Mining Micro-Blogs: Opportunities and Challenges

Yang Liao[1], Masud Moshtaghi[1], Bo Han[1], Shanika Karunasekera[1], Ramamohanarao Kotagiri[1], Timothy Baldwin[1], Aaron Harwood[1], and Philippa Pattison[2]

[1] Department of Computer Science and Software Engineering,
The University of Melbourne, VIC, Australia

[2] Faculty of Medicine, Dentistry and Health Sciences Psychological Sciences,
The University of Melbourne, VIC, Australia

Summary. This chapter investigates whether and how micro-messaging technologies such as Twitter messages can be harnessed to obtain valuable information. The interesting characteristics of micro-blogging services, such as being user oriented, provide opportunities for different applications to use the content of these sites to their advantage. However, the same characteristics become the weakness of these sites when it comes to data modeling and analysis of the messages. These sites contains very large amount of unstructured, noisy with false or missing data which make the task of data mining difficult. This chapter first reviews some of the potential applications of the micro-messaging services and then provides some insight into different challenges faced by data mining applications. Later in the chapter, characteristics of a real-data collected from the Twitter are analysed. At the end of chapter, application of micro-blogging services is shown by three different case studies.

1 Introduction

With the wide uptake of the Internet, micro-blogging services such as Tumblr and Twitter have become popular means of communication. Most of today's popular social networking sites such as Facebook and MySpace also support micro-blogging features. Superficially, the main factor differentiating micro-blogging from traditional blogging is the limited message size, with a hard limit on messages in micro-blogging services typically of around 150 characters. Micro-blogs have evolved to include rich social networking features, however, most notably via the ability to "follow" another user, and thereby receive the feed of all messages posted by them. Via this social networking feature, information propagation in micro-blogs resembles epidemic propagation in social communities. The highly connected nature of these dynamic networks and the explosive nature of message passing leads to rapid and efficient data dissemination and very targeted information fluxes.

Liao, Yang, Masud Moshtaghi, Bo Han, Shanika Karunasekera, Ramamohanarao Kotagiri, Timothy Baldwin, Aaron Harwood and Philippa Pattison (to appear) Mining Micro-Blogs: Opportunities and Challenges, In Ajith Abraham and Aboul Ella Hassanien (eds.) Social Networks: Computational Aspects and Mining, Springer.

Although micro-blogging technologies were originally created as a means of personal communication, they have many unique characteristics that offer opportunities beyond simple communication, and make them a ripe target for data mining. Some applications that have recently been explored over micro-blog data are disaster detection [21], trend identification [4] and online marketing [6]. However, effective mining of micro-blogging sites have associated challenges.

In this chapter we discuss and demonstrate the opportunities and the challenges associated with data mining in micro-blogs, with specific focus on Twitter. We classify the information generated in micro-blogs into three categories, and identify important characteristics of information in each of them. A high-level data mining architecture for micro-blogs is then presented. We identify a number of applications which fall into three application areas: event detection, trend identification, and social behaviour analysis. We analyse characteristics of Twitter data based on a data set which we collected over a four week period. This chapter also presents three case studies based on Twitter data. Based on the observation that number of messages related to an event increases due to elevated user interest in the event, the first case study demonstrates how epidemic models combined with frequency domain deconvolution techniques can be used for event identification. The second case study demonstrates how a Markov chain model and a distance computation algorithm can be used for identifying social clusters. The third case study demonstrates how the frequency of keywords within clusters can be used for trend identification.

Section 2 describes some of the characteristics of micro-blogs. Section 3 discusses how these characteristics provide opportunities for data mining to support different classes of applications. The challenges faced in text mining micro-blogs are discussed in Section 4. Section 5 shows the analysis of Twitter data we collected over a four week period and Section 6 shows three different case studies of using Twitter data for different applications.

2 Characteristics

Micro-blogs provide a means for users to generate content and connections. In this section, we briefly review the general characteristics of micro-blogs and classify these characteristics into three categories: *Users*, *Social Connections*, and *Messages*.

2.1 User Properties

Users are the content generators of micro-blogs. Due to privacy issues and the fear of identity theft, users usually limit the public information they share about themselves to fields such as name, location, and spoken language(s). Privacy also affects the reliability of the shared information. For example, a user providing a bogus location could result in erroneous interpretation of information in content they post. In addition, micro-blogs commonly contain *Virtual Users* with a very short active life as a means of identity obfuscation. For example, a business owner might want to start a rumour about the quality of the products of a rival company using multiple virtual users. This property further reduces the reliability of the data on micro-blogs. It also underscores the importance of selecting data analysis approaches which can deal with uncertainty in the information.

Similar to other types of social networks, behavioural properties of users in micro-blogs change with time. Level of activity, location, and even the language used by the user to share information may change. Also, users come and go with time.

2.2 Social Connections

Social connections create the connectivity in micro-blogs. Social connections in a micro-blog can be classified into two categories: *social links* and *social interactions.* Social connections are highly dynamic, and are created and deleted between users over time.

Social Links: These links are created between two users and last till one or both users decide to end the relation. These links are either *bidirectional* or *unidirectional.* Bidirectional links, also known as friendship links, are mutual connections where both users are interested in the content generated by each other. On the other hand, with unidirectional links — also known as follower links — one user is interested in the content generated by the other, but not vice versa. Social links can also be classified into *direct* and *indirect* links. While direct links directly connect users, it is also possible for users to be connected indirectly through their selection of followed threads.

Micro-blogs usually contain large numbers of social links. The number of social links in a network directly affects the propagation of content, and as such, social structure is an important aspect of micro-blogs.

Social Interactions: Social interactions are one-off links between two users, regardless of social links. Interactions in micro-blogs are made via messages. One form of such interaction is a user posting a message in reply to another post or user in a network. Other forms of interaction are: relaying another user message, and mentioning the name of other users in the body of a message. Social links are potentially more socially meaningful than one-off interactions.

2.3 Message Properties

In micro-blogs, messages — also known as *status updates* or *posts* — are short and usually limited to around 150 characters. These short messages allow users to share personal and professional information, as well as links to other sites. The length limit forces users to be succinct in expressing the information they want to share, which in turn leads to various linguistic devices for abbreviating the message length, and creates challenges for language processing techniques, as we survey in Section 4.

Based on the above classification of the in the data in micro-blogs, we propose the high-level data mining architecture shown in Figure 1 for identification of events and trends from mining micro-blogs. As shown, the main activities involved are data collection, natural language processing (NLP), social semantic analysis, and knowledge extraction. The first step in data mining process is data collection. The data from micro-blogs can be collected using web crawlers and official APIs. Though a complete dump is very useful for knowledge extraction, privacy constraints and the sheer volume of data means there are access restrictions on the data, influencing the design of data mining techniques. Our basic approach is to use NLP techniques and social semantic analysis to extract features from raw data of users and their connections as well as message information in micro-blogs, for further processing for knowledge discovery.

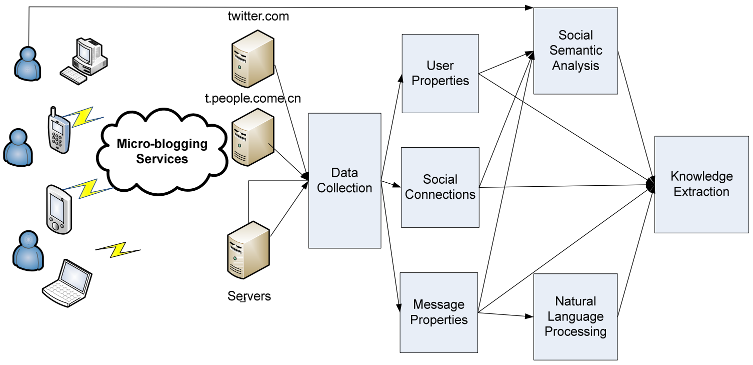


Fig. 1. Schematic representation of the activities related to data mining in micro-blogs

3 Opportunities

Several characteristics of the data available in micro-blogs make them appealing for applications. In this section we outline a number of possible applications, which fall into three broad categories: event detection, trend identification, and social group identification. Events correspond to real-world happenings which usually have a short-term interest period among micro-bloggers; *event detection* techniques try to predict or identify an event based on user messages. *Trend identification* relates to analysis of the effects of real-life developments on user opinions and behaviour. Finally, *social behaviour analysis* deals with relations between users, and tries to group users based on their social links, social interactions, and interests. Note that these three categories might be used in tandem in a single application. For example, an event detection technique might be used to identify an event, then the extent of the event can be predicted by trend inference or identification techniques, while social group analysis could aid in the preparation of a response.

