topics very well. Moreover, different sized moving average can track different period of trends. As the original MACD, the **trend momentum** is defined as:

$$TM(n)=MA(n,k_s)-MA(n,k_l) \tag{3}$$

k_s is a shorter sized time window, while k_l is a longer sized time window. The trend momentum is the just the value subtracting the longer moving average from the short one. However, the characteristic of topics on Twitter is a little different from the price of stocks. When the longer moving average of a topic represented by a key word is high for a while, the topic often last for a long time period as a very hot topic. But, if the price of a stock is high for a while, it is probably just the beginning of falling down. So we redefine the trend momentum as follows:

$$TM(n)=MA(n,k_s)-(MA(n,k_l))^{\alpha} \tag{4}$$

where $0<\alpha<1$.

There are two differences between the definition of trend momentum and the original MACD. First, we use moving average(MA) instead of exponential moving average(EMA). That is because the frequency of a key word is sometimes extremely volatile in different time slices. The EMA more depends on the frequencies of recent time slices, which makes it also volatile extremely sometimes. So we use MA, which is a more smooth way. Second, we give a discount parameter to the longer MA. We consider the hotter news topic should be more focused on, even if it has a similar trend momentum as a cold topic. The hotter news topic often has a bigger long period MA. So, a discount parameter α is given to the long period MA.

However, the value of trend momentum defined in Equation 4, is volatile. So, in practice, we also use moving average to smooth it:

$$Momentum(n)=MA(TM(n),k) \tag{5}$$

In MACD analysis of stock, there are several rules for predicting the future trends of the price. We choose the simplest and most effective rule. When the value of trend momentum turns from negative to positive, the trends of the topic will rise. When it changes from positive to negative, the topic is dying.

IV. Reasons for Trend Variation

When the trend changes its original development direction, it often has reasons. We consider there are three main reasons for it.

A. Key Users

The trend of a news topic becomes hotter. One reason is that some influential users may take part in the discussion of the news topic. So we need to find out the influential users who discuss the news topic at the time when the topic is becoming hotter. We use chi-square criterion value to find the key users.

If the time slice at which the news topic is becoming hotter is denoted as *T*, and the news topics is denoted as *TP*, we can obtain the statistics in Fig. 1.

	TP	$\overline{TP}$
T	A	B
$\overline{T}$	C	D

Fig. 1. Statistics of user number and topic *TP* in time slice *T*.

- A means the user numbers. The user discusses the topic *TP* at time slice *T*;
- B means the user numbers. The user discuss the topic *TP*, but not at time slice *T*;
- C means the user numbers. The user does not discuss the topic *TP* at time slice *T*;
- D means the user numbers. The user does not discuss the topic *TP* and does not publish tweet at time slice *T*, either;

Then, we use Equation 6 to compute the correlation of user *U* and the news topic *TP* at time slice *T*.

$$cor_u(T,TP) = \frac{(A+B+C+D)(AD-BC)^2}{(A+B)(C+D)(A+C)(B+D)} \quad (6)$$

Finally, we rank the users by the relation $cor_u(T,TP)$, and select the top users.

B. Key Words

The trend of a news topic becomes hotter. Another reason is that the news topic is involved with other events or topics. If we use key words to represent the other events or topics, the problem becomes to finding the most related key words, at the time slice when the trend is becoming hotter.

We also use chi-square criterion value to find the key users. The equation to compute the relation $cor_w(T,TP)$ of word *W* and the news topic *TP* at time slice *T* is as follows:

$$cor_u(T,TP) = \frac{(A'+B'+C'+D')(A'D'-B'C')^2}{(A'+B')(C'+D')(A'+C')(B'+D')} \quad (7)$$

A' means the word-topic co-frequency at time slice *T*. B', C', D' are similar as in Equation 6.

Then, we rank the users by the relation $cor_w(T,TP)$, and select the top key words.

C. Topic Interaction

The trend of a news topic falls on Twitter, because fewer and fewer users discuss it. When a topic rises, it attracts more and more attentions. Users focus on the rising topic, therefore, the other news topics will be discussed less. So the topics do interact with each other. It is a very important reason for trend falling of news topic on Twitter.

V. Data Preparation

For the purpose of validating the trend prediction method, we crawled two different datasets from Twitter. One is news event dataset，in which we consider every news event as a topic on Twitter. And the other is Twitter trends (hot topics provided by Twitter itself every hour) dataset.

