A step towards real-time analysis of major disaster events based on tweets

André Dittrich
Karlsruhe Institute of Technology (KIT),
Institute of Photogrammetry and Remote Sensing (IPF),
andre.dittrich@kit.edu

Christian Lucas
Karlsruhe Institute of Technology (KIT),
Institute of Photogrammetry and Remote Sensing (IPF),
christian.lucas@kit.edu

ABSTRACT

The most popular micro blogging platform Twitter has been the topic of a variety of research papers related to disaster and crisis management. As an essential first step and basis for a real-time methodology to exploit Twitter for event detection, localization and ultimately semantic content analysis, a functional model to describe the amount of tweets during a day has been developed. It was derived from a corpus of messages in an exemplary area of investigation. To satisfy the different daily behavior on particular days, two types of days are distinguished in this paper. Moreover, keyword-adjusted data is used to point out the potential of semantic tweet analysis in following steps. The consideration of spatial event descriptions in relevant tweets could significantly improve and accelerate the perception of a disaster. The results from the conducted tests demonstrate the capability of the functional model to detect events with significant social impact in Twitter data.

Keywords

event detection, disaster, social sensor, Twitter

INTRODUCTION

Micro blogging is a form of communication that people use to describe their current status in short posts. The most popular micro blogging tool, Twitter, was launched in October 2006 and has reached the mark of half a billion accounts in June 2012[1]. Twitter posts, also known as tweets, are limited to 140 characters. Approximately 0.5 billion tweets are currently sent per day[2] via smartphones or come from web-based services and applications. Twitters architecture enables real-time propagation of information to a large group of users. In order to use the potential of platforms like Twitter, it is important to understand their usage and the community. Chu, Gianvecchio, Wang, Jajodia (2010) have analyzed the differences in tweeting behavior, content and account settings between diverse categories of users. Java, Song, Finin, Tseng (2007) demonstrate the main types of user intentions to be: daily chatter, conversations, sharing information and reporting news. Above that, events such as disasters, also have an influence on the number of tweets. Consequently, 20 million tweets were sent about the impact of winter storm Sandy and its aftermath. During the first presidential debate 2012 the number of tweets per second rose to over 8000. These facts point out that Twitter offers an ideal environment for the dissemination of breaking-news directly from the news' source. Jackoway, Samet, Sankaranarayanan (2011) show that the identification of live news events is possible by using Twitter and much faster than "conventional news aggregators". Sakaki, Okazaki, Matsuo (2010) present an approach for the detection of earthquakes interpreting tweets as sensor values. Their procedure consists of a semantic and probabilistic temporal analysis and finally a spatial analysis based on a particle filter. The weak point of the approach is the scaling with respect to the number of tweets, which continues to increase quickly. The approach, presented in this paper, focuses on a grid-based worldwide detection of major disaster events. The methodology builds on scale and language independent parameters.

TWEET CORPUS

Twitter provides access to the *Firehose*, the real-time stream of all tweets being sent, through its Streaming API (Application Programming Interface). The single posts are received as JSON documents (JavaScript Object Notation) consisting of compulsory and optional fields containing both textual and numerical information. The relevant fields for the current research hold the message itself, the time the tweet was posted and the position from

1 http://www.statisticbrain.com/Twitter-statistics/, retrieved 01/12/2013

2 http://news.cnet.com/8301-1023_3-57541566-93/report-twitter-hits-half-a-billion-tweets-a-day/, retrieved 01/14/2013

where it was posted. Developers, however, are only able to receive a limited amount of approximately 1% of the *Firehose* for free, according to Twitter employees on the developer homepage. Tests showed that a percentage of about 2% of all tweets contain the necessary location information. In another 90% of these tweets, the location is provided as a geographical position, i.e. longitude and latitude in the WGS84. In this case the coordinates either originate from a GNSS sensor or other positioning methods, e.g. Wi-Fi positioning, of mainly mobile platforms. Thus, they can be located with an approximate accuracy between 10 and 100 meters. The remaining 10% of the georeferenced tweets only hold a geographical place with its surrounding polygon or bounding box as location information and are not considered in this paper. The Streaming API allows filtering the tweets either according to several keywords or a geographical bounding box, but not simultaneously. The spatial filter does not provide the expected results though, i.e. depending on the spatial situation, between 1 and 25% of the tweets are lying outside of the defined bounding box. Hence, for a correct analysis, the streaming outcome needs to be further processed.

