Life Activity Modeling of News Event on Twitter Using Energy Function

Rong Lu, Zhiheng Xu, Yang Zhang, and Qing Yang

Institute of Automation, Chinese Academy of Science, Beijing, China
{rlu,zhxu,yzhang,qyang}@nlpr.ia.ac.cn

Abstract. This research is the first exploration on modeling life activity of news event on Twitter. We consider a news event as a natural life form, and use an energy function to evaluate its activity. A news event on Twitter becomes more active with a burst of tweets discussing it, and it fades away with time. These changes of the activity are well captured by the energy function. Then, we incorporate this energy function into the traditional single-pass clustering algorithm, and propose a more adaptive on-line news event detection method. A corpus of tweets which discuss news events was analyzed using our method. Experimental results show that our method not only compares favorably to those of other methods in official TDT measures like precision, recall etc., but also has better time and memory performance, which makes it more suitable for a real system.

Keywords: life activity modeling, energy function, Twitter, news event detection, single-pass clustering.

1 Introduction

Twitter is a very popular micro-blogging and social-networking service. More than 160 million users around the world are using it to remain socially connected to their friends, family members and co-workers[3]. It allows users to use a short text within a limit of 140 characters as their posts (also called ***tweets***) through many ways, including the mobile phone, the Web and text messaging tools[1] and so on. Twitter also employs a social-networking model called "following"[4], in which the user is allowed to follow any other users she wants to, 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 she receives all the updates from her friends[2].

More than a micro-blogging and social-networking service, Twitter is also like a news media. Many news outlets have accounts on Twitter, such as ABC, CNN, and New York Times. They use their accounts to report news, while many other users follow these accounts to subscribe news coverage. Up to now, New York Times has already have about 3 million followers. This news reporting and reading application is so popular on Twitter, because the short text makes the

P.-N. Tan et al. (Eds.): PAKDD 2012, Part II, LNAI 7302, pp. 73–84, 2012.

news easier to read, and the social-networking functionality makes it faster to diffuse and also provides a good interaction between users.

Unfortunately, Twitter grows too fast. The number of tweets per day is over 200 million. How to obtain the desired news information among the huge mass of tweets becomes a problem. Therefore, a real-time on-line news event detection system of Twitter is necessary. Usually, traditional single-pass algorithm[21] is used to handle this problem. However, there is an ambiguous place of the traditional single-pass algorithm. This algorithm clusters tweets into different clusters as different news events. But, it does not point out when to drop a news event out of the system memory, as there is no news tweet any more. And, this may cause it time consuming and memory exhausting for a real system.

In this paper, we implement the traditional single-pass clustering algorithm by modeling the life activity of news event. First, we use an energy function to model the life activity of a news event. We consider a news event on Twitter as a natural life. For a natural life, it eats different food containing different energy. It absorbs the energy by a certain transform ratio. Then, it grows old with time. Similarly, the tweet is food to a news event on Twitter. So the energy of a single tweet, an energy ***transferred factor*** and an energy ***decayed factor*** are introduced and integrated together as an energy function. The value of the energy function indicates the activity of a news event.

Then, we incorporate the energy function into the traditional single-pass algorithm. The threshold of the traditional single-pass algorithm is a constant. But, we use a variable to replace it. This variable threshold changes with the activity. We also add a time window to determine when a news event should be dropped out of the system memory. This time window changes with the activity of the news event, too.

The rest of this paper is organized as follows: In section 2, we give a review of related works. In Section 3, we describe the concepts and details of the energy function. Then, we incorporate it into the traditional single-pass clustering algorithm in Section 4. Section 5 reports the experiments and Section 6 concludes this paper.

2 Related Work

What is Twitter? Kwak, et al.[2] point out that Twitter is not only a social network, but something more akin to traditional news media. In its follower-following topology analysis, [2] has found a non-power-law follower distribution, a short effective diameter, and low reciprocity, which all mark a deviation from known characteristics of human social networks[10]. Actually, over 85% of trending topics on Twitter are headline or persistent news in nature. Java, et al. also give six main user intentions on Twitter in [11], and reporting news is one of them.