3.1 Event detection

Micro-messages tend to be constrained to essential facts due to the enforced brevity, which makes automated analysis more efficient. These characteristics can be exploited to help identify and provide rapid response in emergency situations, including natural disasters such as bush fires, and accidental or deliberate chemical, biological, and radiological releases. Our first case study in Section 6 (quantitative analysis of information propagation in Twitter) is used for identifying the new events.

Disaster Detection: Micro-blogs can play an important role as an emergency alert system. For example, during the recent Haiti earthquake, Mumbai terror attacks, and political unrest in Iran, many people found Twitter to be a useful means of getting real-time updates on the situation. In some recent natural disasters, these networks are reported to have been able to even beat commercial news networks such as CNN and BBC with situation updates. Because of the effectiveness of these social networks, some organizations (Red Cross and some government agencies in USA) have made use of these networks to promptly provide updates on evacuation routes and other information.

During times of major disasters, telecommunication networks have been known to fail due to traffic overload. Internet micro-messaging communication technologies are able to alleviate traffic overloads due to queued transmission and the limited message size. The failure to provide adequate warnings to affected communities on Black Saturday in Australia was one of the key contributors to the large number of casualties in the bush fire. Therefore, investigating the suitability of these emerging micro-messaging technologies to complement the existing communication techniques is of vital importance.

In recent work, Twitter has been leveraged to detect earthquakes and send alerts to relevant communities ahead of impending shockwaves [21]. In this work, an event detection algorithm was developed on the basis of Twitter users being social sensors for a disaster. First, the posting time and volume of earthquake-related Tweets was modelled as an exponential distribution, and a Kalman filter and Particle filters were applied for location estimation.

Anomalous Change Detection: Internet-based syndromic surveillance using predefined keywords such as disease symptoms has been used in the area of epidemiology, as a possible means of early detection of infectious disease outbreaks [4]. Similarly, rapid increases in message traffic related to specific keywords occurring in micro-messaging systems may be used to detect anomalous events.

Research has shown the potential use of micro-blogs in providing early warning for events like Swine Flu outbreak [17, 20]. In addition, first story detection within the Twitter stream has been addressed in recent work [16]. By finding the relaxed nearest neighbour of a Tweet, an optimized Locality Sensitive Hashing (LSH) algorithm is used to meet the need of high volume data in speed and memory usage.

3.2 Trend identification

In many applications we are interested in finding the effects of a phenomenon, such as its extent or user opinions on it. The effects of a phenomenon can be analyzed through trend identification in micro-blogs. In trend identification, unlike event detection, the trigger is known a priori and the focus is on its consequence.

Opinion Polls: Micro-blogs are a great way for users to express their opinion. Analysis of the sentiments of users and extraction of meaningful information from user messages by means of NLP techniques enables the determination of user opinions. This information can be used as a supplement to traditional polling [3]. Separately, conventional positive/negative sentiment detection over micro-blog messages has been investigated [2].

Marketing: Market analysis provides feedback to companies. Mining micro-blogs may facilitate market analysis by providing cheaper, faster and more comprehensive information about their products and market campaigns. In Section 6, we demonstrate how to identify short-term events using an epidemic model; the same methodology is useful in collecting feedback for short-term marketing campaigns, such as immediate reactions to a new advertisement. On the other hand, the long-term trend of the market may be observed using the approach that we introduce in the third case study, where Twitter messages are used to predict attendance trends at the World Expo 2010 in Shanghai, China.

3.3 Social group identification

It is interesting to study reciprocated and repeated uni-directional relations between users of a micro-blog through analysis of social connections between users, for example to find out who is the most influential user in the network, or what makes one user follow another. User profile and network analysis can offer hints in this regard.

TwitterRank [27] has been proposed to determine the most influential users on different topics in Twitter. The topologies of Twitter users can also be analyzed for relationship distribution [12]. Alternatively, it is possible to develop recommender systems for users to recommend a set of users that a given user is likely to benefit from following [8].

In our second case study, we introduce an experiment for grouping users into clusters by measuring the velocity of information flows between them; the results show that users in the same cluster have same or similar backgrounds, such as geological location, interests and age groups.

4 Text Mining Challenges

Micro-blog messages differ from conventional text. They feature many unique symbols like mentions, hashtags and urls, and the popular use of colloquial words and Internet slang. Message quality varies greatly, from newswire-like utterances (e.g. *The United Nations Security Council will hold an emergency meeting Sunday on tensions in the Korean Peninsula*) to babble (e.g. *O_o haha wow*). In terms of text processing, there are significant research challenges, as outlined below.

First, popular micro-blogs such as Twitter attract users from a variety of language backgrounds, and are thus highly multilingual in content. This language diversity poses obstacles in text processing, as most text processing tools such as word tokenizers and syntactic parsers are language dependent. Thus, it is important to perform language identification before further text processing.

Second, typos, ad hoc abbreviations, phonetic substitutions and ungrammatical structures in micro-blog messages hamper text processing tools [23, 19, 7]. For example, given the Twitter message *I was talkin bout u lol* (or in standard English: *I was talking about you (lol).*), the Stanford parser [11, 5] analyzes *talkin bout u* as a noun phrase rather than a verb phrase. Noise of this type restricts the performance of text mining without proper normalization or in-built robustness in the text processing.

Third, micro-blog data is user generated and as such subjective and, at times, unreliable in content. Anyone who can access the micro-blog service can post a message, possibly containing false or offensive information. The unreliability of micro-blog data can cause grief for applications such as information extraction, and analysis of the authority and trustworthiness of different users/messages is a significant challenge.

In addition, algorithm efficiency is critical for data mining over popular micro-blogs due to the high rate of data generation: around 65 millions tweets were posted on Twitter per day in June 2010, for example [24]. In order to keep pace with the real-time stream of data, processing time must be kept to a minimum, particularly for real-time applications like event alert services where the response time is critical [16].

In summary, although micro-blogs are a highly attractive target for knowledge extraction due to the large amount of real-time data generated by their users, there are many challenges associated with text mining these sites.

5 Analysis: Twitter Data

In this section, we present an analysis of a dataset gathered from Twitter, with focus on the analysis of the characteristics of the messages introduced in Section 2. We start by introducing the data collection process.

5.1 Data Collection

Data from Twitter can be gathered by crawling the Twitter website, or via a set of official APIs. Full public information about the messages and users can only be obtained through APIs. Twitter APIs consist of a REST API and a Streaming API. The REST API supports keyword or user ID-based querying, but is subject to rate limiting, currently set to 350 requests per hour for authenticated users. The Streaming API provides access to a random sample of 5% of public status updates from its users. The results reported in this chapter are based on two different data sets we gathered from Twitter.

Data Set 1: This data set was collected in a one moth period starting from the 26th of October 2010 using the Twitter Streaming API. We collected about 200 million messages generated by around 15 million users during this period. The results reported in the remainder of this section and, expect where explicitly mentioned, the case studies are based on this data set.

Data Set 2: This data set consists of messages in the Chinese language, collected over a period of 5 months from April to September 2010, using an in-house crawler. By searching for particular Chinese keywords relating to the World Expo 2010, we were able to constrain messages to the Chinese language, and specifically posts relating to the Expo.

5.2 User Information

Twitter has a very large number of users spread across different countries. In 2010, 100 million new Twitter accounts were created. These users are from different age groups, despite Twitter being targeted at a young demographic. 10 million visitors to the Twitter website in February 2009 were over 35 years old. In the US, 10 percent of users are aged between 55 and 64, almost equalling the number of users aged between 18 and 24 [18]. This shows the diversity of the population contributing content to Twitter.

The location information provided by users shows their geographical diversity. However, this information is specified by the user and can be both unreliable and inconsistent in format/granularity. For example a user in San Francisco may list their location as *USA*, *CA*, *CA USA*, *San Francisco*, *San Francisco CA*, *San Francisco CA US(A)*, etc. The top-20 locations declared by users in the collected dataset are shown in Figure 2.