MongoDB, a document-orientated database system that uses the mentioned JSON format just as Twitter, is employed as storage technology. Consequently, no complex database scheme had to be. Moreover, MongoDB offers two-dimensional spatial indexing for querying georeferenced data and enables regular expressions for keyword searches. The corpus that is analyzed in this paper, consists of four 24-hour records from the 12/29/2012, 01/05/2013, 01/07/2013 and 01/09/2013 starting around 1 pm Eastern Standard Time (EST)[3] respectively. Another 48 hour record spans over 0 pm 12/31/2012 to 0 pm 01/02/2013 containing New Year's Eve as special event. Each of the 24-hour records contains around 2.3 million posts. The absolute number of daily tweets in a specific area strongly correlates with the population density, the technical standards (e.g. smartphone density) and with the popularity of other Twitter-like platforms[4]. Taking these considerations into account, the data is initially limited to a particular area of investigation for this paper. The applied bounding box approximately comprises the east coast of the USA and is defined via its lower-left corner (longitude: -86°, latitude: 0°) and the upper-right corner (longitude: -67°, latitude: 53°). With the applied longitudes following approximately the borders of the EST zone, an undistorted daily routine of the number of tweets could be sufficiently guaranteed. The next steps will expand these boundaries to cover the whole world, trying to build up a global spatial grid of characteristic daily tweet flow. The resolution of a grid-cell will be variable depending on the specific daily amount of tweets. Additional data consisting of two time periods from 3:30 am 10/29/2012 until 2 pm 10/30/2012 and from 6:30 10/31/2012 until 0 pm 11/03/2012, came from keyword requests without spatial restrictions. This particular dataset covers all tweets containing at least one of the following single or compound disaster relevant keywords:

shelter | winterstorm | weather AND sandy | subway AND flood | sandy AND victims | snowfall | power outages

ANALYSIS

The recorded corpus exhibits several interesting aspects that will be analyzed in the following section. In general, the datasets follow a similar, typical characteristic of the amount of tweets during a day (e.g.

Figure **1**).

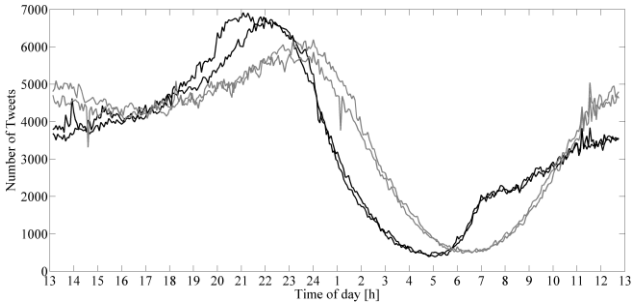

Figure 1. Typical temporal distribution of tweets during a day; gray: 12/29/2012 – 12/30/2012 and 01/05/2013 – 01/06/2013; black: 01/07/2013 – 01/08/2013 and 01/09/2013 – 01/10/2013, all datasets starting at 1 pm respectively

- a peak between 9 pm and midnight (over 6000 tweets in 4 minutes) followed by a rapid decrease
- a minimum between 4 am and 7 am (below 1000 tweets in 4 minutes) followed by a gradual rise

3 All following date and time information will be in terms of Eastern Standard Time.

4 e.g. the Chinese Twitter counterpart *Sina Weibo* with over 400 million users

In the course of a more detailed examination however, the data suggests further differentiation between special types of days, as also shown in (Java et al., 2007). Regarding the current data, the most important aspect influencing the daily amount of tweets, is the percentage of Twitter users being bound to work the next day. Hence, the distinction between at least two main types of days is considered advantageous for further research.