Now, reporting and reading news is one of the most important application on Twitter. As enormous amount of tweets are generated by the users everyday, it is necessary to detect and track news event automatically. Detecting and

Tracking new event was discussed in the project called Topic Detection and Tracking(TDT), which is a DARPA-sponsored activity to detect and track news events from streams of broadcast news stories.In general, clustering techniques are the major methods of TDT. Salton in [9] introduced hierarchical agglomerative clustering(HAC) method, and Yang, et al.[6] speed up the HAC by using the technique of bucketing and re-clustering. However, HAC is not very suitable for the time-ordered data collection. The other clustering method is single-pass clustering[21,5], which processes the input documents iteratively and chronologically. Ron and James discuss the implementation and evaluation of a on-line new event detection and tracking system using the single-pass clustering algorithm in [20]. They proposed a threshold model, which regarded exploiting temporal information would lead to improve the performance.

Besides, Chen, et al. proposed an aging theory to model life cycle of news events in order to improve the traditional single-pass clustering algorithm in [7]. This work is close to ours. However, we go much further. First, we clearly defined an energy function to evaluate the activity of a news event and give a iterative algorithm to solve the parameters. Second, we use the activity of a news event to determine the threshold of the single-pass clustering algorithm, and how long a news event should stay in the system memory. Third, [7] divides the news events into short-term and long-term events, while we treat all news events the same way. Finally, our work focuses on the stream of tweets instead of traditional news reports or stories.

With the rise of Twitter, researches of event detection on Twitter have already attracted some attention. Sakaki, et al. gave a real-time event detection algorithm, which monitors tweets to detect a target event like earthquake in [15]. In their work, they also found out the tweets of a news event follows an exponential distribution with time. [22] proposed a topic detection technique that permits to retrieve in real-time the most emergent topics. [23] introduced a method to collect, group, rank and track breaking news on Twitter, and developed an application called "Hotstream".

Another issue worthwhile to note is that the tweets are much shorter and noisier. The tweets stream is mixed by News events, Conversations[12], work communication[13], business information[14] and so on. [8] made an attempt to select the tweets that discussed news events only.

3 Modeling Life Activity Using Energy Function

In this section, the details of the energy function are described. We consider a news event on Twitter as a natural life form. To track its life activity, we use the concept of energy function. Like the endogenous fitness of an artificial life agent[16], the value of the energy function indicates the activity of a news event.

3.1 Definition of Energy Function

A news event on Twitter becomes popular with a burst of tweets discussing it and it fades away with time. A tweet discussed a news event is called **news**

tweet. A news tweet to a news event is like food to a natural life. It provides energy to the news event by a certain transform ratio. However, the life energy of the news event also diminishes with time as the same as natural life grows old. Figure 1 shows this life process.

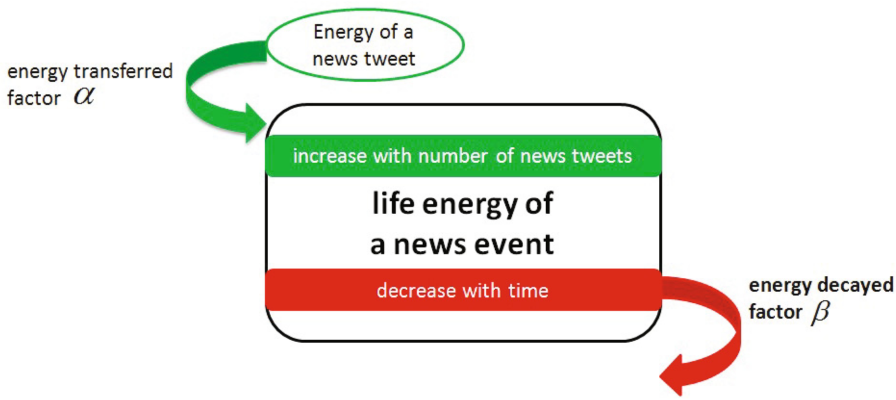

Fig. 1. Life process of a news event on Twitter

α is **energy transferred factor**, and β is **energy decayed factor**. Energy of a single news tweet is denoted by *itemEng*. Then we divide the whole life span(the time span from the first news tweet to the last one) of a news event into several successive and equal sized time slice. In time slice t, E_t represents the net energy that a news event obtains, including the energy absorb from news tweets and the energy lost with time in this time slice. A news tweet is denoted as d, and the news tweet set in this time slice is denoted as D_t. So, the net energy in this time slice t is defined as follows:

