Spatiotemporal Social Media Analytics for Abnormal Event Detection and Examination using Seasonal-Trend Decomposition

Junghoon Chae*   Dennis Thom†   Harald Bosch†   Yun Jang‡   Ross Maciejewski§
Purdue University   University of Stuttgart   University of Stuttgart   Sejong University   Arizona State University

David S. Ebert*   Thomas Ertl†
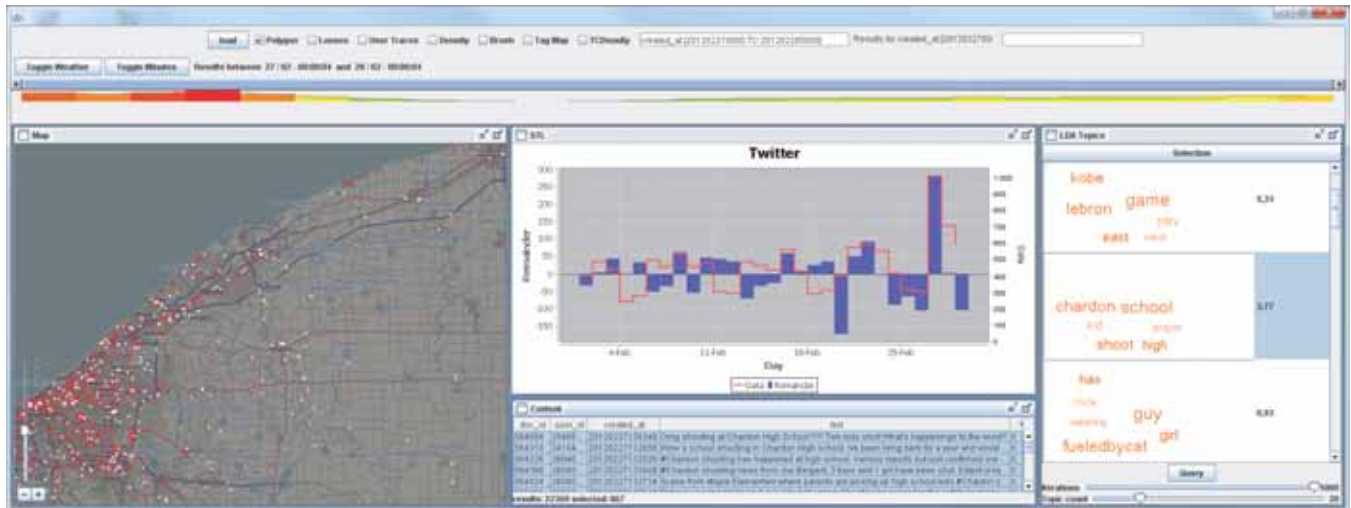Purdue University   University of Stuttgart

Figure 1: Social media analysis system including message plots on a map, abnormality estimation charts and tables for message content and topic exploration. It can be seen, how the Ohio High School Shooing on February 27, 2012 is examined using the system. The selected messages, marked as white dots on the map, show retrieved Tweets that are related to the event.

Abstract

Recent advances in technology have enabled social media services to support space-time indexed data, and internet users from all over the world have created a large volume of time-stamped, geo-located data. Such spatiotemporal data has immense value for increasing situational awareness of local events, providing insights for investigations and understanding the extent of incidents, their severity, and consequences, as well as their time-evolving nature. In analyzing social media data, researchers have mainly focused on finding temporal trends according to volume-based importance. Hence, a relatively small volume of relevant messages may easily be obscured by a huge data set indicating normal situations. In this paper, we present a visual analytics approach that provides users with scalable and interactive social media data analysis and visualization including the exploration and examination of abnormal topics and events within various social media data sources, such as Twitter, Flickr and YouTube. In order to find and understand abnormal events, the analyst can first extract major topics from a set of se-

*e-mail: {jchae|ebertd}@purdue.edu
†e-mail: {forename.surname}@vis.uni-stuttgart.de
‡e-mail: jangy@sejong.edu
§e-mail: rmacieje@asu.edu

lected messages and rank them probabilistically using Latent Dirichlet Allocation. He can then apply seasonal trend decomposition together with traditional control chart methods to find unusual peaks and outliers within topic time series. Our case studies show that situational awareness can be improved by incorporating the anomaly and trend examination techniques into a highly interactive visual analysis process.

