Erschienen in: Proceedings of the 3rd ACM SIGMOD Workshop on Databases and Social Networks : DBSocial 2013; New York, NY, USA, June 23 2013 / Kristen LeFevre, Ashwin Machanavajjhala, Adam Silberstein (Conference Chairs). - New York, NY : ACM, 2013. - S. 1-6. - ISBN 978-1-4503-2191-4
http://dx.doi.org/10.1145/2484702.2484703

Event Identification for Local Areas using Social Media Streaming Data

Andreas Weiler, Marc H. Scholl
Database and Information Systems Group,
University of Konstanz
Box D188, 78457 Konstanz, Germany
{firstname.lastname}@uni-konstanz.de

Franz Wanner, Christian Rohrdantz
Data Analysis and Visualization Group,
University of Konstanz
Box D78, 78457 Konstanz, Germany
{firstname.lastname}@uni-konstanz.de

ABSTRACT

Unprecedented success and active usage of social media services result in massive amounts of user-generated data. An increasing interest in the contained information from social media data leads to more and more sophisticated analysis and visualization applications. Because of the fast pace and distribution of news in social media data it is an appropriate source to identify events in the data and directly display their occurrence to analysts or other users. This paper presents a method for event identification in local areas using the Twitter data stream. We implement and use a combined log-likelihood ratio approach for the geographic and time dimension of real-life Twitter data in predefined areas of the world to detect events occurring in the message contents. We present a case study with two interesting scenarios to show the usefulness of our approach.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

event detection, information extraction, social media analytics

1. INTRODUCTION AND MOTIVATION

The ongoing emergence of new web services, such as social media platforms and technologies impose new challenges on the way such data volumes are analyzed and prepared for users and analysts. Because a lot of users of information services are typically interested in current events and actual happenings of the world, but are unable to follow the large amount of data, it is a crucial task to support these users in the identification of events. A leading player in producing a large volume of data as a continuous stream is the social network platform Twitter. Over 140 million registered users and about 340 million messages – called "tweets" – per day let Twitter become the undisputed number one in microblogging today. In its initial stage, Twitter prompted the users to answer a simple question "What are you doing?" and so the users reported about their actual activities, feelings, and experiences of their everyday life. As Twitter gained significance and users started exchanging on matters beyond personal things, in November 2009 the question changed to a more general one "What's happening?"[1]. The intention of the new question is to also report about current news and events. The consequence of this change is that Twitter has developed into an expanded information source.

Due to the diversity of provided information, Twitter even plays an increasingly important role as a source for news agencies. In fact, news agencies use Twitter for two important functionalities in their daily work. On the one hand, agencies use Twitter as a publication and distribution platform for current news articles with a high throughput rate. For example, any reproduction of a tweet ("retweet") reaches an average of about 1,000 users [10]. On the other hand, news agencies, such as BBC, are constantly increasing the usage of Twitter as a reference in their daily news reports [21]. A further characteristic of Twitter is its vibrant user community with a wide range of different personalities from all over the world. Apparently, this whole spectrum can be subdivided into few categories of Twitter usage patterns, such as daily chatter, information and URL sharing, or news reporting [8]. Further research undertaking has discovered that the majority of users publish messages focusing on their personal concerns and matters, whereas a smaller set of users publish for information sharing [14].

In this paper we present an approach of using a combined significance measure calculated with the log-likelihood ratio for geographic and time dimension to identify events in the content of tweets located in predefined local areas. We support analysts or news reporters in identifying local events in self-defined areas and keeping the overview about event evolution over time. Our contribution is the effective elimination of random noise out of the data, which results in clearly transcribed events and an application to detect, track, summarize, and associate events identified in the Twitter stream in time. The case study (see Section 4) shows two interesting scenarios in which we apply our approach.

2. RELATED WORK

Extreme popularity of Twitter and the availability to access its public stream have resulted in the multiplication of

[1] `http://blog.twitter.com/2009/11/whats-happening.html`

Twitter-related scientific, industrial and governmental research initiatives. The work related to our contribution is overviewed in this section.

Bontcheva et al. [2] presented an overview about sense making of social media data, which also includes current event detection methods in social media streams. They classified detection methods into three categories: clustering-based, model-based, and those based on signal processing.

