2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

Breaking News Detection and Tracking in Twitter

Swit Phuvipadawat, Tsuyoshi Murata
Department of Computer Science, Graduate School of Information Science and Engineering
Tokyo Institute of Technology, Japan
swit.p@ai.cs.titech.ac.jp, murata@cs.titech.ac.jp

***Abstract*—Twitter has been used as one of the communication channels for spreading breaking news. We propose a method to collect, group, rank and track breaking news in Twitter. Since short length messages make similarity comparison difficult, we boost scores on proper nouns to improve the grouping results. Each group is ranked based on popularity and reliability factors. Current detection method is limited to facts part of messages. We developed an application called "Hotstream" based on the proposed method. Users can discover breaking news from the Twitter timeline. Each story is provided with the information of message originator, story development and activity chart. This provides a convenient way for people to follow breaking news and stay informed with real-time updates.**

***Keywords*-Twitter, Topic Detection and Tracking, Real-time text-mining, Information Retrieval**

I. Introduction

Twitter is a social networking service that allows users to share information, which is described by Twitter as "What's happening?" in a form of short texts (140 characters). Main characters of Twitter are: brevity—contents are in short length and simultaneousness—contents are updated frequently. Twitter has transformed the way people convey information especially in the areas of news.

In June 2009, Twitter has played an important role in delivering user-generated contents from the Iranian citizen in the Iran election. We see that people with technology played a role of journalists in the situation where news reporting in a conventional way has been made difficult [1]. Anyone who is not associated to the media industry can also deliver news. Thus, Twitter presents a highly effective way to discover what is happening around the world.

Breaking news is defined by Wiktionary [2] as *"news that has either just happened or is currently happening. Breaking news may contain incomplete information, factual error or poor editing because of rush."* With this definition Twitter can fit the needs of breaking news delivery.

However, news posted in Twitter requires an effort to discover it. Firstly, users often have problems of deciding which users to follow. That is, to find users with interesting tweets [3]. Secondly, users need to read through status updates and follow links to obtain further information. To ease these problems and to deliver breaking news effectively, we propose a method to collect, group, rank and track breaking news in Twitter. This work is a contribution to the area of Topic Detection and Tracking (TDT) [4]. The tasks we focus are first story detection, cluster detection, and tracking.

II. Characteristics of breaking news in Twitter

As a preparatory experiment for analyzing characteristics of breaking news, we collected messages from Twitter using the Twitter API. The data contains 121,000 messages from public statuses and 33,000 messages from a selected group of 250 users who contribute to breaking news postings in Twitter. We selected users who use a breaking news hash tag (*#breakingnews*) in their messages.

Table I
Characteristics of messages in Twitter based on 154,000 message samples

Characteristic	No. of occurrences	Percentage
Tag a user	79,469	51.6%
Embed a link	50,404	32.7%
Retweet	29,935	19.4%
Use a hash tag	20,348	13.2%

Table I shows characteristics of messages and the number of occurrences. In contribution to breaking news detection, these characteristics help us find more facts about a message. From user tags, we can identify conversations between users. From embedded links, we can follow them to find more information. Retweet means to repost another user's message. From a number of retweets, we can determine popularity or importance of a message. And from hash tags, we can group together related messages. A retweeted message often contains the information of message originator and previous message. There are two aspects to consider when detecting the breaking news in Twitter: Single message aspect and Timeline aspect. The two aspects are described in details as follows.

A. Single message aspect

There are two important elements in a message: emotions and facts. The inclusion of emotions in the message makes news delivered in Twitter, different from news delivered by professional journalists. Although there are cases where emotions are conveyed in conventional news, expression of emotions occurs much more often in Twitter messages. Emotions are expressed through the use of symbols (mainly