For the news dataset, we first crawled the headlines of top news events from the RSS of The Associated Press website. At the same time, the tweets which match all the words in the headlines are crawled from Twitter. In this dataset, there are 1118 headlines, and more than 450 thousand tweets

For the Twitter trends dataset, we use the Stream API provided by Twitter to sample about 1% tweets of all all public statuses on Twitter. Meanwhile, we also crawled the Twitter trends in the same period. In this dataset, there are more than 20 million tweets, and 1072 trends topics. The two datasets is used for experiments and evaluation in next section.

There are two interesting characteristics of these two datasets. First, most users are not so active. Fig. 2 shows the distribution of tweets per user which basically follows a power-law distribution.

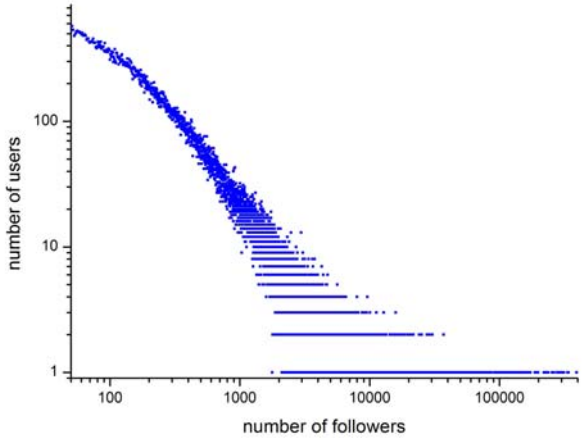

Fig. 2. Number of users vs. number of tweets

Second, if a news event has more tweets discussed it, it will have a longer life span, which also meets our common sense. This characteristic is shown in Fig. 3.

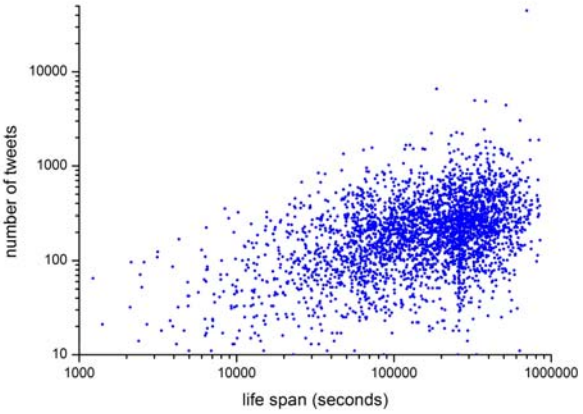

Fig. 3. Number of users vs. life span on news event

VI. Experiments and Evaluation

In this section, we first describe two cases to show how the proposed trends prediction method works, and then a quantitative evaluation is given. Second, we analyze the reasons of important trend changing.

A. Trend Prediction Method Evaluation

In Fig. 4, this case is selected from the news dataset. And the news headline is "PayPal cuts WikiLeaks from money flow", which was released by the Associated Press at 13:15 on December 4th, 2010. At about 80 minutes after the news first appeared on Twitter, the trend momentum of this news event first time changes from negative to positive. Then the tweets talk about the news event is significant increase after that time point. And at about the 230 minutes, the trend momentum changes from positive to negative. Then, very few tweets still talk about it.