An event detection system dedicated to earthquakes was presented by Sakaki et al. [18]. In contrast to our approach, they use the keyword search feature provided by the Twitter API to gather data in specified time intervals. In [19] Schühmacher et al. proposed another domain-specific event detection method on microblogs to support forensic analysis. They trained a linear classifier to detect suspicious posts. In 2011 Weng et al. [22] used wavelet analysis on the frequency-based raw signals of the terms from tweets for detecting events. They used a keyword-filtered dataset to show their practical usage for identifying events during the Singapore General Election in 2011. In the same year Marcus et al. [12] demonstrated an application called "TwitInfo", which identifies and labels event peaks for given search queries related to the event.

Recently in 2012, Ritter et al. [16] presented the first approach for open domain event extraction from Twitter. Their approach is based on latent variable models and proceeds by first discovering event types, which match the data, and then using these results to classify aggregate events. However, no discussion about applying this approach directly to the streaming data is included. In the same year Alvanaki et al. [1] proposed a system "enBlogue", which analyzes statistics about tags and tag pairs for identifying unusual shifts in correlations. Further recent work proposed by Nishida et al. [15] shows a classification model of tweet streams for identifying changes in statistical properties on word basis, which is used for topic classification. Also in the same year Zimmermann et al. [23] propose a text stream clustering method that detects, tracks and updates large and small bursts of news in a two-level topic hierarchy. In contrast to our approach, they use the term "local" not as geographic property, but rather as a burst, which occurs in a previous global burst.

3. SYSTEM DESIGN

In this section we present the tasks and all components of our proposed system. First we describe the used data source, second the tasks we aim to solve with our system, third the used approach for event identification and last the visual representation of the identified events to support the analyst or news reporter in identifying and following events.

3.1 Data Source

The Twitter platform provides direct access to the public live stream of Twitter. By using the Twitter Streaming API[2] with the so-called "Gardenhose" level, we are able to collect 10% of the public live stream. To increase the number of geo-tagged tweets in the dataset, we additionally merge the 10% sample with tweets from geo-filtered streams. This results in a duplicate free dataset with the highest possible amount of tweets. The geographic information is used to assign the different tweets to the belonging areas and is set automatically by the mobile device of the Twitter user. An exemplary evaluation of a representative sample of days shows that we are able to receive an average of over 1.5 million tweets per hour with the average of 20.000 tweets per minute. We can also conclude that about 2.000 to 4.000 of the incoming tweets per minute have geographic information available. The recent observations show that the amount of geo-tagged tweets is steadily increasing.

[2]`https://dev.twitter.com`

3.2 Tasks

Our system design is encouraged by Dou et al. [5]. They define four tasks for "Event Detection in Social Media Data": "New Event Detection", "Event Tracking", "Event Summarization", and "Event Association".

Both event tracking and association are tasks, which our application provides. But also event summarization is a task we take into account through our visualization. The problem with all these tasks around the term *event* is, that an event itself is often not defined exactly. The level of granularity of the underlying data in what you are looking for an event is crucial. If an event is a single tweet and the task of event detection refers to the detection of this tweet as first story, then we are not able to detect it. Our methods base on statistical significance of reoccurring terms in space and time and therefore we find a story only as a composition and aggregation of content of several tweets. So if *story* means a statistical significant term or hashtag, then we are able to detect it and to enrich it with further terms.

3.3 Event Term Identification

Event detection is a classical problem in computer science and has been under discussion for many years in various research areas. A lot of research deals with detecting anomalies or novelties in different data sources ([3, 13]), considering those phenomena to be an indication of an event. Further related research deals with the detection of changes or drifts in data streams ([6, 9]).

Taking into account a vast number of Twitter messages generated each second, it becomes a crucial task to group messages by topics or events. Because there is no explicit knowledge about current or future events, the framework has to identify events in an on-line fashion without limitation to any domain. We suggest that only by means of aggregation it is possible to handle large amount of data and gain important insights into it. Therefore, we use the detected high-level representations to compress the tweets in a meaningful manner.