978-0-7695-4191-4/10 $26.00 © 2010 IEEE
DOI 10.1109/WI-IAT.2010.205

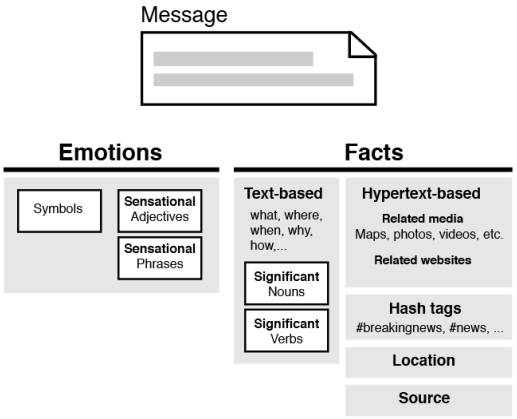

Figure 1. Emotions and facts in a message

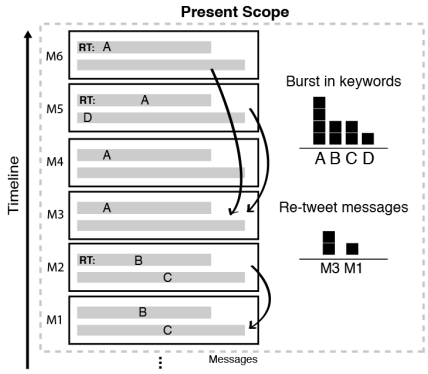

Figure 2. Burst in Keywords and Re-tweet messages

the use of exclamation mark '!'), the use of sensational adjectives and phrases: crazy, amazing, great, terrible, wonderful, shocking, oh my god, etc.

Facts are provided in text-based, hypertext-based, and through location and source information of the message originator. Text-based information is highly significant as it helps interrogate the details of the news in terms of 'what', 'where', 'when', 'how', etc. We can identify keywords from facts that contribute to news story. These keywords are identified as significant nouns and verbs. Significant nouns include keywords found in conventional news, names of famous places, people and events such as Japan, US president, emergency and airplane. Significant verbs are, for examples, fire, crash, bomb, survive, rescue, win, etc. Users often tag their message with a hash symbol (#) followed by keywords for examples, #breakingnews, #haiti, etc. as a mean to group together messages related to the keywords. Hypertext-based facts provide related information from external sources. To cope with the limitation of the message length, users often use a link shortening services like TinyURL[1] and bit.ly[2]. In addition to texts, users often include maps and pictures. Maps are provided by online services like Google Maps[3] and Yahoo Maps[4]. Pictures are hosted on services like TwitPic[5] , yfrog[6], etc.

B. Timeline aspect

From the timeline aspect shown in figure 2, we can observe the burst in keywords and the number of retweeted messages through the passage of time. Interesting or important messages tend to be retweeted more than the others. For the case of breaking news we can see the development of news story through the series of messages. We will use these facts to determine the ranking of messages.

[1]http://tinyurl.com
[2]http://bit.ly
[3]http://maps.google.com
[4]http://maps.yahoo.com
[5]http://twitpic.com
[6]http://www.yfrog.com

III. Methodology

In this paper, we present a method to collect, group, rank and track breaking news. Tasks are divided into two stages: story finding and story development. In this paper, we focus on facts part of messages. Emotions are left for our future works. The overview of the process is shown in figure 3.

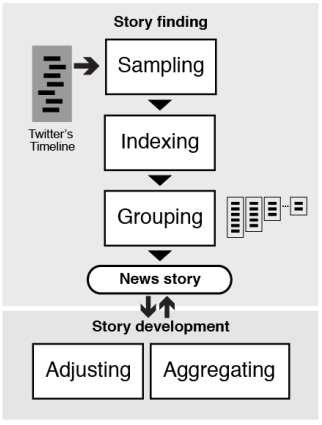

Figure 3. Two stages: story finding and story development

A. Story finding stage

In this stage, the tasks are presented in three steps: sampling, indexing and grouping.

1) *Sampling*: In this experiment, messages are fetched through the Twitter streaming API [5] using pre-defined search queries to get near real-time public statuses. Pre-defined search queries are, for example, hash tags users often use to annotate breaking news e.g. #breakingnews and "breaking news" keyword.
2) *Indexing*: To accommodate the process of grouping similar messages, an index based on the content of

messages is constructed. In this experiment, we use Apache Lucene [6].

3) *Grouping*: Messages that are similar to each other are grouped together to form a news story. Similarity between messages is compared using TF-IDF [7], [8].

The similarity between two messages is defined as:

$$sim(m_1, m_2) = \sum_{t \in m_1} [tf(t, m_2) \times idf(t) \times boost(t)] \quad (1)$$

$$tf(t, m) = \frac{count(t \text{ in } m)}{size(m)} \quad (2)$$

$$idf(t) = 1 + \log \left[\frac{N}{count(m \text{ has } t)} \right] \quad (3)$$

$boost(t)$ is raised for proper noun terms e.g. China, England, Eurostar, Haiti and Twitters artifacts like hash tags and usernames to improve the score on identifiable keywords. We use the Stanford Named Entity Recognizer (NER) [9] for the classification of proper nouns. NER provides a general implementation of linear chain Conditional Random Field (CRF) sequence models, coupled with well-engineered feature extractors for Named Entity Recognition.