$$E_t = \sum_{d \in D_t} (\alpha \cdot itemEng(d)) - \beta \tag{1}$$

The total energy of a news event at the nth time slice is the sum of the net energy of all time slices before, so the **Energy Function** is:

$$E(n) = \sum_{t=1}^{n} E_t = \sum_{t=1}^{n} (\sum_{d \in D_t} (\alpha \cdot itemEng(d)) - \beta) \tag{2}$$

The value of $E(n)$ just indicates the activity of a news event at the nth time slice. It is easy to see $E(0) = 0$. If the news event has N time slices in total, another constraint of the energy function is:

$$E(N + 1) = 0 \tag{3}$$

The meaning of Equation 3 is also obviously. When the news event is over, its energy value should turn to 0 again.

From Equation 2, when a burst tweets discuss a news event, the energy value increases, and the news event becomes more active. As time goes by, less and less tweets discuss the news event, the decayed factor β plays a more effective role, the energy value decreases, and the news event becomes less and less active.

3.2 Energy of A Single Tweet

As mentioned above, a news tweet to a news event is like food to natural life. Different food contains different nutrition. Similarly, different news tweets also contain different energy. The news tweet posted by more influential user will get more attention. If more users can read the news tweet, the news event will have more chance to become popular. It means a news tweet from a more influential user contains more energy.

As a social-networking service, the relationships between users construct a directed map, all users are the nodes of this graph. Researches[17,18,19] have already studied on how to measure the influence of a node in the directed graph or a person in a social network. One simple method of measuring the influence of users on Twitter is the **In-degree** method in[4], it measures the influence of a user by the number of her followers. This measurement currently also employed by Twitter and many other third-party services, such as *twitterholic.com* and *wefollow.com*.

Therefore, we also choose this measurement. f denotes the number of followers of a Twitter user, f_{max} denotes the maximum number of followers that a user has in our dataset. As a result, the influence of a Twitter user is defined as:

$$if_{user}(f) = \frac{\log(f)}{\log(f_{max})} \tag{4}$$

It is obvious that $0 \leq if_{user} \leq 1$.

The energy value of a single news tweet is denoted by (**itemEng**), which has already shown up in Equation 1 and 2. It is defined as:

$$itemEng(d) = \lambda_1 + \lambda_2 \cdot if_{user} \tag{5}$$

where, $0 \leq \lambda_1 \leq 1$, $0 \leq \lambda_2 \leq 1$, and $\lambda_1 + \lambda_2 = 1$.

3.3 Constant Growth and Decay

One particular case of the life cycle of a news event is constant growth and decay, which means, no matter how active the news event is, the energy transform ratio and the loss of energy are the same. In another word, the transferred factor α and the decay factor β are both constants. As a result, the energy function of Equation 2 can be reduced to a simpler form as:

$$E(n) = \sum_{t=1}^{n} E_t = \alpha \sum_{t=1}^{n} \sum_{d \in D_t} itemEng(d) - n\beta \tag{6}$$

There are two parameters α and β in Equation 6. We need two equations to solve them. Therefore, we let t_1, t_2 be two different time slices in the life span of a news event, and s_1, s_2 be the life energy at respective time slice. Then, we have:

$$\begin{cases} E(t_1) = \sum_{t=1}^{t_1} E_t = \alpha \sum_{t=1}^{t_1} \sum_{d \in D_t} itemEng(d) - t_1 \beta = s_1 \\ E(t_2) = \sum_{t=1}^{t_2} E_t = \alpha \sum_{t=1}^{t_2} \sum_{d \in D_t} itemEng(d) - t_2 \beta = s_2 \end{cases} \quad (7)$$

let

$$y(n) = \sum_{t=1}^{n} \sum_{d \in D_t} itemEng(d) \quad (8)$$

where $y(n)$ means the total energy the news tweets provide to the news event till nth time slice.