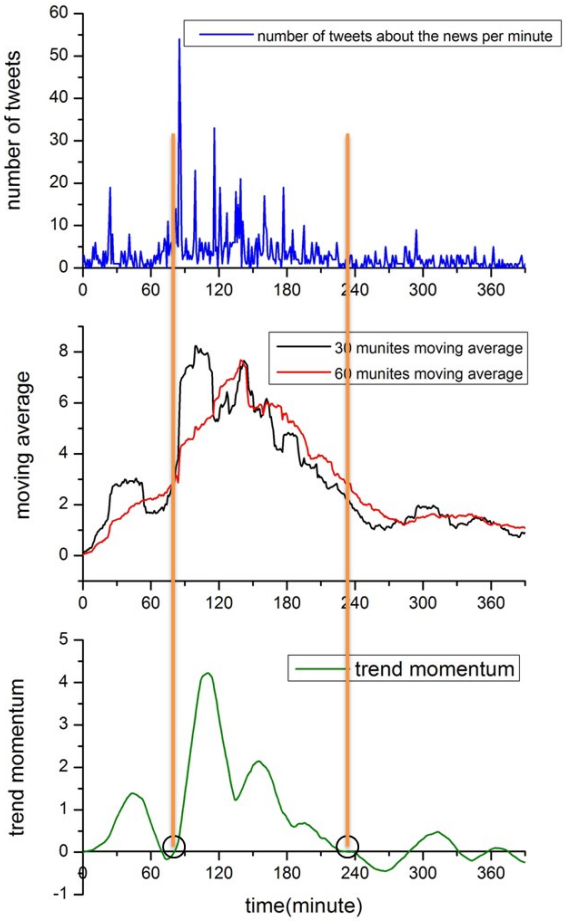

Fig. 4. Trend momentum of a news coverage.

The second case is from the Twitter trends dataset. The key word is "ipad", which becomes the Twitter trends topic at 22:00, March 24th. Its trend momentum is shown in Fig. 5. At the 34th hour(10:00, March 24th), the value of trend momentum changes from negative to positive. According to our method, we predict the trends will rise at this time point. The moving average then stay high after that as expected. After 12 hours, Twitter chose the word "ipad" as a hot topic, when the topic "ipad" had already begun to cool down. This case shows a strong contrast between the hot topics selection method of Twitter and ours. Our method has a very good forward-looking feature, while Twitter's method is lagging far behind.

The two examples above show how to use trend momentum to predict the trends of a topic or a news event on Twitter. Then, we present a quantitative evaluation

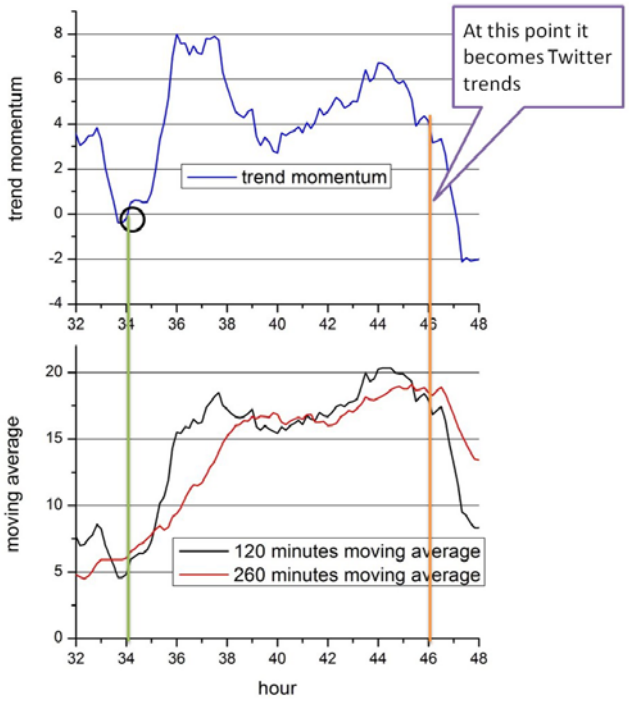

Fig. 5. Trend momentum of a Twitter trends("ipad")

Normally, if a news event or a topic gets a lot attention, it will have at least one concentrated discussion period. In this period, the density of tweets talked about the news event or topic is the highest. So, for all the news in the news dataset, we compute the tweets density in the life span of every news event, and define the highest point as the main peak. Then, we use our trend momentum method(TM) to predict the main peak. If at a certain time point, our method gives an indication that the main peak is arriving, we call this time point the prediction point. The baseline method is a fixed threshold method(FT), in which a fixed threshold is given. If the tweets density exceeds the threshold, it considers the main peak is arriving. Table I shows the results.

TABLE I: TREND MOMENTUM METHOD VS. FIXED THRESHOLD MODEL

methods	Time span from prediction point to the main peak(minutes)	omission rate
TM	24.35	10.98%
FT $\theta = 10$	65.80	15.04%
FT $\theta = 15$	53.16	25.20%
FT $\theta = 20$	36.13	40.65%
FT $\theta = 25$	24.49	42.68%
FT $\theta = 30$	16.95	68.26%
FT $\theta = 35$	15.34	75.61%
FT $\theta = 40$	14.63	77.24%

Our method has the lowest omission rate. Sometimes, the prediction point appears behind the main peak, in this situation, a missing prediction happens. The rate of all missing predictions is called omission rate.