The algorithm for a message-group assignment is shown below.

Algorithm 1 Assign message m into a group in G

```
for g in G do
    Score[g] ← Sim(m, g.firstDoc, g.topTerms)
end for
if Max(Score) > MergeThreashold then
    Assign(m, Max(Score).groupId)
else
    groupId ← Group.create()
    Assign(m, groupId)
end if
return G
```

To ensure that messages in the groups are related to the first story and to allow further messages to develop upon previous messages we will compare a message with the first message in a group and the top k terms in that group. In this experiment we set k to 10. We then assign a message to a group if the score exceeds a pre-defined threshold called $MergeThreshold$. The results to this point are groups of messages or news stories. The score for each group is computed as follows:

$$S = w_1 \sum_{u \in g_i} No.Follower(u_i) + w_2 No.Retweet(g_i) \quad (4)$$

$$Score(g_i) = \frac{1}{Z} \sum_{n=1}^{l} \left[\frac{S}{\log(\Delta_n + 2)} \right] \quad (5)$$

$$\Delta_n = t_{current} - t_n \quad (6)$$

A raw group score S in (4) is based on reliability and popularity factors. Reliability is determined from the numbers of followers from all the users who posted messages in the group. Popularity is determined from the numbers of retweet within the group. The final group score in (5), is adjusted based on the freshness of messages. To this effect a group that has newer messages receives a higher score than a group with older messages. The computation is done on l last messages in a group. Δ_n in (6) is the difference between the current time and the time where a message is created. Z is the normalizing factor.

B. Story development stage

In the subsequent stage, each news story is adjusted with appropriate ranking through a period of time. In addition new findings from external source (outside of Twitter) such as news articles of reliable news source or media like photos and video footages can be aggregated to existing news stories.

IV. Experiment

We show the result when similar news containing messages are grouped together. The sample messages have been collected in February 2010. Each message is given with number and group label. For the purpose of demonstration, we selected 10 messages $M0_{(label)} - M9_{(label)}$ and give 5 labels as follows: (1) for Toyotas brake problems. (2) for Michael Jacksons Doctor. (3) for Heavy snow storm in the U.S. (4) for U.S. military base issues in Okinawa, Japan. (5) for escaped prisoners in Haiti. In this experiment, we show that raising the importance of proper nouns can improve the grouping result.

We employ the grouping method described in section III. Table II shows results from 4 configurations: No boost for proper nouns, with boost for proper nouns by raising the term score to the power of 1.5, 1.7 and 2 respectively.

Table II
Grouping results based on proper noun boost values

(a) No boost	
G0	$M3_{(2)}$ $M4_{(2)}$ $M5_{(2)}$
G1	$M7_{(4)}$ $M8_{(4)}$
G2	$M0_{(1)}$ $M1_{(1)}$
G3	$M2_{(2)}$
G4	$M6_{(3)}$
G5	$M9_{(5)}$

(b) $boost(t) = 1.5$	
G0	$M2_{(2)}$ $M3_{(2)}$ $M4_{(2)}$ $M5_{(2)}$
G1	$M7_{(4)}$ $M8_{(4)}$
G2	$M0_{(1)}$ $M1_{(1)}$
G3	$M6_{(3)}$
G4	$M9_{(5)}$

(c) $boost(t) = 1.7$	
G0	$M0_{(1)}$ $M1_{(1)}$ $M7_{(4)}$ $M8_{(4)}$
G1	$M2_{(2)}$ $M3_{(2)}$ $M4_{(2)}$ $M5_{(2)}$
G2	$M6_{(3)}$
G3	$M9_{(5)}$

(d) $boost(t) = 2$	
G0	$M2_{(2)}$ $M3_{(2)}$ $M4_{(2)}$ $M5_{(2)}$ $M9_{(5)}$
G1	$M0_{(1)}$ $M1_{(1)}$ $M7_{(4)}$ $M8_{(4)}$
G2	$M6_{(3)}$

It is necessary to raise an importance of proper nouns because the length of messages in Twitter is short, and may

not have enough information for comparison. For example a web page usually contains a larger set of terms when compared with Twitter messages. The average term size after the removal of stop words for messages in Twitter according to our data set is 7. If we do not place the importance on particular terms then we cannot find related messages that have weak similarity based on a traditional TF-IDF scheme. One such example is shown in table II(a). $M2$ does not have a similarity score high enough to be grouped with $M3$, $M4$ and $M5$. However if we raise the score for proper nouns, grouping results can be adjusted. By raising $boost(t)$ to 1.5, we can achieve the correct grouping. It is important to choose the appropriate value for $boost(t)$, as it defines how sensitive the grouping is. Table II(c) and (d) show results when grouping based on proper nouns is too sensitive.