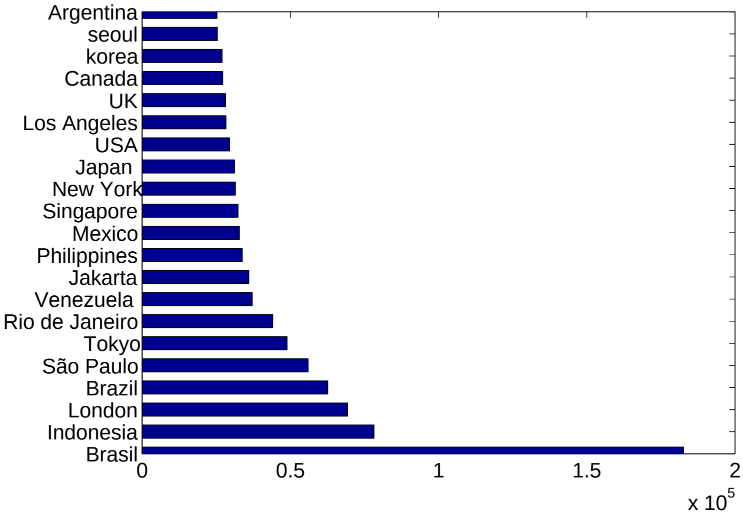


Fig. 2. Top locations specified by Twitter users.

Figure 3 (left) depicts the distribution of spoken language(s) using user-declared information. Our own analysis of Twitter messages based on automatic language identification [1] over a random subset of 600,000 messages points to a higher diversity of languages used on Twitter (see Figure 3 (right)). This language diversity shows the global reach of Twitter, and emphasises the need for language-specific methods for processing Twitter data.

5.3 Connection Information

An important characteristic of Twitter users is their interconnectivity or *social links.* In Twitter, a user can create a social link by *following* another user or adding a user as a *friend.* Figure 4 shows the distribution of the number of followers and friends in our data set. The data contains over 2 billion social links.

Social interactions are more complicated to infer from messages and it requires analysis of the message content. Twitter provides some additional information that

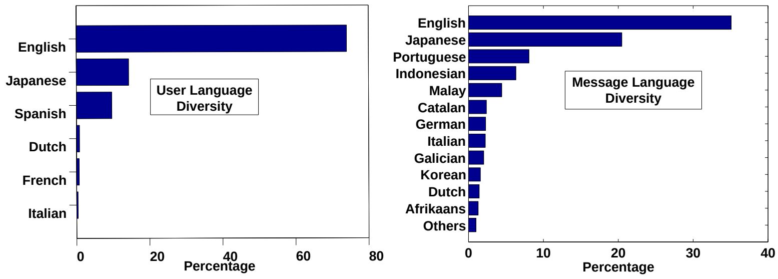


Fig. 3. The language diversity among twitter users specified by the users (left) and obtained by a language identification tool (right).

helps identification of social interactions. These additional fields indicate whether a message is in response to another message or user. 34% of the messages in our data set are responses. Twitter users themselves first introduced this feature informally by prefixing user names with @, but this is now officially supported by Twitter.

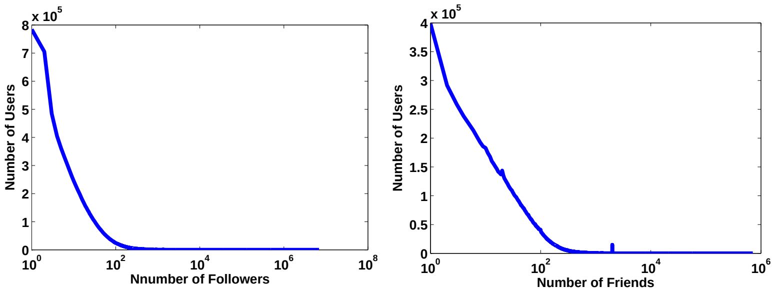


Fig. 4. Distribution of the number of followers (left) and friends (right) in the Twitter data set.

5.4 Message Information

Messages in Twitter, known as *tweets*, are limited to 140 characters. A tweet can be *retweeted* by other users, to share the content with followers. Retweeting is a simple mechanism of message dissemination. Beside the content of the message, Twitter APIs provide additional information about the message. In this section, we briefly review these complementary fields, that can aid data analysis and knowledge extraction.

Complementary Information

The date and time of each message is an attribute that is included with each tweet. This data can be used along with the message content to find popular trends among users and how they evolve over time. Another piece of information optionally provided with each tweet is the geolocation of the user when posting the message. This data is important for many applications, notably event detection and emergency response. Many modern smart phones have built-in GPS, which facilitates geotagging. Therefore, the number of tweets with geolocation information is expected to rise with the increasing number of users accessing Twitter via a smart phone. Figure 5 shows a 10% decrease in the number of users accessing Twitter via the web interface, and an increase in the usage of smart phones and custom-made applications to access Twitter from 2008.

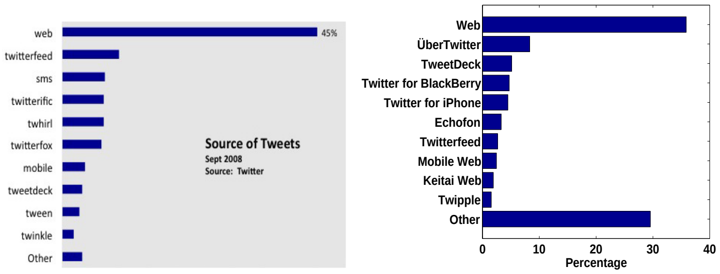


Fig. 5. Client applications used to access Twitter reported by Twitter 2008 (left) and obtained from our data set (right).

Message Content

To conclude this section, we analyze the contents of English messages. Messages contain a lot of information, and are usually the focus of data mining applications. Natural language processing and text mining techniques are widely used to extract useful information from the messages. Challenges faced by these techniques are discussed in Section 4. In Twitter, user-generated features such as specifying the topic of a tweet with hashtags (#) help to categorize the message content. According to [15], around only 2% of tweets in 2008 contained hashtags, while in our collection 12% of tweets contain hashtags. Hashtags can be used to identify long- and short-term trends. Short-term trends quickly reach a peak number of messages and then drop off. Long-term trends, on the other hand, take longer to reach their peak and then to dissipate. When a subject becomes popular among users, hashtags related to that subject become frequent in the data. Figure 6 (left) shows top 15 popular hashtags in the data set. If we want to see short-term trends, we need to look at frequent hashtags in a smaller time window. Figure 6 (right) shows popular hashtags on the 27th of October 2010. In comparison to the most popular hashtags overall, we can see hashtags related to more short-term events such as those corresponding to a natural disaster in Indonesia and election in Brazil. Therefore, in order to capture short-term trends, smaller time windows should be considered for data analysis.

Though hashtags are an important feature in data mining applications, they are not usually sufficient for trend analysis. NLP techniques and keyword analysis

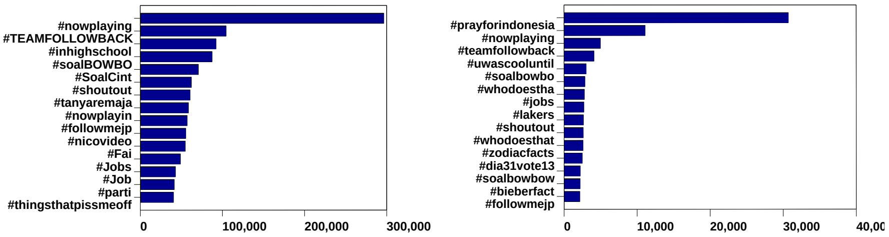


Fig. 6. Popular hashtags and their corresponding frequency in the whole data set (left) and in a one day period (right).

are required to complement this analysis. In terms of language-specific information, the part-of-speech distribution of 3.6 million words[3] is listed in Figure 7. The Penn part-of-speech tagset [22] used by the Stanford parser treats different morphological variations of the same word class as different types, e.g. verbs are separated by inflectional type.

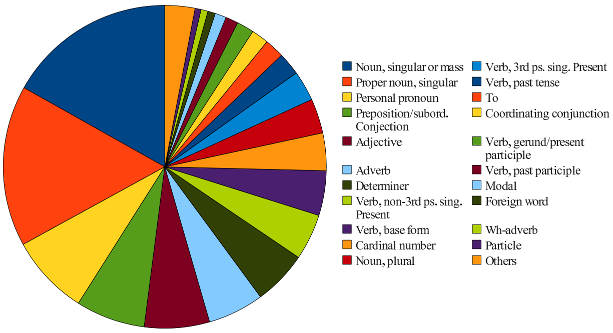


Fig. 7. Part-of-Speech distribution in our sample of English tweets

Unsurprisingly, nouns and proper nouns occupy the top-2 positions in our data, followed by personal pronouns, prepositions and adjectives.