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces—GUI; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Information filtering, relevance feedback

1 Introduction

Social media services, e.g, Twitter, Youtube, Flickr, provide a rich and freely accessible database of user-generated situation reports. As advances in technology have enabled the widespread adoption of GPS enabled mobile communication devices, these reports are able to capture important local events observed by an active and ubiquitous community. The different forms of social media content provided by the users, such as microposts, images or video footage, can have immense value for increasing the situational awareness of ongoing events.

However, as data volumes have increased beyond the capabilities of manual evaluation, there is a need for advanced tools to aid understanding of the extent, severity and consequences of incidents, as well as their time-evolving nature, and to aid in gleaning investigative insights. Due to the

143

large number of individual social media messages it is not straightforward to analyze and extract meaningful information. For example, in Twitter, more than 200 million Tweets are posted each day [31]. Thus, in a developing event, the relevant messages for situational awareness are usually buried by a majority of irrelevant data. Finding and examining these messages without smart aggregation, automated text analysis and advanced filtering strategies is almost impossible and extracting meaningful information is even more challenging.

To address these challenges, we present an interactive spatiotemporal social media analytics approach for abnormal topic detection and event examination. In order to find relevant information within a user defined spatiotemporal frame we utilize Latent Dirichlet Allocation (LDA) [1], which extracts and probabilistically ranks major topics contained in textual parts of the social media data. The ranks of the categorized topics generally provide a volume-based importance, but this importance does not reflect the abnormality or criticality of the topic. In order to obtain a ranking suitable for situational awareness tasks, we discard daily chatter by employing a Seasonal-Trend Decomposition procedure based on Loess smoothing (STL) [5]. In our work, globally and seasonally trending portions of the data are considered less important, whereas major non-seasonal elements are considered anomalous and, therefore, relevant.

However, due to the large volumes of data, the very specific syntax and semantics of microposts and the complex needs of situational analysis, it would not be feasible to apply these techniques in the form of a fully automated system. Therefore, our whole analysis process, including the application of automated tools, is guided and informed by an analyst using a highly interactive visual analytics environment. It provides tight integration of semi-automated text-analysis and probabilistic event detection tools together with traditional zooming, filtering and exploration following the Information-Seeking Mantra [24].

The remainder of this document is structured as follows: Section 2 is a review of related work. The automated methods to find and examine unusual topics and events are described in Section 3. In Section 4 we briefly introduce our visual analytics system Scatterblogs, which was already featured in previous works, and explain how the automated methods are integrated within a sophisticated iterative analysis loop. Finally we demonstrate the performance of our system based on selected case studies in Section 5 and discuss the approach in Section 6.

2 Related Work

In recent research, social media services have become a popular and influential data source for many domains. Researchers in the fields of data mining and visual analytics have found through studies among users and domain experts, that the analysis of such data can be essential for spatiotemporal situational awareness [15, 23]. Thus, as the size of social media data increases, scalable computational tools for the effective analysis and discovery of critical information within the data are a vital research topic. This section presents previous work that has focused on spatiotemporal and event related social media analysis.

2.1 Spatiotemporal Social Media Data Analysis

As social media platforms move towards location-based social networks (LBSNs) researchers have proposed various approaches to analyze spatiotemporal document collections, in general, and spatiotemporal social media data, in particular.

VisGets [7] provides linked visual filters for the space, time and tag dimensions to allow the exploration of datasets in a faceted way. The user is guided by weighted brushing and linking, which denotes the co-occurrences of attributes. Further works demonstrate the value of visualizing and analyzing the spatial context information of microblogs for social network users [9] or third parties like crime investigators [22] and urban planners [33]. With Senseplace2, MacEachren et al. [15] demonstrate a visualization system that denotes the message density of actual or textually inferred Twitter message locations. The messages are derived from a textual query and can then be filtered and sorted by space and time. Their work also has shown that social media can be a potential source for crisis management. With ScatterBlogs [2], our own group developed a scalable system enabling analysts to work on quantitative findings within a large set of geolocated microblog messages. In contrast to Senseplace2, where the analysts still have to find and manage the appropriate keywords and filters to gather relevant messages in the high volume of insignificant messages, we propose a semi-automatic approach that finds possibly relevant keywords and ranks them according to their 'abnormality'.