We prototyped a web application based on the method described in section III called Hotstream. The purpose is to construct a real-time news portal featuring breaking news and popular stories from Twitter. Figure 4 (a) shows the front page of Hotstream accessed on February 5th, 2010. The page contains the list of top stories within 24 hours.

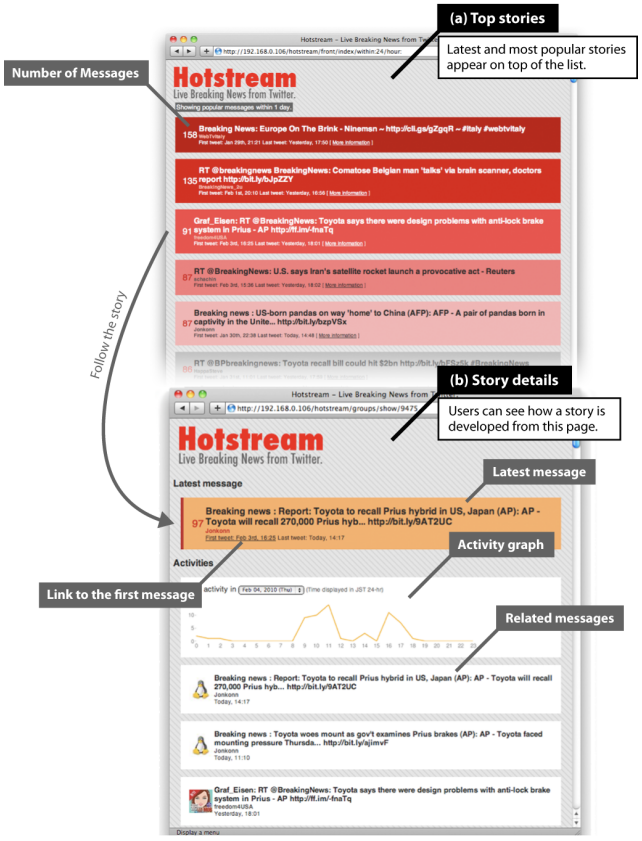

Figure 4. Hotstream, "Live breaking news from Twitter", application screenshots. Top (a) front page showing top stories, bottom (b) story details.

Users can click on each story to find more details. Figure 4 (b) shows a timeline of a story. We show the number of messages along with the first message in the story. The top message is the latest message in the story. In addition, the activity graph based on the number of messages in the story with respect to time is displayed.

V. Conclusion and Future Works

In this work, we discussed the characteristics of breaking news in Twitter and presented a method to collect, group, rank and track breaking news from Twitter. To improve the similarity comparison for short-length messages, we put an emphasis on proper nouns. There are several factors to rank the news story. In this experiment we use reliability, popularity and freshness for the ranking factors.

An application based on the proposed method called Hotstream is developed. This application shows a high potential for an intelligent news portal based on Twitter. We believe that such application can present user generated contents to the mass audience efficiently.

For future works, we will explore ways to utilize emotion information from messages and consider the network structure of retweet messages. With these counterparts, we can understand the users perception to news stories, impacts to the mass audience and the pattern in which the information spreads.

References

[1] E. Morozov. Iran Elections: A twitter Revolution? *The Washington Post*, June 17, 2009. http://www.washingtonpost.com/wp-dyn/content/discussion/2009/06/17/DI2009061702232.html.

[2] Wiktionary. http://en.wiktionary.org/wiki/breaking_news. Accessed February 1, 2010.

[3] R. Mateosian. Micro Review: Twitter. *Micro, IEEE*, 29, Issue 4:87–88, July-August 2009.

[4] J. Allen. *Topic Detection and Tracking*, pages 17–30. Kluwer Academic, Norwell, Massachusetts, 2002.

[5] Twitter Streaming API. http://apiwiki.twitter.com/Streaming-API-Documentation. Accessed February 1, 2010.

[6] Apache Lucene. http://lucene.apache.org. Accessed February 1, 2010.

[7] C. D. Manning, P. Raghavan, and H. Schütze. *Introdution to Information Retrieval*, pages 108–115. Cambridge University Press, New York, 2008.

[8] Similarity (Lucene 3.0.0 API). http://lucene.apache.org/java/3_0_0/api/all/org/apache/lucene/search/Similarity.html. Accessed February 1, 2010.

[9] J.R. Finkel, T. Grenager, and C. Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370, 2005.