In addition, the top-30 most frequent nouns, verbs and adjectives are listed in Figures 8, 9 and 10, respectively. These figures indicate the language preference of Twitter users, and reflect the trends of public attention at the time of data collection. Especially in Figure 10, sentiment-bearing adjectives like *good, happy, great, bad, nice funny, amazing* and *awesome* are commonplace.

[3] The data is based on the output of the Stanford parser over a 0.3 million sample of English tweets, as identified using automatic language identification over our primary data set.

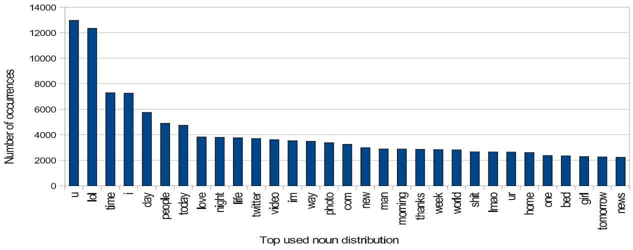


Fig. 8. Top 30 nouns in our sample of English tweets

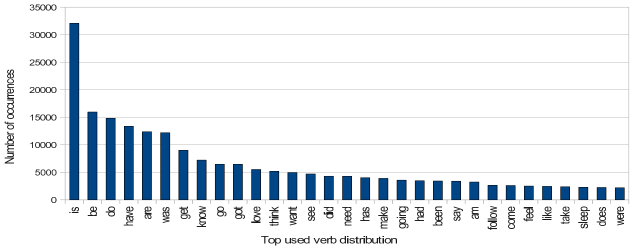


Fig. 9. Top 30 verbs in our sample of English tweets

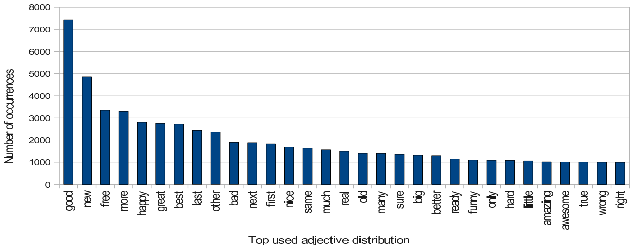


Fig. 10. Top 30 adjectives in our sample of English tweets

6 Case Studies

In this section, we describe three case studies, each from one of the opportunity areas identified in Section 4. Although the data sets used in these case studies were from Twitter, the same techniques can be used in data from other micro-blogs that have a structure similar to Twitter.

6.1 Case Study 1: Event Identification

Background

In this case study, we use an Epidemic Model to characterize the intensity of information flow between micro-blog users, and reveal the sequence of events behind the information flows using frequency-domain deconvolution techniques. The methods shown in this case study are useful in identifying new events by mining micro-blogs.

The Epidemic Model

Epidemic models [25] characterize the way epidemics propagate in communities, which may resemble information flow in Cyberspace. Different epidemic models have been proposed, including the Susceptible-Infected-Recovered (*SIR*) model, Susceptible-Infected-Recovered-Susceptible (*SIRS* model and the Susceptible-Infected-Susceptible (*SIS*) model. We used the SIR model to analyse patterns of interest in events. In the SIR model, the population is considered to be in one of three states:

- Susceptible – when an individual is yet to be infected, but is exposed to the risk of being infected.
- Infected – when an individual has been infected, and is a source of infection.
- Recovered – when an individual has been infected and recovered. A recovered individual is considered to be immune to reinfection.

Based on [25], equations for predicting the number of the people in the population in the different states at a given time point are:

$$\frac{dS(t)}{dt} = -\beta S(t)I(t), \tag{1}$$

$$\frac{dI(t)}{dt} = \beta S(t)I(t) - \lambda I(t), \tag{2}$$

$$\frac{dR(t)}{dt} = \lambda I(t), \tag{3}$$

where the coefficient β denotes the expected number of people an infected individual is in contact with, and λ denotes the rate of infected individuals recovering in a given period; $S(t)$, $I(t)$ and $R(t)$, respectively, denote the number of people that are susceptible, infected and recovered at time point t. Assuming a constant size for the population S_0, and that a given individual must be in only one of the three states at a given point in time, $S_0 \equiv S(t) + I(t) + R(t)$. The basis of $S(t)$, $I(t)$ and $R(t)$ come from [9]; different equations may be used in real world applications to approximate the number of individuals at different states at discrete time points.

If the epidemic is seen as an impulse to the social system, $I(t)$ can be seen as a reaction function to the impulse. Figure 11 shows an example of the variation of $I(t)$ over time for the case of $S_0 = 10000$, β=0.01, λ=0.13.

Information propagation in micro-blogs resembles the course of an epidemic. A new event is perceived (experienced directly or learned about from external sources) by a user, who is the first infected individual. The messages reporting the event are contagious to all users who were not aware of that event. After a period of cooling, a user may lose interest in the event, i.e. recover from the contagion. After losing interest, subsequent exposure to the same event will not raise the interest level of the user again, i.e. reinfection doesn't occur. Based on these observations, we can apply epidemic models to study the flow of information.

We assume that each user who is interested in an event posts messages about the event at a constant frequency, so that the number of messages that mention an event in a period is proportional to the number of the interested users (i.e. the "infected" individuals in the epidemic model). We name the time-dependent function of message numbers $m(t) = c \times I(t)$, where c denotes the frequency of messages. Figure 12 illustrates the relationship between the weekly numbers of messages containing the two keywords *earthquake* (left) and and *expo* (right) in our primary dataset, and

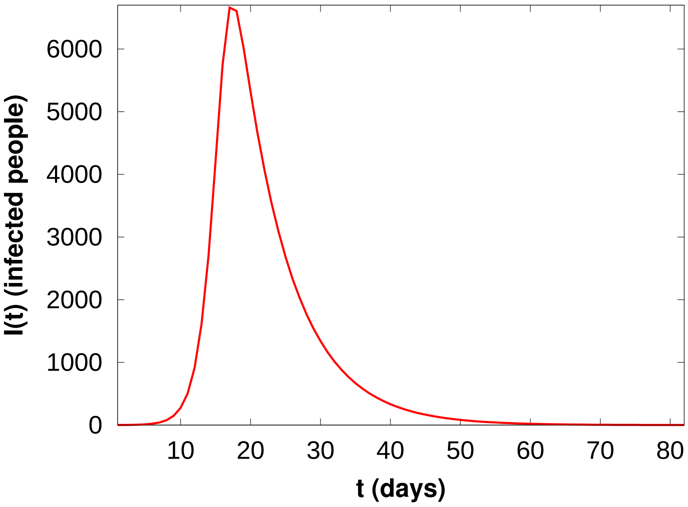
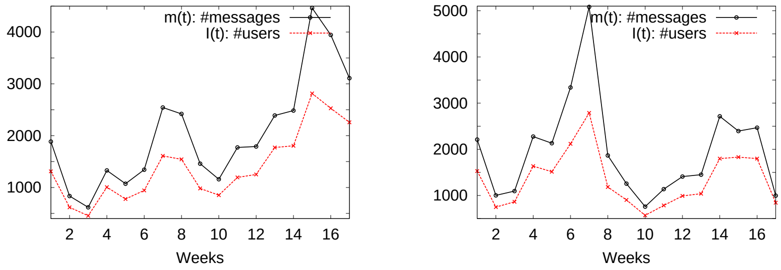


Fig. 11. $I(t)$ for an ideal epidemic

the corresponding numbers of individual users who posted these messages per week, namely $m(t)$ and $I(t)$ respectively. The very high correlation-coefficients between $m(t)$ and $I(t)$ give support to our assumption — c in both cases is equal to 1.46.