Type 1: days during the week, where the majority of users has to work or go to school the next day

Type 2: days on the weekend or before a public holiday, where the majority of users has the next day off

The exact number of subtypes and their informative significance has not been defined yet due to the limited quantity and variety of recorded datasets. First tests at least indicate more subtypes though, e.g. days where a considerable amount of users typically is on vacation apart from public holidays.

Functional model

To provide the functional analysis with a solid basis, the time step with the highest stability concerning the mean deviation to the following time period during the course of a day is initially identified. The development of the mean deviation as a function of the used time step is visualized in Figure 2. The minimum of the regression resides between 2 and 4 minutes. Representing the best compromise between a sufficient time increment and stability, 4 minutes were selected as a fixed value in the functional model for the given area of investigation. Thus, it was possible to calculate a robust histogram to approximate the temporal distribution of the number of tweets. Besides, following steps will involve monitoring Twitter activities aiming to apply this time step as an interval for real-time analyses and event detection.

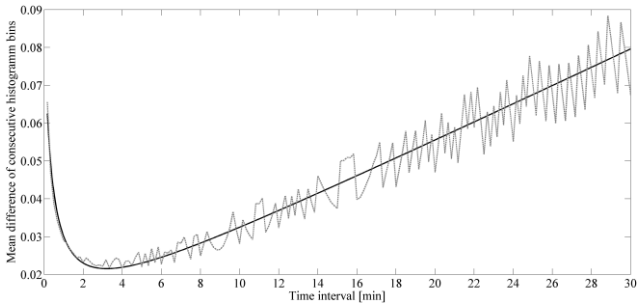

Figure 2. Determination of the most stable time step

The calculated histogram is used as input in a non-linear regression based on a least-squares adjustment to estimate a parameterized functional model. The tests conducted with different functional models clearly indicate that the temporal distribution of tweets during a day is best described by the Fourier series. The basic formula is defined as follows:

$$f(x) = a_0 + \sum_{i=1}^{n} \left(a_i \cos(nwx) + b_i \sin(nwx) \right)$$

Overfitting is avoided by applying a simple constraint to the goodness of the fit, given as the root mean square error (RMSE). The mean deviation of the input datasets to the functional model must not rise above the double RMSE respectively, corresponding to a confidence bound of nearly 95%. Thus, the value of $\boldsymbol{n}$ can be set to 4 resulting in 10 parameters describing one model type (see Table 1). Therein the parameter $\boldsymbol{a_0}$ holds the intercept of the model and represents the average amount of tweets in a 4 minute time step during one day.

	a_0	a_1	b_1	a_2	b_2	a_3	b_3	a_4	b_4	w	RMSE
Type 1	**3463,6**	**928,1**	**-2062,3**	**998,1**	**-704,2**	**353,2**	**-205,5**	**30,2**	**-78,2**	**0,27**	**151,1**
Type 2	3602,9	781,1	-1887,9	1251,1	239,5	-3,7	0,7	178,2	92,5	0,26	92,8

Table 1. Parameters and RMSE of the Fourier series from days of type 1 and 2

Figure 3 shows the models of type 1 (black) and type 2 (gray) and reveals their essential differences as being mainly a translation in time. On a day of type 1 the peak lies near 10 pm compared to the model of type 2 reaching its maximum around 11:30 pm. The minimums are even displaced by nearly 2.5 hours (type 1: 4 pm; type 2: 6:30 pm). However, while the peak in the model of type 1 exceeds its analog of type 2 by more than 800 tweets, the two minimums are almost the same concerning the number of tweets (~500). The only part of the models where they match numerically, temporally and concerning their slope lies between 4 and 6 pm. For subsequent work, this period of time could be an improved boundary to a 24-hour routine of tweets.