In the Twitter trends dataset, we also examine the rate that how many Twitter trend topics have the feature that the trend momentum changes from negative to positive before it becomes the Twitter trend topics. Results are shown as Table II:

TABLE Ⅱ: TIME RANGE VS. RATE THAT HAS THE FEATURE

Time range	rate
0 ~ 4 hours	10.22%
0 ~ 4 hours	24.34%
0 ~ 4 hours	29.69%
0 ~ 4 hours	10.21%

Before the topic becomes a Twitter trend topic, about 74.46% trend topics, their trend momentum experienced the process from negative to positive in the last 16 hours.

B. Reasons Analysis of Important Variation of Trend

In this section, we will analyze the reasons of important changes of Trend of news topics on Twitter. We proposed three main reasons in the previous section, two of them may be the reasons for trend rising, and one may be the reason for trend falling.

First, the reasons for trend rising is given in Fig. 6 and Fig. 7 We monitor a trend topic "Yankee" in our Twitter trend dataset from March 30th, 2011 to April 4th, 2011, and use the prediction method based on trend momentum described above. It give four prediction that the trend will rise at the 13th, 36th, 91st and 111th hour.

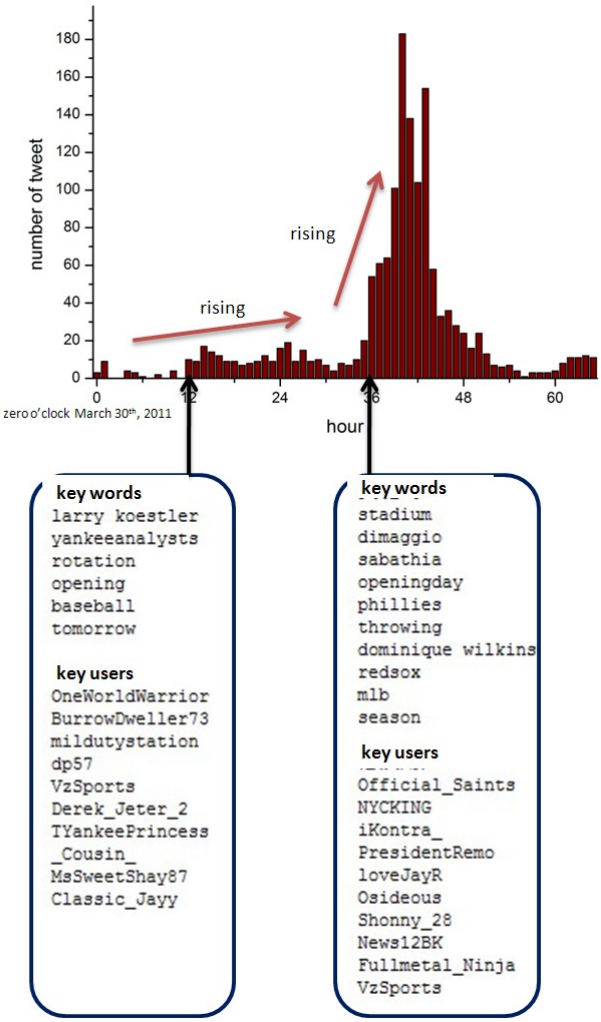

Fig.6. Possible reasons for trend rising – part 1

Fig. 6 shows the first 66th hours from 0 o'clock on March 30th, 2011. In the 13th hour, "Yankee" was most related to the key words: "rotation", "opening", "tomorrow" and so on. Because the opening day of The World Series of Major League Baseball (MLB) is March 31st. However, at the 36th hour, it is 12 o'clock on March 31st, so the key word "tomorrow" is gone. Key words such as "stadium", "opening day" and the names of some players are most related to "Yankee".

In Fig. 7, as the league goes on, "Yankee" most related to other baseball team or some players. Every time, the trend of "Yankee" rises with different related key word. Our method reveals this phenomenon.