(a) *Earthquake*

(b) *Expo*

Fig. 12. The relationships between numbers of messages about an event and the number of individual users who posted these messages

We show in Figure 13 two examples of significant events which generate reactions on a micro-blog. Since each message is published at a time point, we need to define a time granularity for grouping and counting the messages, so that the discrete messages are converted into a time-dependent density function. In our case, the granularity is set to a day. The two curves in the figure show the density of messages mentioning the keywords *A380* and *earthquake*, respectively, in the primary dataset. As shown, the frequency of messages about *A380* was clearly raised by the occurrence of a mid-air incident involving a Qantas A380 on the 3rd of November. There are two peaks, differing in significance, on the curve for *earthquake*: the first peak, the more significant one, can be attributed to an earthquake, with a magnitude of 7.2, in New Zealand on the 25th of October, while the second peak can be attributed to a minor earthquake, with a magnitude of 6.1, in Indonesia on the 3rd of November. One may notice that the three peaks and the ideal epidemic model in Figure 11 have similar shapes. The two earthquakes differ in magnitude, and the different demographic backgrounds of the two countries further result in differences in the number of people who are potentially concerned. As a result, we have two peaks with different S_0 and β for the two earthquakes.

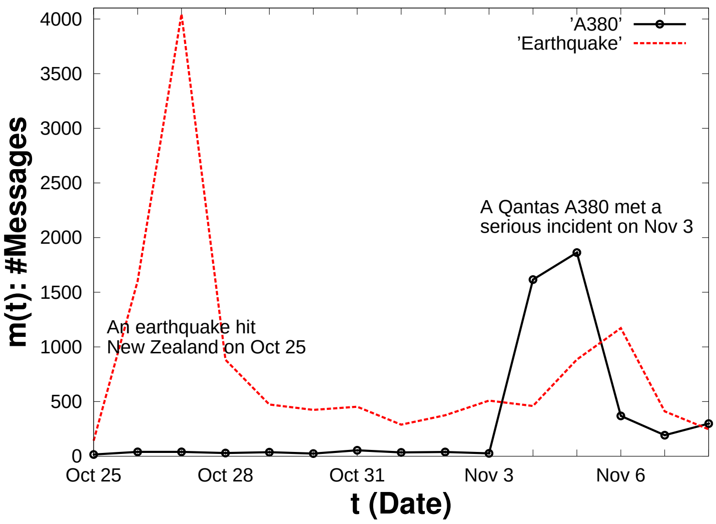


Fig. 13. The density of messages containing the keywords *A380* and *earthquake*

Deconvolution

The epidemic model characters the process of a single event arousing reactions from users. However, there is usually a very large number of events happening in the real world in any short period, each of which can be seen as an impulse to the system; the output of the system in that period (e.g. the number of messages mentioning these events at each time point) is thereby the accumulation of influence from all these events. Hence, each peak of message numbers may not be exclusively mapped to an event, and the outcome from minor events may be overshadowed by major events. To recover the original events from the mixture of these footprints, we used the method of deconvolution to convert the time-series data into the frequency domain, wherein we recovered the strengths and the time points of the potential events.

The reader will recall that $m(t)$ is a time-dependent function of message numbers. This function can be seen as the convolution of pulses (the event) and reaction functions (the epidemic model). The significance of each event is the energy of the pulse, and the impressiveness/attentiveness of the event determines the parameters of the epidemic model. Therefore:

$$m(t) = w(t) * e(t) + noise(t), \tag{4}$$

where $w(t)$ is the event function to be recovered, and $e(t)$ is the reaction function to each impulse; it equals to I when there is only one impulse to the system. Hereafter, we will use $*$ to denote the operation of convolution.

Considering the noise-free case, we used the Fourier Transform to restore the original $w(t)$ from $s(t)$ by calculating:

$$W(\omega) = \frac{M(\omega)}{E(\omega)}, \tag{5}$$

where $W(\omega)$, $M(\omega)$ and $E(\omega)$ are the frequency domain functions corresponding to the three time-domain functions. From $W(\omega)$, we calculate the approximation of $w(t)$ by the Inverse Fourier Transform:

$$\hat{w}(t) = FT^{-1}\{W(\omega)\} \tag{6}$$

Note that there are three parameters of $e(t)$: S_0, β and λ, which are unknown. However, a large number of optimistic algorithms, e.g. the Expectation-Maximization (EM) algorithm, can be used to determine $e(t)$ and $w(t)$, as demonstrated in [13] and [14] if additional information about the events is available. We used a simple approach, by trying a number of parameters and selecting the set of parameters that trims $w(t)$ most effectively.

Natural Language Processing

As mentioned above, micro-blogs provide no or very limited meta-data about each message. The only available information is the date of publication, and for some newer applications, the location of the publisher and a handful of tags. Therefore, NLP is necessary to extract more useful information from each message, including the topic of any events, the names of people who are involved in the event, the location where the event took place (which may differ from the location of the message author), and the date/time of the event (which may also differ from the time when the message is published).

NLP is particularly critical for processing micro-blog messages in languages such as Chinese and Japanese, which are non-segmenting languages which do not represent word breaks in their orthography. In our case, we use NLP to split sentences into words to identify the names of locations that are mentioned in the messages. In this case study, we used the location name as a keyword filter to efficiently removing irrelevant messages.

Data analysis

The events were identified by a two-step routine. First, we calculated the mean message frequency for each location as a baseline, so that the locations that temporarily received markedly high attention could be spotted. Second, those locations spotted

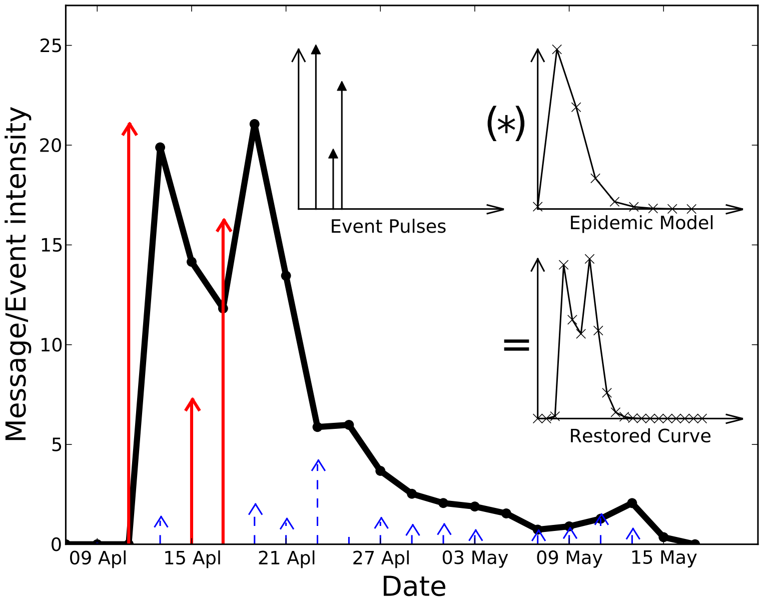


Fig. 14. Using deconvolution to rationalize the events behind the messages that mentioned *Yushu* in the given period. The symbol $*$ denotes convolution. Note that the small chart of the restored curve is the result of convolution of the identified events and the ideal epidemic model, which is very similar to the trend of the observed messages.

in the first step were analyzed using the epidemic model to evaluate whether events happened in these locations.

There was a number of locations mentioned in our primary data set that emerged as being unusual, amongst which the most significant location which attracted a great deal of momentary attention was Yushu. We then analyzed the messages about that location using deconvolution, and show our result in Figure 14. The main chart in the figure illustrates the relationship between the number of messages mentioning *Yushu* each day, and the potential background events, based on deconvolution. The three smaller charts show how the outcome function, $m(t)$, is accumulated by the reactions to the three events, each expressed as a peak based on the epidemic model. Note that the "restored curve" is the result of re-calculating the convolution of $w(t)$ with $e(t)$, but not the observation results. The restored curve has a very similar shape to the observed curve, lending support to the validity of our methodology.

We mapped the resulting events to real world incidents by manually reviewing the messages. The significant event behind these messages is an earthquake in Yushu on the 14th of April, resulting in thousands of fatalities. There are three potential events identified. The first event was the earthquake itself which generated an initial flood of messages from concerned people; the second event was reporting of the high death toll; and the third event was the official nationwide mourning which was announced on the 20th of April and conducted on the 21st.[4] Note that the date labels are based on Coordinated Universal Time.

Challenges

Reliably identifying events based on messages from micro-blogs has several associated challenges, which are identified below.