For our work, we define an event as a term, which occurs significantly higher in a certain local area than in other areas of the world and significantly higher in this local area for the current time frame than in a past time frame. Therefore an event term has a specific location where it occurs and a specific time when it occurs.

To extract the terms of the tweets, we filter the incoming stream for tweets with geographic information only. Then we check if the tweet object is inside or outside the predefined area and categorize the tweet. Afterwards the tweet content is tokenized and analyzed using a tokenizer and Part-Of-Speech Tagger tailored especially for Twitter [7]. Only nouns, proper nouns, and hashtags will remain and all tokens with unknown characters or contained in a standard English stop word list are discarded. To avoid wrong identification of events by continuous repetition of the same term,

we keep the list of terms per tweet content duplicate-free and only take tweets into account, which are not retweeted in the moment of the analysis. To obtain the historical values we apply the same approach applied to data of a pre-defined time frame from the past.

In order to detect *events* we monitor the occurrence frequencies of terms in tweets across time and space. Users can select geographic areas of interest. Whenever the frequency of a certain term bursts, i.e. becomes much higher than expected, within a user-selected geographic area, it may potentially point us to a local event. We approach the detection of such bursts from two different angles in that we require two different criteria to be fulfilled:

1. Within the tweets of the selected area, we require the term to be statistically significantly much more frequent in the recent past (e.g., last hour) than we would expect it to be from considering the complete history of the stream. We leave it up to the user to define what s/he would like to consider as recent past.

2. Taking into account only this recent past, we require the term to be statistically significantly much more frequent in the user-selected area than outside this area.

The combination of both criteria, salience in time and salience in space, assures that the method is robust with respect to random noise in the data.

In our approach to measure the significance of an incoming term we use the log-likelihood ratio test, which operates on a contingency table and has been used before to measure the strength of word collocations [11]. This method is also used by Rohrdantz et al. [17] to identify context-coherence of terms in consumer feedback comments. In contrast to this approach, we build two contingency tables (see Table 1 and Table 2), one for the geographic dimension and one for the time dimension. By using the geographic dimension we measure the significance of terms in a pre-defined region (e.g. see Figures 1) against all other areas of the world. We calculate the current frequency of terms in the pre-defined area and the frequency of terms in the rest of the world. With this knowledge we create the contingency table (see Table 1) and use the log-likelihood ratio equation to calculate the significance measure. Hereby we can identify region specific terms, which are not included in tweets around the world.

The time dimension is used to identify the significance of terms in the pre-defined area over time. Here we make use of historical data and build the contingency table (see Table 2) with a historical time frame and the current time frame. By applying the log-likelihood ratio equation we can identify significant terms in the current time frame and therefore eliminate terms like "Birthday", which occur very frequent all the time. Furthermore we can eliminate specific local terms like "Boston", which also always occur very frequent in a pre-defined area.

The document counts were used to calculate the significance value with the equation of the log-likelihood ratio, where A; B; C and D correspond to the four cells in Table 1 or Table 2. For the continuous execution of the analysis we integrate the current values after the closing of the time frame into the historical values. The log-likelihood ratio is

	$TWEETS \in R$	$TWEETS \notin R$
$T \in TWEETS$	A	B
$T \notin TWEETS$	C	D

Table 1: Contingency table showing the number of tweets *TWEETS* out of a set of all tweets depending on a certain term *T* and a certain region *R*.

	$TWEETS \in TI$	$TWEETS \notin TI$
$T \in TWEETS$	A	B
$T \notin TWEETS$	C	D

Table 2: Contingency table showing the number of tweets *TWEETS* depending on a certain term *T* and a certain time frame *TI*. This table contains only tweets within the pre-defined region *R*.

calculated by using the following equation:

$$\textbf{log-likelihood ratio} =$$

$$2 * \left(A\ log\ \left(\frac{A/(A+B)}{(A+C)/(N)} \right) + B\ log\ \left(\frac{B/(A+B)}{(B+D)/(N)} \right) \right.$$
$$\left. +C\ log\ \left(\frac{C/(C+D)}{(A+C)/(N)} \right) + D\ log\ \left(\frac{D/(C+D)}{(B+D)/(N)} \right) \right)$$
$$with\ N = A+B+C+D$$