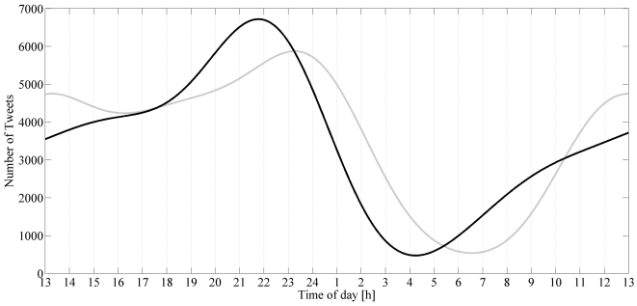

Figure 3. Comparison of functional models of days of type 1 (black) and type 2 (gray)

All of the described effects highlight the typical daily routine of Twitter users and people in general. On weekends people usually stay up longer and get up later the next morning. Additionally, the broader minimum of type 2 denotes the commonly extended time people stay in bed on Sundays and holidays.

EVENT BASED ANALYSIS

The preceding section has introduced an approach for functional modeling of a typical daily routine of the amount of localized tweets in a specific region. As a result, the derived model allows identifying events by their influence on the amount of tweets. In case of an event, this leads to a deviation between the observed values and the expected values given by the particular model. A first indicator for a possible event is a deviation which is higher than the double RMSE of the respective model. This indicator matches approximately the two sigma border, ensuring its statistical significance. Because the estimators are not reliably unbiased, a hypothesis test is also required. An example for such a significant peak is drawn in Figure 4. This histogram shows the amount of tweets in 4 minute time steps in the area of investigation. The period covers approximately a daily routine of 24 hours between 12/31/2012 1 pm and 01/01/2013 1 pm. Accordingly, the event at midnight is New Year's Eve. This events characteristic is similar to that of a disaster event with punctiform extent. Therefore, it is suitable to demonstrate the approach. The black line in Figure 4 marks a day of type 2 as derived in the previous sections. Type 2 was chosen because New Year's Day is a public holiday. Between 7 pm and 11:30 pm, the histogram shows a decline of tweets preceding the event compared to the model. It is assumed that most of the users are part of a New Year's Eve party at this time and therefore tweet less.

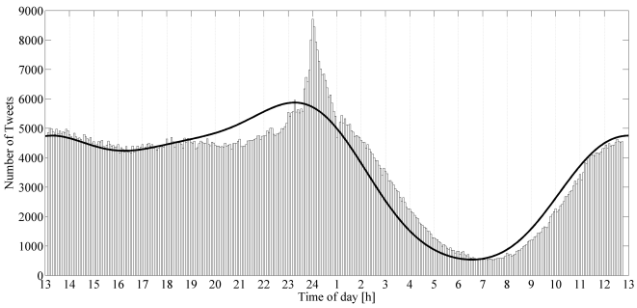

Figure 4: Impact of New Year's Eve on the daily routine of tweets; bars: histogram with 4 minute time steps from 12/31/2012 1 pm to 01/01/2013 1 pm; black line: functional model of type 2

A real disaster to show the impact of such events on the tweets' content, is represented by winter storm Sandy. Sakaki et al. (2010) showed that a storm causes similar tweeting behavior as unexpected events (e.g. earthquakes).

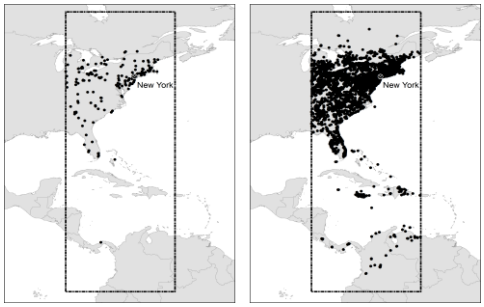

Figure 5. Spatial distribution of tweets containing the disaster relevant keywords; right: 10/29/2012 1 pm to 10/30/2012 1 pm; left: 12/29/2012 1 pm to 12/30/2012 1 pm