- Some users simply re-post news stories via Twitter instead of reporting their own experience. These messages may reflect the social reaction raised by the media, but they have little, if any, effect on the identification of new events before they are reported in traditional ways.
- Some users do not care about authenticity of information; they just re-post messages that they found interesting. Thereby, rumours are propagated along with true information; rumors may be propagated even faster because of the characteristic implications of scandal.
- Information propagation of some events does not perfectly match the curve derived from the epidemic model: sometime there are multiple peaks for a single event, sometime there is a long tail after the event happened, and sometimes the curve rapidly dissipates. Also, noise may come from outside sources, such as biased media coverage, variable user interest levels, weekdays and weekends, and even the different time-zones users live in.
- Some of the assumptions in our SIR model are too simplistic, which could affect the reliability of our technique. For example, a recovered user may be reinfected by the bombardment of media reports, and users may have very different cooling down periods. By using a more sophisticated model, we may discover more latent factors behind the messages.

6.2 Case Study 2: Social Cluster Identification

Background

This second case study presents our approach for identifying social clusters in micro-blog users. A cluster is defined in graph theory as a set of vertexes in a graph between which there is a complete sub-graph. Our definition of user clusters resembles the definition in graph theory, by characterizing a user cluster to be a set of users who interact with each other intensively, such that information can rapidly propagate between the users. However, we do not expect the complete sub-graph between all members of a cluster because information can still quickly propagate between a given pairing of users which is not directly linked but shares a large number of friends; instead, we used the Markov chain model and random walks to measure the distance between each pair of users.

The forms of social connections and social links differ in the ability of reflecting the relationships between users. In [10], the authors analyze the forms of interactions

[4] There is a minor impulse on the 25th of April. This impulse cannot be explained by any single event, but by the accumulation of a large number of minor events in the aftermath of the earthquake

between the groups of users, and reveal that links made by mentioning (i.e., retweeting and replying), rather than by following, have more significant correlation to the user groups; hence, we followed only the links by mentioning in this case study.

Markov Chain Model

The Markov chain model is widely used for measuring the distances between vertexes in a graph. The probability of a user being reached in τ steps of propagation is defined by:

$$U_\tau = \begin{bmatrix} p_\tau(u_1, u_r|x) \\ p_\tau(u_2, u_r|x) \\ \vdots \\ p_\tau(u_m, u_r|x) \end{bmatrix}, \tag{7}$$

where $p_\tau(u_k, u_r|x)$ is the probability that the k-th user is reached by message x generated by user r in τ steps. This probability vector is derived using an iterative equation:

$$U_\tau = \mathbf{A}U_{\tau-1}. \tag{8}$$

In this formula, matrix $\mathbf{A}$ is the adjacency matrix defining the probability of a step taken between any two directly connected users. The initial vector, U_0, is a zero vector, but with the i-th element being 1, where i denotes the starting node in a random walk. Adjacency matrix $\mathbf{A}$ is derived by the normalized weight of out-degrees of each user, which is defined by the number of messages from this user and mentions of each other user.

We note that the damping factor that is defined in the PageRank algorithm may be important to characterise the means of the random walk, namely a walk terminates at any node with a given probability. However, we did not introduce this factor in our case.

Distance Measurement

The Markov chain model provides a measurement of the probability with which a walk stops at each node after a given number of steps. Given enough time, the probability of a walk reaching any node will converge to one if there is a path between the target node and the original node in the graph, and the probability of the walk stopping at a node converges to a value that is independent of where the walk started. On the other hand, we need an approach for distance measurement, which is only determined by the structure of the network, and is sensitive to the starting node.

Considering the epidemic model mentioned in the first case study, an individual is infected when the epidemic reaches this individual for the first time. The time cost for propagating the epidemic can be measured by the expected number of steps taken in a walk in the social network, from a given node to reach a target node for the first time. This expectation is only determined by the structure of the network, and is more suitable for measuring the ease of information propagating between two nodes; hence, we used these expectations as the distance measurement for clustering users. An algorithm, proposed in [26, 28], was used to calculate these expectations.

By knowing the distance between each pair of users, one may use an existing clustering algorithm to group the users that are close to each other; we used the k-means algorithm in our case study, which requires the manual specification of the value k. We tried a number of different values of k, and selected the one that results in no predominantly large clusters. Note that, in a practical application, the process of finding k can be done using an optimistic algorithm.

For evaluating the validity of our method, we assume that, being closer to each other on the path of information propagation, users in the same cluster have similar backgrounds. We manually evaluate the clusters identified in our data set, and show the similarity between the profiles of their members.

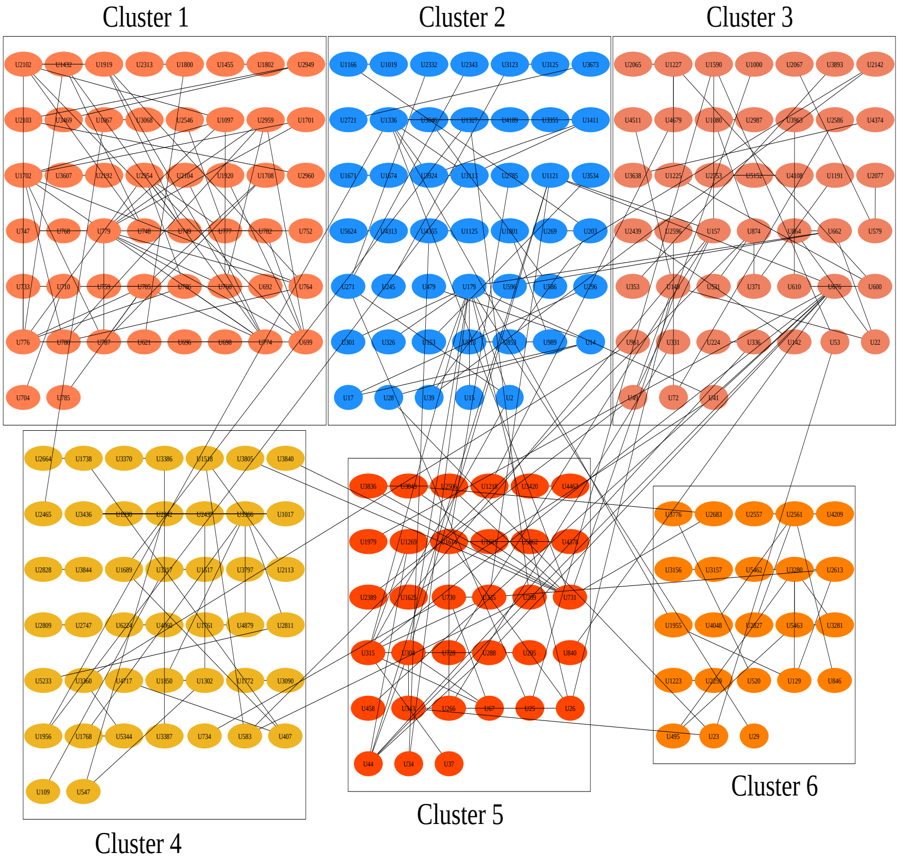


Fig. 15. The clusters identified in our Twitter data set. From the large numbers users in each cluster, we selected only those who have the largest numbers of connections to others in drawing the graph. Cluster are rendered by bounding rectangles. Clusters that have very few members are not rendered, for readability purposes. The user identities are not shown in the graph.

Revealed clusters

We ran our algorithm on the primary data set, and identified six major user clusters that have many members. Figure 15 shows members of these clusters, as well as inner- and inter-connections between these members. The number of inter-cluster connections in the figure are far fewer than those connecting users in the same cluster; which gives the basic support to the validity of the distance measurement and clustering algorithms.

Each link shown in the figure denotes at least 14 re-tweets or replying messages between the two users. The structure of the connections with each cluster is rather like the conjunction of star structures: some users have significantly more connections to other users, and there are relatively fewer links between inactive users. This confirms to Zipf's law, which is often used to analyze the popularity of objects in social systems, such as people in social networks and websites on the Internet. It also worth noting that we referenced not only the links shown in the figure in clustering, but also links that have low weights, i.e. instances of less than 14 interactions between a given pairing of two users. These links are omitted from the figure for readability purposes. From Table 1, we can see that users in a given cluster may share the same social background, the same occupation, the same political ideology and even the same geological location. This results in them having similar interests, such that a message from a cluster member has a high probability of being noticed by other users in the same cluster. Please note that the clusters are roughly divided and the description may be not applicable to all members of the cluster.

Mark	Members	Description
Cluster 1	366	Young people living in Taiwan and Hong Kong
Cluster 2	507	People living in mainland China who tend to talk about critical politic issues
Cluster 3	669	Young people living in mainland China, sharing popular things
Cluster 4	369	IT workers in China who tend to talk about technical issues
Cluster 5	667	People in their 30's living in mainland China, sharing deeper and more intellectual subjects
Cluster 6	365	Similar to the third cluster
Sum	2493	

Table 1. Descriptions of the clusters that are identified using the random walk model; the descriptions are summarized by manually reviewing the messages and the self-descriptions of the cluster members.