If both significance values are greater than the critical value 10.83, we multiply both values with each other and get a total significance value per term. After the analysis of all terms we rank the terms by the significance value and take the top 10 significant terms per time frame. We also combine these terms with the top 5 most co-occurring terms of that term. We start with the most significant term and analyze the co-occurrence terms, if a term in this set is also identified as a significant term the co-occurrence terms of this hit will also be added to the result set. We assume that there is a relationship between two event terms, if at least two co-occurrence terms or the event term itself (or the corresponding version with or without hashtag) occur in the co-occurrence terms of another event term. Hereby we are able to summarize some of the significant terms and their co-occurrence terms as one resulting event. The terms describing an event are also limited to a maximum of 10 terms per time frame. For subsequent time frames we check the currently existing events and analyze if the newly occurred events are related to any of the existing ones. This analysis also takes the event term and the co-occurrence terms into account. If an event only consists of terms with a hashtag at the beginning, we consider this event as not very useful and discard it. Note, all defined thresholds can be easily changed and adapted to the corresponding use case.

3.4 Result Visualization

The result visualization (see Figures 2 & 3) is designed in a timeline fashion with the starting point on the left and all subsequent time frames aligned on the next position on the right. For each resulting event, which is merged from several event terms, a new line is added at the bottom of the visualization. In the first occurrence of the event at most 10 co-occurring and significant terms defining that event are displayed. In the following time frames only new occurring terms for the event are displayed. Hereby we support the an-

alyst in keeping the overview of new emerging terms around and easily detect the development of an event. The rounded rectangle is colored by using a heat map color scheme depending on the amount of events in the time frame and the significance value of the single event. By using the color scheme, we are able to display the significance rank of the events.

4. CASE STUDY

In this section we present two scenarios of applying our framework to identify event terms in the incoming stream of tweets. For both scenarios the average amount of terms, which have both significance values above the critical limit, is approximately 100 and the average amount of co-occurrence terms per event is about 12 per hour. The historic dataset contains about 44 million of tweets with geographic information (about 280 million in total). About 2 million of tweets are located in the specified areas for both scenarios. Each hour of the live analysis contains about 200.000 tweets with geographic information and about 10.000 located in the specified area. The live analysis of an hour of streaming Twitter data takes about half an hour. The analysis of the historic data for a pre-defined area takes about three minutes on a *3.2 GHz Intel Core i3 iMac* by using *2GB of RAM*.

The first scenario deals with the area of the Northeast of the USA (see Figure 1 (a)) on 7/8 February 2013 from 11 PM to 03 AM, GMT Timezone (6 PM to 10 PM Local Time, EST). We use one week of historic data (February 1st to 7th) to analyze the time specific terms for the chosen area. Figure 2 shows the resulting visualization for the corresponding time frame.

In the first two hours the three events *storm*, *#nyfw*, and *Islanders* occur. The most significant event that emerges in the first hour and lasts for at least four hours is a *storm*, which takes place in the Northeast of the USA. The co-occurrence terms of the event indicate that the *storm* is a *blizzard* and the terms *nemo* and *#nemo* describe the given name of the snowstorm. During the next hours co-occurrence terms like *warning*, *food*, *lines*, and *bread* describe the situation around the *storm*. These terms indicate that the people on-site should be well prepared for the upcoming *storm*. Hereby it is possible to have a closer insight into the event and the analyst can see how the situation around an event evolves over time.

The second identified event in the first hour is *#nyfw*, which means the "New York Fashion Week" (*#nyc*, *fashion*), where *Mercedes-Benz* was eponym for the event (*#mbfw*). This event however is only significant for the first two hours of the analysis. The third identified event is about the term *Islanders*, a New York hockey team, which played at 7 PM EST against the New York Rangers (*Rangers*). The event is already identified one hour before the beginning of the game. This is mainly due to the fact that shortly before the beginning of the match, a *trade* of *goalie Timmy Thomas* from the "Boston Bruins" (*Bruins*) to the "New York Islanders" was made. The co-occurrence terms indicate that the game took place at *Madison Square Garden*, a *goal* was shot in the second hour, and some player names (*Miller*, *Timmy Thomas*, *Kreider*) are mentioned.