Then, solve Equation 7, we get α and β as follows:

$$\alpha = \frac{s_1 \cdot t_2 - s_2 \cdot t_1}{t_2 \cdot y(t_1) - t_1 \cdot y(t_2)} \quad (9)$$

and

$$\beta = \frac{s_1 \cdot y(t_2) - s_2 \cdot y(t_1)}{t_2 \cdot y(t_1) - t_1 \cdot y(t_2)} \quad (10)$$

4 Single-Pass Clustering with Energy Function

If news events were to be sought from a time-ordered static collection, one solution would be to use document clustering techniques[9,6] to cluster the collection, and then to return the document from each cluster containing the earliest timestamp. However, we are interested in the strict on-line data, which has real-time constraints and imposes a single-pass restriction over the incoming data stream of tweets. The traditional single-pass clustering algorithm for news event detection on Twitter, is described as follows:

1. Build a term vector representation for the tweets and news events. The term vector of a news event is represented by the geometric center of all term vectors of its news tweets.
2. Compare a new tweet against the previous news events in memory.
3. If the tweet does not trigger any previous news events by exceeding a threshold, flag the tweet as containing a new event, and add the news event into the memory.
4. If the tweet triggers an existing news event, flag the tweet to this news event, and update the term vector of the news event by recomputing the geometric center.

There are two shortcomings of the traditional single-pass clustering algorithm. First, the threshold of single-pass clustering method is a constant, which is not very reasonable. When a news event is hot and its energy value is high, there are a lot discussions on Twitter. Therefore, the threshold should turn smaller. So that, tweets about the same news event with different contents can be clustered into one news event. When the news event is dying, the news tweet is few. The threshold should turn bigger, in case of other news tweets are clustered in this news event.

Second, the traditional single-pass clustering algorithm does not mention how long a news event should stay in the memory. It wastes the system memory and also increases the time cost. Because, a new tweet still need to be compared to the dead news event, and it even has a small chance to be flagged to the dead news event. In a word, this shortcoming could reduce the performance of a real system in all aspects.

We modify the traditional single-pass clustering algorithm with energy function to conquer the two problems described above. For the first one, we make the threshold denoted by θ a variable, which changes with the energy value as follows:

$$\theta = \begin{cases} \theta_{max}, & E > E_2 \\ \dfrac{\theta_{max} - \theta_{min}}{E_2 - E_1} \times E + \dfrac{\theta_{min} \times E_2 - \theta_{max} \times E_1}{E_2 - E_1}, & E_1 \leq E \leq E_2 \\ \theta_{min}, & E < E_1 \end{cases} \tag{11}$$

E represents the energy value of a news event. θ changes linearly with E. And it has an upper bound θ_{max} and a lower bound θ_{min}, when E reaches E_2 and E_1.

For the second one, we check the time of the last tweet of every news event in the memory periodically. At the check point, the time is T, and the time of the last tweet of a news event e_i is T_i. A time window W is given. If $T - T_i > W$, e_i should be dropped out of the memory. This time window is also change with the energy value computed by the energy function as follows:

$$W = \begin{cases} w_{max}, & E > E_2 \\ \dfrac{w_{max} - w_{min}}{E_2 - E_1} \times E + \dfrac{w_{min} \times E_2 - w_{max} \times E_1}{E_2 - E_1}, & E_1 \leq E \leq E_2 \\ w_{min}, & E < E_1 \end{cases} \tag{12}$$

w_{max} and w_{min} are the upper bound and lower bound of W, respectively.

5 Experiments and Evaluation

Before experiments and evaluation, we give a brief description of the dataset for this research work. Then, we solve the energy transferred factor α and decayed factor β of the energy function of our dataset. Finally, our news event detection method is compared with others.

5.1 Data Preparation

For the purpose of this research work, we crawled 900 headlines of news reports from November 2, 2010 to January 10, 2011 through the RSS(Really Simple Syndication) of the Associated Press website[1]. For each news report, we crawled the tweets which matches all the words in the headline. Because all the news reports are published by one news outlet, they seldom talk about the same news event. We suppose that each news report represents an independent news event. Besides, there were other tweets we did not crawl also discussed the news event, only because they did not match all the words in the headlins of news report. However, the tweets we crawled can be regarded as a sample of the whole news tweets set. In this dataset, there are more than 400 thousand tweets in total, and these tweets were posted by more than 130 thousands users.