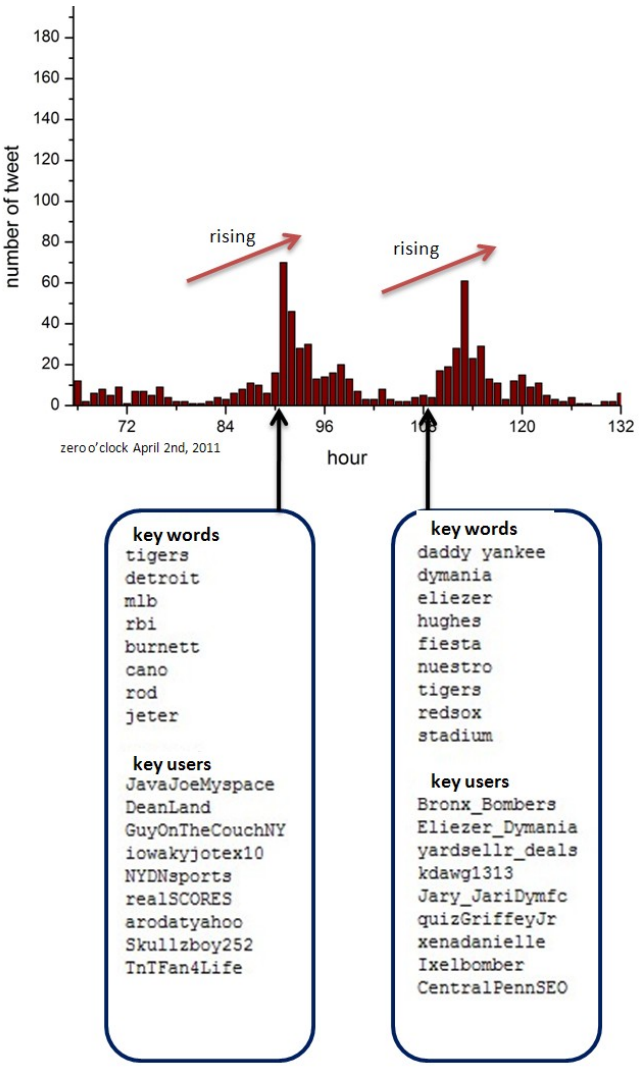

Fig. 7. Possible reasons for trend rising – part 2

Second, we monitor the rising and falling of trend of news topics on Twitter and find the interaction of each other topics. Fig. 8 shows it. In Fig. 8, it shows changing of 8 trend topics on April 4th, 2011. They almost rise one by one. It is obvious that, when a topic rises, another topic falls.

It is interesting that, the trend topic "wrestlemania" and "undertaker" are very connected in real world. Topic "wrestlemania" is a big celebration of wrestling. And, topic "undertaker" is a very famous wrestler. When the celebration begins the key word "wrestlemania" rises very quickly. However, as the celebration goes on, users turned their attention to a particular wrestler, gradually. So the trend topic "wrestlemania" falls. So the interactions of different news topics do exist.

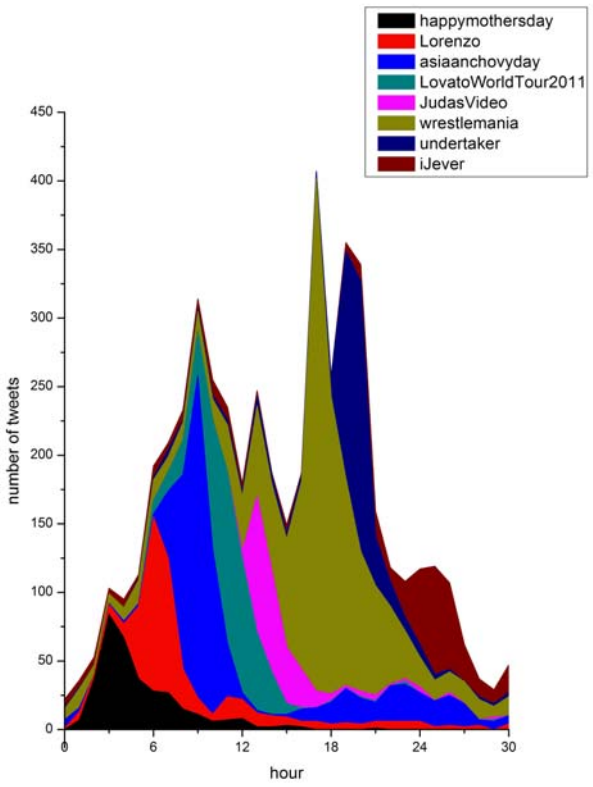

Fig. 8. Possible reasons for trend falling

In this section, we verify three main reasons for the trend changing of new topics on Twitter. They all affect the trend to some extent.

VII. Conclusion

In this paper, we present a trend prediction method for news event or news topics on Twitter. Experimental results show that the method is simple and effective. Then, we also propose and analyze several possible reasons for the trend rising and falling of news topics on Twitter.