In the last two hours of the analysis four new events *#housofanubis*, *Celtics*, *#iloveitwhen*, and *#impactlive* occur. Whereas *#houseofanubis* is a TV series and the users

(a) Chosen area of the Northeast of the USA for the first scenario.

(b) Chosen area of the Southwest of the USA for the second scenario.

Figure 1: Chosen areas for the scenarios.

talk about the broadcast and the persons occurring in it, the third and fourth event seem to be an exclusively occurring phenomena on Twitter. We have found no corresponding real-life event, which belongs to the terms and hashtags forming out these events. The event *Celtics* describes a NBA game between the "Boston Celtics" and the "Los Angeles Lakers". The game started at 8 PM EST and therefore the analysis identifies the event in the correct time frame. In the last hour this event supersedes the *storm* and becomes the most significant event. The co-occurrence term *#beatla* suggests that the perspective of the fans from Boston is reflected in the data. Further co-occurrence terms mention player names of both teams (*Kobe Bryant*, *Dwight Howard*, *Rajon Rondo*) and the arena called "TD Garden".

The continuous appearance of the same event (e.g., *storm*) shows that this event is still getting more and more significant for that area. We can derive that a natural disaster, like the *storm*, is longer and more significant in the social media than social events, like fashion shows or sport events.

The second scenario deals with the data of the same time span (3 PM to 7 PM Local Time, PST). but this time the area selection contains the Southwest of the USA (see Figure 1 (b)) for the analysis.

In the first hour the most interesting event is about the term *Dorner*. This event is about the killing incident by the former *LAPD* officer *Christopher Dorner*. He *shot* four people and was found after a long lasting *manhunt*. The *manifesto* he wrote before he died is also mentioned in the event. The other two identified events (*freeze*, *tracking*) are about a poker game and an automatic reporting system for graffiti and are therefore not of much interest.

An interesting observation is the event *Celtics*. It describes the same event as mentioned before in the scenario from the Northeast. However, it is identified one hour earlier and reflects the perspective of the fans from Los Angeles (*#golakers*).

In the third hour there is also a weather event identified. The event *showers* is about an expected *t-storm* with *snow* in the San Francisco Valley. Furthermore the approach identifies an exhibition from *#kubrick* and some social "virtual" events like *#codymmag* and *#poppunk*.

5. CONCLUSIONS

In this paper, we present a method for identifying events in real-world social media data streams. We have shown

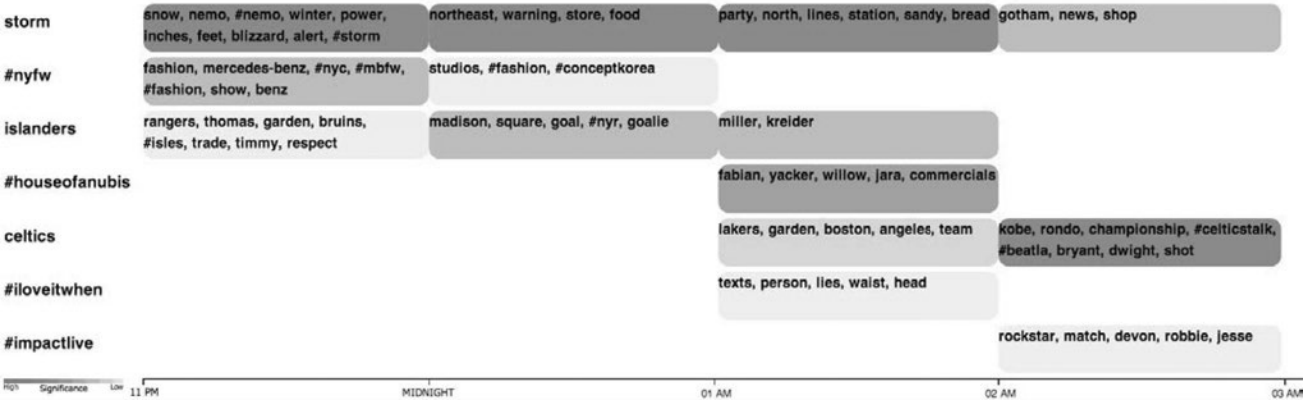

Figure 2: Results from 11 PM to 03 AM, GMT (6 PM to 10 PM Local Time, EST) on 7/8 February 2013 for the Northeast area of the USA.