The exclusiveness of the clusters strongly supports the validity of our algorithm. The figure shows that the second cluster, formed mainly from people living in Taiwan and Hong Kong, is the most exclusive cluster, due to its intense inner links and very few outer links. The imbalanced connections between clusters reflect the extend to which users share same interests. Different economic, cultural and political backgrounds make the second cluster share few interests with other groups. Another exclusive cluster, the first one, is a turnoff for other users because of the detailed technical subjects.

Even though the unique backgrounds of each cluster tend to exclude irrelevant users, it does not mean that all topics in the cluster are unique. We noted that more than half of the topics in every cluster are irrelevant to the background and circumstance of the cluster members; however, people are more likely to share general topics of interest with users in the same cluster.

Challenges

The validity of the methodology introduced in this section may be influenced by factors including:

- Clusters may be temporal in nature. A new event may lead to interactions between a particular set of users, who interact solidly over a short time period; however, as their interest in the event fades, the cluster may dissipate.
- The hard cluster membership defined in our model may be over simplistic. A particular combination of circumstances may result in a user belonging to multiple clusters, between which there is very low similarity. Ignoring such overlaps in membership can result in inappropriate merging of two unrelated clusters.

6.3 Case Study 3: Trend Identification

Background

In this case study, we present a technique for identifying trends in long-term events by observing trends in micro-blog messages mentioning them. Expo 2010 took place in Shanghai from May to November in 2010. In the 184 days, there were around 70 million visitors to the exhibition park, and disclosure of daily and hourly updates on visitor numbers on the official website. This publicly available data served as a good reference for analyzing how micro-blog messages reflect the public perception of events. Figure 6.3 shows the number of visitors and the number of Twitter messages referencing the event. In contrast to the figures shown in the first case study, long-term events do not usually raise a single peak of interest, but a rather smooth and continuous curve of attention. Nevertheless, the curve still reflects the interests of the users, so that it may reveal event trends.

Methodology

We used the following three step process to identify the relationship between the number of messages mentioning Expo 2010 and the daily visitors to the event.

- The interlinks between the users were extracted from the messages for clustering the users into clusters, using the clustering technique presented in Case Study 2.
- From each cluster, we extracted messages that potentially mention experience and/or intention of visiting Expo 2010, and also counted the number of messages from that cluster.
- We calculated the correlation coefficients between the relative proportion on Expo-related messages from each cluster and the official visitor numbers. A higher correlation coefficient denotes the cluster being a better indicator of visitor number prediction.

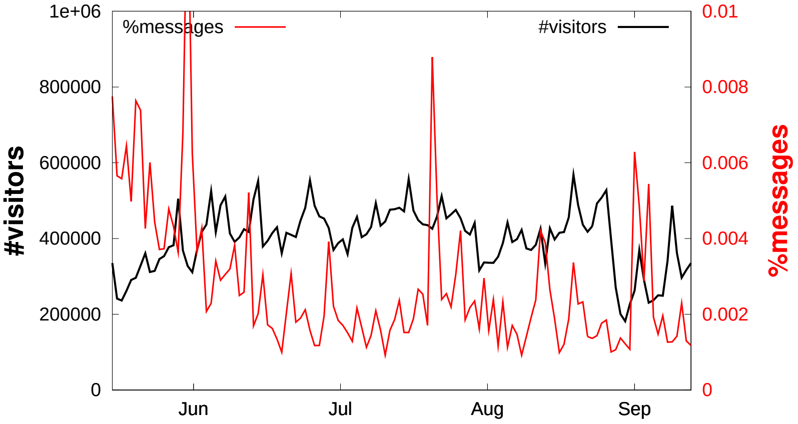


Fig. 16. The number of the visitors to Expo 2010 on each day, versus the proportion of messages mentioning the event. The three abnormal peaks on the message number curve correspond to three short-term events: a stampede incident on the 30th of May, a rumour about injuries in the very congested park in late July, and a critical news report which was published in early September and disclosed unpleasant behaviour of visitors and aroused discussion.

Micro-blog Message Refinement

From the 13.5 million messages in our primary data set, 35 thousand explicitly mentioned *Expo 2010*. We divided the number of messages each day mentioning *Expo* by the total message number for the day, in order to remove the variance resulting from accessibility to Twitter in mainland China.

As mentioned in the first case study, some messages posted on micro-blogs may not reflect the experiences and intentions of the users. We further reviewed messages from Twitter mentioning Expo 2010, and found that users have different motivations for writing messages mentioning an event:

- Some users copy stories from outer websites or re-tweet other user's message about the event.
- Some users post comments on news reports or to others' messages about the event.
- Some users report the intention or experience of other people to the event, who may have no access to Twitter.
- Some users report the intention or experience of themselves with regard to the event.

Only the last two message categories, which report the experience and the intention in visiting Expo 2010, are valuable to us. We distinguished these messages using the following heuristics:

- These messages are more authentic, and as such are re-tweeted by fewer, if any, users.

- These messages contain more subjective and relative temporal expressions like *I*, *we*, *will*, *go*, *today* or *tomorrow.*

Only messages which satisfy these two heuristics (i.e. which weren't retweeted, and contain one of a small set of keywords) were considered for analysis.

Data analysis

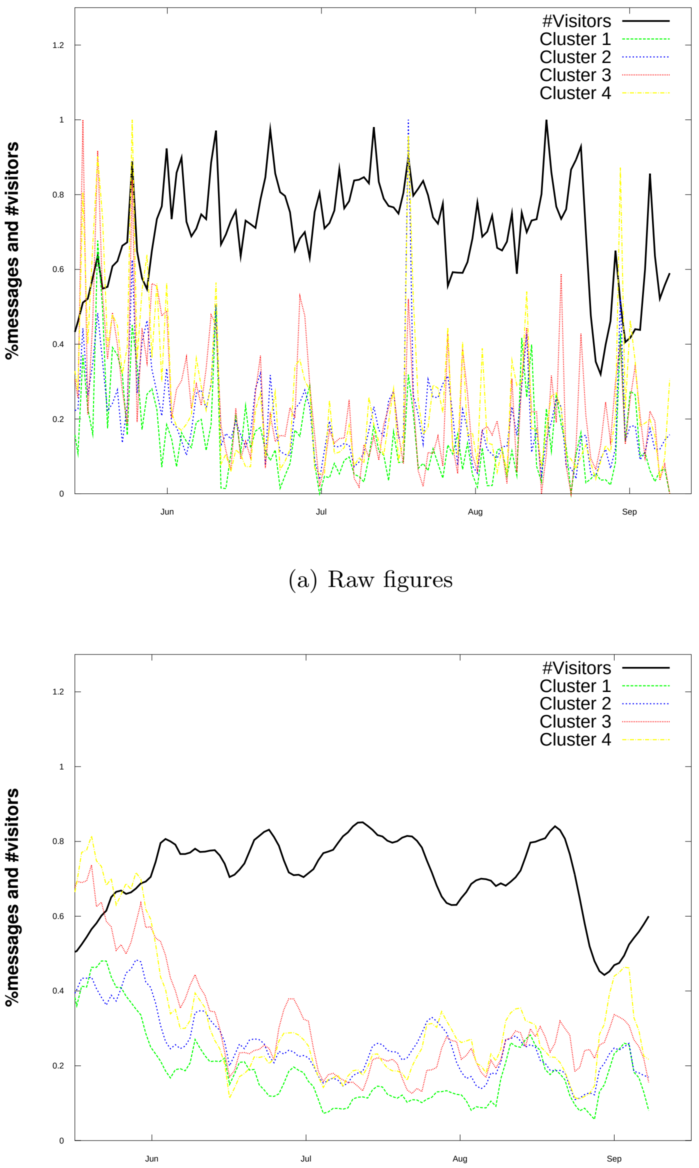
17(a) shows the real visitor numbers and the proportion of messages from each cluster based on our technique. This figure seems very noisy because of the three reasons: (1) high-frequency noise over the weekly period, (2) the great enthusiasm at the very beginning of the exhibition, and (3) the impulses from the three special events described above. We manually removed these abnormal peaks, and plot the result as a seven-day moving average in Figure 17(b).