Then we divided the dataset into two sets. One is training set, which is used to train the energy transferred factor α, decayed Factor β, and threshold for the single-pass clustering algorithm. We randomly chose 259 news events as the training set. The rest 641 news events constitute the testing set, which is used for evaluation and comparison.

5.2 Training Energy Transferred Factor and Decayed Factor

In this subsection, an iterative algorithm is proposed to solve the energy transferred factor α and decayed factor β. Before that, one more point to add is the maximum energy value of each news event, which will be used to solve the energy transferred and decayed factors. We suppose the maximum energy of every news event is proportional to the its activity. As a result, for news events e_1 and e_2, they have c_1 and c_2 news tweets, their whole life span are l_1 and l_2 hours, and their maximum energy values are $max(e_1)$ and $max(e_2)$, the assumption can be expressed as:

$$\frac{max(e_1)}{c_1/{l_1}^{\mu}} = \frac{max(e_2)}{c_2/{l_2}^{\mu}} \tag{13}$$

where, $0 < \mu < 1$. If $\mu = 1$, $c_1/{l_1}^{\mu} = c_1/l_1$ is the average activity of news event e_1; If $\mu = 0$, $c_1/{l_1}^{\mu} = c_1$ is the total activity. So, when $0 < \mu < 1$, it can be regarded as the mixture of the average and the total activity. In our experiment, μ is set to 0.6.

Therefore, in the training set, if we set the maximum energy value of the news event e_{max} to 1.0, for other news event e, the maximum energy value is:

$$max(e) = \frac{c_e}{c_{e_{max}}} \times (\frac{l_{e_{max}}}{l_e})^{\mu} \tag{14}$$

The iterative algorithm to solve the energy transferred factor α and energy decayed factor β are described as follows:

[1] http://hosted.ap.org/lineups/TOPHEADS-rss_2.0.xml?SITE=ILMOL&SECTION=HOME

1. compute the maximum energy value $max(e)$ of every news event e in the training set by Equation 14.
2. compute the energy of every tweet in the training set by Equation 5, where λ_1 and λ_2 are set to 0.3 and 0.7, empirically.
3. for every news event e in the training set:
   (a) initialize $t_1 = 0.3l_e$, $s_1 = 0.7max(e)$, $t_2 = N + 1$, $s_2 = 0.0$, $t_{max} = 0$
   (b) repeat (c)(d)(e), until t_{max} does not change any more
   (c) compute α and β by Equation 8, 9, 10
   (d) find the maximum energy value and the time slice t_{max} using α and β above by Equation 2
   (e) reset $t_1 = t_{max}$, $s_1 = max(e)$, $t_2 = N + 1$, $s_2 = 0.0$
4. compute the average value of all α and β of all news event in the training set as the final results.

The final results are: $\alpha = 0.00110091$, $\beta = 0.00654238$. We also give all results of α and β for all news events in the training set in Figure 2.

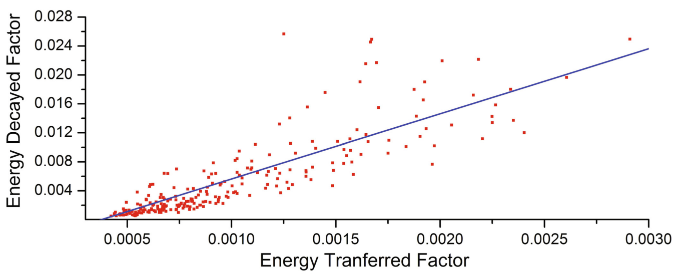

Fig. 2. Energy decayed factor vs. energy transferred factor

It is clear to see that there is a obvious linear correlation between the energy transferred factor and decayed factor. The main reason may be that all news events follow almost the same tweets distribution with time. [15] considered this distribution is an exponential distribution. So, using the average value of all energy transferred factors and decayed factors as the final results is appropriate.

5.3 News Event Detection Comparisons

In this experiment, our method(A) is compared to two other methods. The baseline method(B) is the traditional single-pass algorithm. The other is a fixed time-window single-pass clustering method(W). This fixed time-window method is a traditional single-pass clustering method added with a **fixed time window** W_{fixed}. In this method, it also check the news event in the memory periodically.