that by means of aggregation it is possible to handle large volumes of data and gain important insights into it. We believe that under ideal conditions the data streamed by Twitter would provide a faster support for detecting events than the reports by news agencies. We obtained the data through the Twitter API that provides just 10% of the total data stream. Therefore, our access is somewhat random and the total stream may in fact contain more tweets about an event. Our scenarios show that we are able to identify actual and ongoing events, which belong to a certain region and are outstanding over time, by using the log-likelihood ratio approach.

As pre-evaluative study and for comparison reasons, we also identified the top 10 most frequent terms of the time frames for the pre-defined areas for both scenarios by using a standard IDF [20] measurement. The result shows that the standard approach also detects some of our "real world" events like *storm* and *showers*. However, we can derive that this approach identifies much more noisy and always repeating terms like "Dinner", "Hour", and "Birthday" in most of the time frames. We further can derive that in both areas specific local terms are identified. In the Northeast terms like "York", "Boston" and in the Southwest terms like "Vegas", "Francisco", "Angeles", "California", and "Diego". All of theses terms are city or region names and always repeating terms for the belonging areas and therefore not important for our event identification.

The combination of the time and the geography dimension of the data ensure that our results only contain significant terms for the chosen region and time frame. Our contribution is the effective elimination of random noise out of the data, which results in clearly transcribed events and an application to detect, track, summarize, and associate events from Twitter data over time.

6. FUTURE WORK

The integration of additional data sources could be a first extension of our framework. Stock exchange markets, weather forecasts, data from news agencies, RSS feeds, and further social media services offer contents that can be retrieved in different ways as streams and could also enrich our event identification and tracking analysis. For example, even the social media photo sharing platform Flickr was recently used as data source for event detection [4]. Further information about the events could also help to differentiate between "real world" and "virtual" events. The scenarios show the first interesting results and therefore it is planned to evaluate these results against other event identification methods.

A further really interesting extension is the analysis of local and global reactions to identified events. Additionally it would be from interest to track local and global drifts in the reactions about an event. Hereby we could improve the situational awareness of analysts and news reporters for events.

Furthermore we plan to give the analyst more possibilities for an exploratory analysis and to develop a more space saving visualization. The visualization should fulfill all the requirements to support the analyst coping with the mentioned tasks. Since we integrate actual results of the analysis into the historic dataset, we are only able to track events, which have an ongoing significant behavior. This could also be extended to automatically track all ever-occurred events till the end of the event.

7. REFERENCES

[1] F. Alvanaki et al. See what's enblogue: real-time emergent topic identification in social media. In *Proceedings of the 15th International Conference on Extending Database Technology*, EDBT '12, pages 336–347, New York, NY, USA, 2012. ACM.

[2] K. Bontcheva and D. Rout. Making Sense of Social Media Streams through Semantics: a Survey. *Semantic Web*, 2012.

[3] V. Chandola et al. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3), 2009.

[4] L. Chen and A. Roy. Event detection from Flickr data through wavelet-based spatial analysis. In *Proceedings of the 2009 ACM International Conference on Information and Knowledge Management (CIKM '09)*, 2009.

[5] W. Dou et al. Event Detection in Social Media Data. In *IEEE VisWeek Workshop on Interactive Visual Text Analytics - Task Driven Analytics of Social Media Content*, 2012.