However, this study is not only very important but also very unique. There are several further problems and research directions in this domain.

First, the trend prediction problem needs a more clear definition and a more classical evaluation. The trend prediction problem in this paper was not very well described. In the future, we will give a more mathematical definition. The evaluation method should be also improved.

Second, the trend prediction method is too simple. The MACD indicator is very mature in price prediction of stocks. Many rules are developed. So, we can use more MACD prediction rules in our method, and improve them to suit the Twitter platform.

Third, we will integrate more characteristics of Twitter into our method. For example different user post tweets at different time, the influences are not the same. Simple frequency count may not be very reasonable.

References

[1] B. A. Huberman, D. M. Romero, and F. Wu, "Social networks that matter: Twitter under the microscope," in *Social Computing Laboratory, HP Labs*, 2008.

[2] S. Milstein, A. Chowdhury, G. Hochmuth, B. Lorica, and R. Magoulas, "Twitter and the micro-messaging revolution: Communication, connections, and immediacyc140 characters at a time," *in O'Reilly Report*, November 2008.

[3] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*, New York, 2010, pp. 261–270.

[4] 4] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *WWW '10: Proceedings of the 19th international conference on World wide web*, Raleigh, 2010, pp. 591–600.

[5] G. Appel, "Become Your Own Technical Analyst," *The Journal of Wealth Management*, vol.6, no. 1, pp. 27-36, 2003

[6] A. Kontostathis, L. Galitsky, W.M. Pottenger, S. Roy, and D.J. Phelps.: *Survey of Text Mining: Clustering, Classification, and Retrieval.Survey of Text Mining*, 185-224(2003)

[7] R. Swan, D. Jensen, "Timemines: Constructing timelines with statistical models of word usage," in *KDD-2000 Workshop on Text Mining*, Boston, 2000, pp. 73-80

[8] Lent, B., Agrawal, R., Srikant, R.: Discovering trends in text databases. Proc. of the 3rd Internatonal Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, California, 1997, pp. 227-230

[9] W. M. Pottenger, Y. B. Kim, D. D. Meling, "HDDI™: Hierarchical Distributed Dynamic Indexing," *Journal of Data mining for scientific and engineering applications*, 2001.

[10] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, 2003, pp. 993-1022.

[11] D. M. Blei, J. D. Lafferty, "Dynamic topic models," in *Proc. of the 23rd ICML*, Pittsburgh, Pennsylvania, 2006, pp. 113-120.

[12] C. Wang, D.M. Blei, "Continuous time dynamic topic models," in *Proc. of UAI*, 2008.

[13] X. Wang, A. McCallum, "Topics over time: a non-Markov continuous-time model of topical trends," in *Proc. of the 12th ACM SIGKDD*, Philadelphia, PA. USA. 2006, pp. 424-433.

[14] N. Kawamae, "Trend analysis model: Trend consists of temporal words, topics, and timestamps," in *Proc. of 4th WSDM*. 2011, Hong Kong, pp. 317-326.

[15] A. J. Lawrance and P.A.W. Lewis, "The exponential autoregressive-moving average earma (p, q) process," *Journal of the Royal Statistical Society*, Series B (Methodological), 1980, pp.150-161, 1980.

[16] T. Aspray. Obv, *Perfect indicator for all markets*, 2011.

[17] J. W. Wilder. *New Concepts in Technical Trading Systems. Trend Research*, 1978.

[18] J. Bollinger. *Bollinger On Bollinger Band*, McGraw-Hill, 2001.

[19] G. Appel and F. Hitschler. *Stock Market Trading Systems*. Traders Press, 1990.

Rong Lu was born in 1985. He received the bachelor's degree in computer science and technology from University of Science and Technology of China in 2003. Now he is a PHD student in Institute of Automation of Chinese Academy of Sciences. He has authored more than 6 published technical papers in data mining and knowledge management. His current research activities mainly focus on news information mining in the social-networking service.

Qing Yang was born in 1970. He received the bachelor's degree in automation department of University of Science and Technology of China in 1990. He received the doctor's degree of pattern recognition and artificial system in Institute of Automation of Chinese Academy of Sciences in 1998. He has authored more than 50 published technical papers in computer vision and data mining. His current research activities mainly focus on data mining in the social-networking service.

He worked as a computer scientist in Lawrence Berkeley National Laboratory from 1999 to 2004. After that, he works as a research fellow in Chinese Academy of Sciences till now.