The corpus contains the keywords described in section Tweet Corpus. The dataset covers all tweets in the 24 hours from 10/29/2012 1 pm to 10/30/2013 1 pm. This corresponds to the period during which winter storm Sandy hit the east coast of the USA. Figure 5 (right) shows the set of tweets with relevant keywords in the investigated area during winter storm Sandy. Additionally, Figure 5 (left) displays the tweets of a reference day (12/29/2012 to 12/30/2012) limited by the same relevant keywords. Hence, the two maps illustrate the increase of tweets including disaster relevant keywords in case of a disaster. Because a storm is a predicted and more continuous event, an increased degree of event-related tweets is visible in the entire timeline (shown in Figure 6 as gray bars). Thus, the increase was already evident since 10/29/2012 and decreased slowly to 02/11/2012.

The highest peak is given at 9 am on October 30 with 3274 tweets per hour. In contrast, only 10 tweets with relevant keywords were sent in the same time period on the reference day (shown in Figure 6 as black bars).

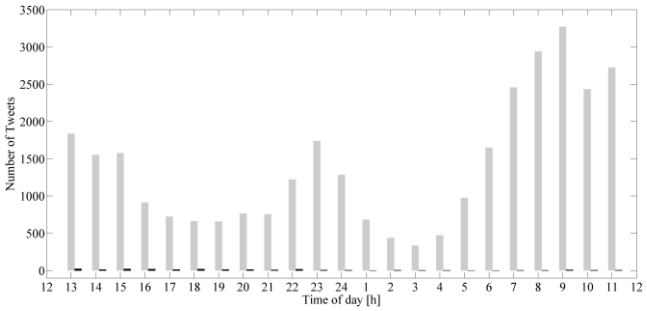

Figure 6. Impact of winter storm Sandy on the daily routine of tweets with respect to particular event-related content; gray bars: 10/29/2012 1 pm to 10/31/2012 1 pm; black bars: 12/29/2012 1 pm to 12/30/2012 1 pm

CONCLUSION AND NEXT STEPS

The paper demonstrated that the flow of daily tweets can be formalized using a functional model based on the Fourier series. Now it is possible to subject ongoing Twitter flow to statistically robust hypothesis testing. Thus, current events producing a significant deviation can be detected. To satisfy the various daily routines, two differing sets of parameters for the model have been introduced. Additionally, the extensive impact of natural disasters on the amount of tweets mentioning event-related keywords was shown on the example of winter storm Sandy.

The next steps towards an automatic, real-time event detection, localization and situation analysis prototype, will involve building up a global, spatial grid with variable cell sizes containing the corresponding functional model and its goodness to enable dynamic hypothesis testing and hence robust event detection.

ACKNOWLEDGMENTS

The work has been funded by the Center for Disaster Management and Risk Reduction Technology (CEDIM). Above that we want to thank Joachim Fohringer (GFZ German Research Centre for Geosciences, Potsdam) for his suggestions as well as for recording the tweet dataset of winter storm Sandy.

REFERENCES

1. Chu, Z., Gianvecchio, S., Wang, H. and Jajodia, S. (2010) Who is tweeting on Twitter: human, bot, or cyborg? , Proceedings of the 26th Annual Computer Security Applications Conference, ACM. Austin, Texas, 21-30.
2. Jackoway, A., Samet, H. and Sankaranarayanan, J. (2011) Identification of live news events using Twitter. in: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, ACM. Chicago, Illinois, 25-32.
3. Java, A., Song, X., Finin, T. and Tseng, B. (2007) Why we twitter: understanding microblogging usage and communities. in: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, ACM. San Jose, California, 56-65.
4. Sakaki, T., Okazaki, M. and Matsuo, Y. (2010) Earthquake shakes Twitter users: real-time event detection by social sensors. in: Proceedings of the 19th international conference on World wide web, ACM. Raleigh, North Carolina, USA, 851-860.