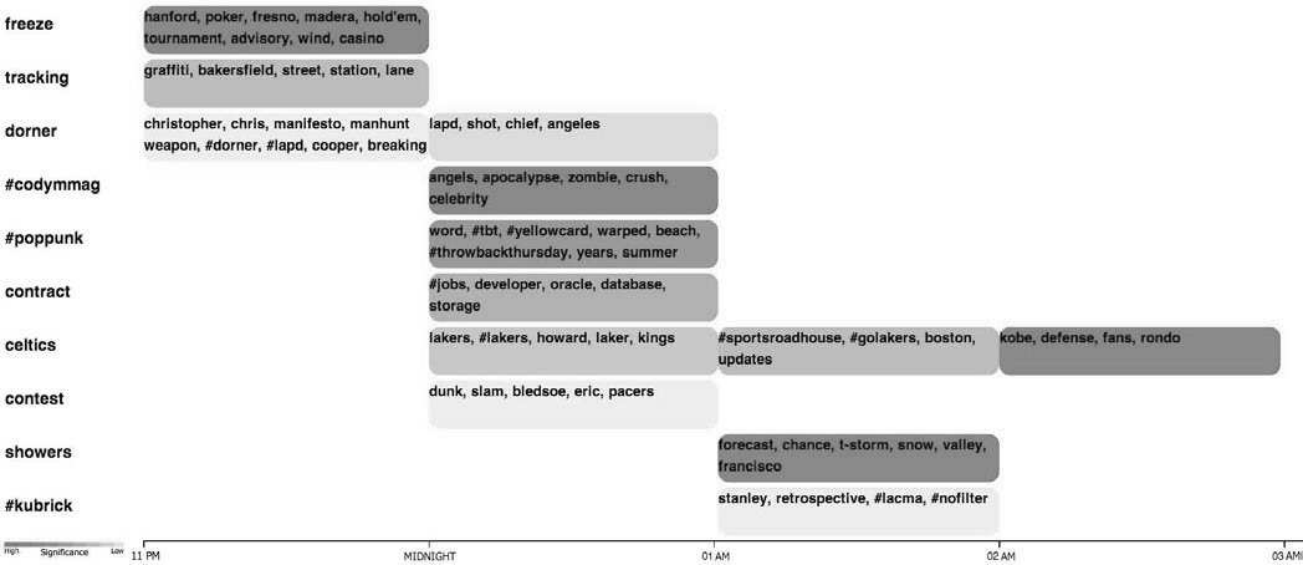

Figure 3: Results from 11 PM to 03 AM, GMT (3 PM to 7 PM Local Time, PST) on 7/8 February 2013 for the Southwest area of the USA.

[6] A. Dries and U. Rückert. Adaptive concept drift detection. *Stat. Anal. Data Min.*, 2:311–327, 2009.

[7] K. Gimpel et al. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In *ACL (Short Papers)*, pages 42–47, 2011.

[8] A. Java et al. 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*, pages 56–65. ACM, 2007.

[9] D. Kifer et al. Detecting change in data streams. In *Proceedings of the Thirtieth international conference on Very large data bases - Volume 30*, VLDB '04, pages 180–191. VLDB Endowment, 2004.

[10] H. Kwak et al. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600. ACM, 2010.

[11] C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA, 1999.

[12] A. Marcus et al. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 227–236. ACM, 2011.

[13] M. Markou and S. Singh. Novelty Detection: A Review - Part 1: Statistical Approaches. *Signal Processing*, 83:2003, 2003.

[14] M. Naaman et al. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, pages 189–192. ACM, 2010.

[15] K. Nishida et al. Improving tweet stream classification by detecting changes in word probability. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 971–980, New York, NY, USA, 2012. ACM.

[16] A. Ritter et al. Open domain event extraction from Twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pages 1104–1112, New York, NY, USA, 2012. ACM.

[17] C. Rohrdantz et al. Feature-based visual sentiment analysis of text document streams. *ACM Trans. Intell. Syst. Technol.*, 3(2):26:1–26:25, Feb. 2010.

[18] T. Sakaki et al. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 851–860. ACM, 2010.

[19] J. Schühmacher and C. Koster. Signalling events in text streams. volume 40 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 335–339. Springer, 2009.

[20] K. Sparck Jones. *A statistical interpretation of term specificity and its application in retrieval*, pages 132–142. Taylor Graham Publishing, 1988.

[21] E. Tonkin et al. Twitter, information sharing and the London riots? *Bulletin of the American Society for Information Science and Technology*, 38, 2012.

[22] J. Weng et al. Event Detection in Twitter. Technical report, HP Labs, 2011.

[23] M. Zimmermann et al. Discovering global and local bursts in a stream of news. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, SAC '12, pages 807–812, New York, NY, USA, 2012. ACM.