If there is no new web document for a news event more than W_{fixed} time, this simple modified method will consider the news event is over, and delete it from memory.

All the three methods group the tweets in the test set into several clusters. Five official TDT measures[5] including: precision(p), recall(r), miss(m), false alarm(f) and F1-measure($F1$) are used to evaluate the results of these three methods.

Table 1 shows the results. W4, W8, W12 are the fixed-time-window method with a fixed time window of 4, 8, 12 hours. In our method, the W_{min} and W_{max} in Equation 12 are set to 4 and 12 hours. So it is reasonable to compare our method with W4, W8 and W12.

Table 1. Results of TDT measures

	p	r	m	f	$F1$
B	0.877939	0.904714	0.095286	0.000285	0.891125
W4	0.947464	0.499588	0.500412	0.000065	0.654215
W8	0.941955	0.692312	0.307688	0.000096	0.798066
W12	0.932875	0.779466	0.220534	0.000124	0.849298
A	0.914556	0.876216	0.123784	0.000301	**0.894976**

In Table 1, all fixed time-window methods out-performance the baseline method a little in precision. However, their recalls are too low to accept. Thus, the baseline method has a better F1-measure than all fixed time-window methods. For our method, it achieves both reasonable precision and recall, which results in the best F1-measure of all methods.

Besides, our method also has acceptable time and memory performance. Figure 3 shows the time and memory performance of all three methods. The red lines are our method.

In Figure 3(a) and 3(b), The time cost of the traditional single-pass clustering method increases as the square of the number of tweets processed, while the fixed time-window method and our method are increase almost linearly. Our method is a little slower than the fixed time-window method. Because it needs a few more computational works of the changing threshold and time window, which is worthwhile. As there is a big improvement in official TDT measures, especially in recall.

In Figure 3(c) and 3(d), the memory cost is represented by the number of clusters in memory. The number of clusters of Traditional single-pass method increases linearly with the number of tweets processed, while the fixed time-window method and our method both fluctuate around a small constant.

Generally speaking, our method has the best results in the official TDT measures and it also has quite acceptable time and memory performance, which makes it suitable for a real system.

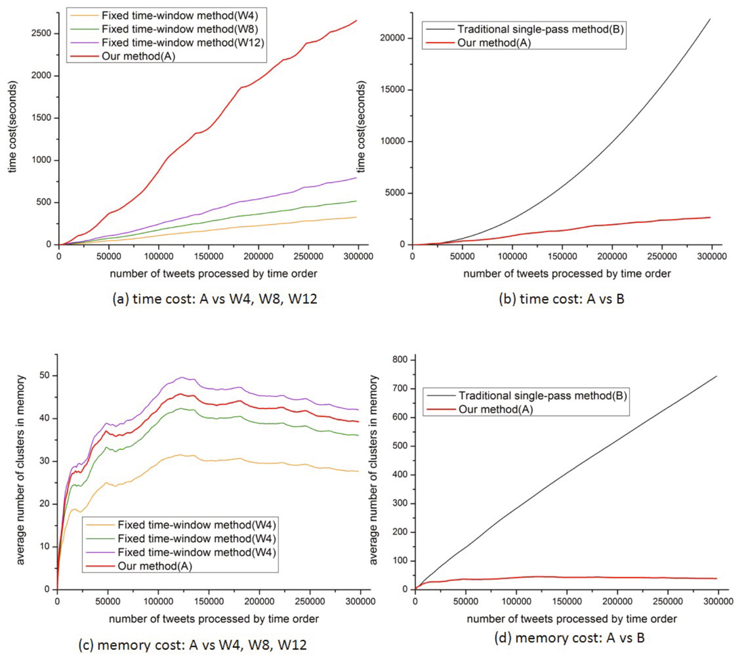

Fig. 3. Comparison of time and memory cost

6 Conclusions

In this paper, we report a novel news event detection method of Twitter. Experimental results show that it performances well not only in the official TDT measures, but also in time and memory cost.