The correlation between the numbers of messages from clusters and the official numbers of visitors is perceivable, while each cluster has a different lag, either positive or negative, over the trend of the official visitor numbers. We show in Table 2 the correlation coefficient with the official visitor number for the message numbers of each cluster and also the overall Twitter population community. We ignored the first 30 days to remove noise from the great enthusiasm at the beginning of the Expo.

Lag	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total Msgs
-6	0.3656	0.3475	0.0554	0.2775	0.0708
-5	0.4139	0.3719	0.0761	0.2875	-0.0086
-4	0.4319	0.3939	0.0944	0.2946	-0.1010
-3	0.4210	0.4098	0.1138	0.2879	-0.1990
-2	0.3869	0.4216	0.1320	0.2724	-0.2717
-1	0.3366	0.4324	0.1459	0.2468	-0.3298
0	0.2847	0.4449	0.1681	0.2191	-0.3575
1	0.2282	0.4521	0.2007	0.1938	-0.3391
2	0.1817	0.4554	0.2394	0.1722	-0.2856
3	0.1527	0.4534	0.2851	0.1451	-0.1878
4	0.1191	0.4413	0.3102	0.1283	-0.0974
5	0.0968	0.4287	0.3342	0.1179	0.0028
6	0.0688	0.4035	0.3366	0.0995	0.1075
7	0.0208	0.3604	0.2913	0.0620	0.2090

Table 2. The Correlation coefficients between the official visitor numbers and the the message numbers from the four clusters, as well as the total number of messages mentioning the Expo. The Correlation coefficients are calculated over the seven-day moving average values. Note that a negative lag denotes that the message numbers can be used in predicting future visitor numbers, while a positive lag denotes that the message numbers can be used in recovering past visitor numbers.

This table shows that the clusters vary in their ability in predicting or recovering the numbers of visitors. The first cluster is useful as an indicator of the users'



(a) Raw figures

(b) Seven-day moving average after trimming

Fig. 17. The number of messages from the four major user clusters versus the official numbers of visitors. Note that the message numbers are normalized; from the four clusters, the maximum average number of messages is approximately five times larger than the minimal number.

intention in four days later because the maximum Correlation Coefficient is with a four-day lag, whereas the the second and the third clusters may be used to recover the attendance trend in following days. It is also shown that, without user clustering, the total numbers of messages have very little correlation, if any, with the trend of the event.

We further tested our methodology by comparing trends of messages containing other irrelevant keywords to the trend of Expo 2010. The keywords that we selected are *officer* and *house price*, both of which are popular but intuitively irrelevant to Expo2010. The results showed that the maximum positive correlation coefficients of these irrelevant message trends to the numbers of attendants are 0.12 and 0.21 respectively, much lower than the Correlation Coefficients of the message trends that are relevant to Expo 2010; namely, the message trends for irrelevant keywords are largely meaningless for predicting or recovering the trend of attendance to Expo 2010.

Challenges

This case study provides only a preliminary example of trend identification by mining micro-blogs. We point out some issues to be further improved for making our methodology more practical in real mining applications.

- As we pointed out in Case Study 2, user clusters may change from time to time. This may influence the approach to utilizing the cluster-wise information in two ways: (1) the membership of clusters may change with time; and (2) the correlation and the lag between the behaviour of a cluster and the trend of the event may differ for a long-term event.
- The method for clustering users may be further guided by the ability of the resulting clusters to predict event trends. Given that users who have similar reactions to events are more likely to share the same interests, these clusters will be better suited to event trend prediction.
- In this case study, we manually removed noise the data. Nonetheless, as we showed in the first case study, this task can be automatically done using deconvolution and the EM algorithm. Further studies are needed to build the model for analyzing the mixture of long-term trends and short-term events.

7 Conclusion

This chapter has served as an introduction to data mining of micro-blogs, with a particular focus on Twitter. We first described characteristics of micro-blogs, and different data mining tasks that can be performed over them, including both opportunities and challenges. We then presented an analysis of Twitter data, based on data collected over a four week period. Finally, we presented three case studies using Twitter data: event identification using an epidemic model, social cluster identification using a Markov chain model, and trend identification using keyword frequencies within user clusters.

References

1. Baldwin, T., Lui, M.: Language identification: The long and the short of the matter. In: Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010), pp. 229–237. Los Angeles, USA (2010)
2. Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Posters Volume, pp. 36–44. Beijing, China (2010)
3. Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: Linking text sentiment to public opinion time series. In: Proceedings of the Second International AAAI Conference on Weblogs and Social Media, pp. 122–129. Washington DC, USA (2010)
4. Culotta, A.: Towards detecting influenza epidemics by analyzing Twitter messages. In: Proceedings of the KDD 2010 Workshop on Social Media Analytics. Washington DC, USA (2010)
5. de Marneffe, M., MacCartney, B., Manning, C.D.: Generating typed dependency parses from phrase structure parses. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006). Genoa, Italy (2006)
6. Goorha, S., Ungar, L.: Discovery of significant emerging trends. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 57–64. ACM (2010)
7. Han, B., Baldwin, T.: Lexical normalisation of short text messages: Makn sens a #twitter. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011). Portland, USA (to appear)
8. Hannon, J., Bennett, M., Smyth, B.: Recommending Twitter users to follow using content and collaborative filtering approaches. In: Proceedings of the Fourth ACM Conference on Recommender Systems, pp. 199–206. Barcelona, Spain (2010)
9. Hethcote, H., Tudor, D.: Integral equation models for endemic infectious diseases. Journal of Mathematical Biology **9**(1), 37–47 (1980)
10. Huberman, B., Romero, D., Wu, F.: Social networks that matter: Twitter under the microscope. First Monday **14**(1), 8 (2009)
11. Klein, D., Manning, C.D.: Fast exact inference with a factored model for natural language parsing. In: Advances in Neural Information Processing Systems 15 (NIPS 2002), pp. 3–10. Whistler, Canada (2003)
12. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web, pp. 591–600. Raleigh, USA (2010)
13. Lane, R.: Methods for maximum-likelihood deconvolution. JOSA A **13**(10), 1992–1998 (1996)
14. Likas, A., Galatsanos, N.: A variational approach for Bayesian blind image deconvolution. IEEE Transactions on Signal Processing **52**(8), 2222–2233 (2004)
15. 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 (2008). O'Reilly Radar Report

16. Petrović, S., Osborne, M., Lavrenko, V.: Streaming first story detection with application to twitter. In: Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010), pp. 181–189. Los Angeles, USA (2010)
17. Quincey, E., Kostkova, P.: Early warning and outbreak detection using social networking websites: The potential of twitter. In: Electronic Healthcare, vol. 27, pp. 21–24. Springer, Heidelberg, Germany (2010)
18. Reuters-Web: Twitter older than it looks (2009). URL `http://blogs.reuters.com/mediafile/2009/03/30/twitter-older-than-it-looks/`. Reuters MediaFile blog
19. Ritter, A., Cherry, C., Dolan, B.: Unsupervised modeling of Twitter conversations. In: Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010), pp. 172–180. Los Angeles, USA (2010)
20. Ritterman, J., Osborne, M., Klein, E.: Using prediction markets and Twitter to predict a swine flu pandemic. In: Proceedings of the 1st International Workshop on Mining Social Media (2009)
21. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web, pp. 851–860 (2010)
22. Santorini, B.: Part-of-speech tagging guidelines for the Penn Treebank project. Tech. rep., Department of Computer and Information Science, University of Pennsylvania (1990)
23. Sproat, R., Black, A.W., Chen, S., Kumar, S., Ostendorf, M., Richards, C.: Normalization of non-standard words. Computer Speech and Language **15**(3), 287–333 (2001)
24. Twitter: Big goals, big game, big records (2010). `http://blog.twitter.com/2010/06/big-goals-big-game-big-records.html`. Retrieved 4 August 2010.
25. W. O. Kermack, A.G.M.: A contribution to the mathematical theory of epidemics. In: Proceedings of the Royal Society A, vol. 115, pp. 700–721 (1927)
26. Wasow, W.: A note on the inversion of matrices by random walks. Mathematical Tables and Other Aids to Computation **6**(38), 78–81 (1952)
27. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM '10, pp. 261–270 (2010)
28. Yen, L., Vanvyve, D., Wouters, F., Fouss, F., Verleysen, M., Saerens, M.: Clustering using a random walk based distance measure. In: Proceedings of the 13th Symposium on Artificial Neural Networks (ESANN 2005), pp. 317–324 (2005)