Although the proposed method is quite good for a real system, there are still two major points needed to be improved. First, the energy transferred factor and decayed factor could also change with the energy value itself. When the news event is active, the energy transferred factor could be a little bigger, while the energy decayed factor could be a little smaller. Second, the user influence, which measures the energy of a single tweet, could use a more reliable and effective model. Moreover, our method can also be more generalized in other time sequential data mining, in the future.

References

1. Milstein, S., Chowdhury, A., Hochmuth, G., Lorica, B., Magoulas, R.: Twitter and the micro-messaging revolution: Communication, connections, and immediacy–140 characters at a time. O'Reilly Radar Report (2008)
2. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: WWW 2010, pp. 591–600 (2010)

3. Huberman, B.A., Romero, D.M., Wu, F.: Social networks that matter: Twitter under the microscope. Social Computing Laboratory, HP Labs (2008)
4. Weng, J., Lim, E.P., Jiang, J., He, Q.: TwitterRank: finding topic-sensitive influential twitterers. In: WSDM 2010, pp. 261–270 (2010)
5. Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yang, Y.: Topic detection and tracking pilot study: Final report. In: Proc. of the DARPA Broadcast News Transcription and Understanding Workshop (1998)
6. Yang, Y., Pierce, T., Carbonell, J.: A study of retrospective and on-line event detection. In: SIGIR 1998, pp. 28–36 (1998)
7. Chen, C.C., Chen, Y.-T., Sun, Y., Chen, M.C.: Life Cycle Modeling of News Events Using Aging Theory. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) ECML 2003. LNCS (LNAI), vol. 2837, pp. 47–59. Springer, Heidelberg (2003)
8. Lu, R., Yang, Q.: Extracting News Topics from Microblogs based on Hidden topics discovering and Text Clustering. In: CCIR 2010, pp. 291–298 (2010)
9. Salton, G.: Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley Longman Publishing Co., Inc. (1989)
10. Newman, M.E.J., Park, J.: Why social networks are different from other types of networks. Phys. Rev. E 68(3), 36–122 (2003)
11. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: Proc. of 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, pp. 56–65 (2007)
12. Honey, C., Herring, S.C.: Beyond Microblogging: Conversation and Collaboration via Twitter. In: Proc. of the 42nd Hawaii International Conference on System Sciences, pp. 1–10 (2009)
13. Zhao, D., Rosson, M.B.: How and why people Twitter: the role that micro-blogging plays in informal communication at work. In: Proc. of the ACM 2009 International Conference on Supporting Group Work, pp. 243–252 (2009)
14. Coon, M., Reeves, B.: Social Media Marketing: Successful Case Studies of Businesses Using Facebook and YouTube With An In-Depth Look into the Business Use of Twitter (2010)
15. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: WWW 2010, pp. 851–860 (2010)
16. Menczer, F., Belew, R.K., Willuhn, W.: Artificial Life Applied to Adaptive Information Agents. In: AAAI 1995 (1995)
17. Kempe, D., Kleinberg, J., Tardos, É.: Influential Nodes in a Diffusion Model for Social Networks. In: Caires, L., Italiano, G.F., Monteiro, L., Palamidessi, C., Yung, M. (eds.) ICALP 2005. LNCS, vol. 3580, pp. 1127–1138. Springer, Heidelberg (2005)
18. Leavitt, A., Burchard, E., Fisher, D., Gillbert, S.: New approaches for analyzing influence on twitter. A publication of the Web Ecology project (2009)
19. Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Network and ISDN Systems, 107–117 (1998)
20. Papka, R., Allan, J.: On-Line New Event Detection using Single Pass Clustering. University of Massachusetts, Amherst (1998)
21. Van Rijsbergen, C.J.: Information Retereval, 2nd edn. Butterworths, Massachusetts (1979)
22. Cataldi, M., Di Caro, L., Schifanella, C.: Emerging topic detection on Twitter based on temporal and social terms evaluation. In: Proc. of the Tenth International Workshop on Multimedia Data Mining, pp. 1–10 (2010)
23. Phuvipadawat, S., Murata, T.: Breaking News Detection and Tracking in Twitter. In: International Conference on Web Intelligence and Intelligent Agent Technology, pp. 120–123 (2010)