Special LBSN for certain domains, like Bikely[1] and EveryTrail[2] have an even stronger focus on the sharing and tracing of user locations. Ying et al. [37] present various location based metrics using spatial information of these LBSNs to observe popular people who receive more attention and relationships within the network. Similarly, there are many related works for non-spatial temporal document collections, for example IN-SPIRE [36], which is a general purpose document analysis system that depicts document clusters on a visual landscape of topics.

2.2 Social Media Event Detection and Topic Extraction

One of the major challenges in analyzing social media data is the discovery of critical information obscured by large volumes of random and unrelated daily chatter. Due to the nature of microblogging, message streams like Twitter are very noisy compared to other digital document collections. Recently, many researchers have tried to solve this challenge by means of automated and semi-automated detection and indication of relevant data.

Sakaki et al. [23] propose a natural disaster alert system using Twitter users as virtual sensors. In their work, they were able to calculate the epicenter of an earthquake by analyzing the delays of the first messages reporting the shock. Weng and Lee [34] address the challenge by constructing a signal for each word occurring in Twitter messages using wavelet analysis, thereby making it easy to detect bursts of word usage. Frequently recurring bursts can then be filtered by evaluating their auto-correlation. The remaining signals are cross correlated pairwise and clustered using a modularity-based graph partitioning of the resulting matrix. Due to the quadratic complexity of pairwise correlation, they rely on heavy preprocessing and filtering to reduce their test set to approx 8k words. As a result, they detected mainly, large sporting events, such as soccer world cup games, and elections. Our approach, in contrast, provides a set of topics through a probabilistic topic extraction algorithm which can be iteratively applied to subsets and subtopics within user selected message sets.

Lee and Sumiya [14] as well as Pozdnoukhov and Kaiser [19] present methods to detect unusual geo-social events by measuring the spatial and temporal regularity of

---

[1] http://www.bikely.com/
[2] http://www.everytrail.com/

Twitter streams. Lee and Sumiya propose a concept to detect unusual behavior by normalizing the Twitter usage in regions of interests which are defined by a clustering-based space partitioning. However, their results are mainly a measurements of unusual crowd behavior and do not provide further means for analyzing the situation. Pozdnoukhov and Kaiser observe abnormal patterns of topics using spatial information embedded in Twitter messages. Similar to our approach, they apply a probabilistic topic model (Online Latent Dirichlet Allocation) as a means of analyzing the document collection. A Gaussian RBF kernel density estimation examines the geo-spatial footprint of the resulting topics for regularities. The usual message count of identified areas is then learned by a Markov-modulated non-homogeneous Poisson process. The spatial patterns are shown as a static heat map. The resulting system does not provide interactive analytics capabilities.

Recently, researchers have applied LDA topic modeling to social media data to summarize and categorize Tweets [39] and find influential users [35]. Zhao et al. [39] demonstrate characteristics of Twitter by comparing the content of Tweets with a traditional news medium, such as the New York Times. They discuss and adapt a Twitter-LDA model and evaluate this model against the standard topic model and the so-called author-topic model [25], where a document is generated by aggregating multiple tweets from a single user, in terms of meaningfulness and coherence of topics and Twitter messages. In this work, we do not use the author-topic model, since a users Tweet timeline is usually a heterogeneous mixture of unrelated comments and messages and not a homogenous framework of interrelated topics like a traditional document. Furthermore, the evaluation of Zhao et al. [39] shows that the standard model has quite reasonable topic modeling results on Tweets, although the Twitter-LDA model outperforms the standard model. Works from Ramage et al. [20] also show promising results in LDA based Twitter topic modeling by evaluating another type of LDA model (Labeled LDA) [21]. ParallelTopics [8] also extracts meaningful topics using LDA from a collection of documents. The visual analytics system allows users to interactively analyze temporal patterns of the multi-topic documents. The system, however, does not not deal with spatial information, but takes an abnormality estimation into account.

In our previous work [26], we proposed a spatiotemporal anomaly overview based on a streaming enabled clustering approach that is applied for each term in the dataset individually. The resulting clusters can be used to generate a spatially and temporally explorable term map of large amounts of microblog messages as an entry point for closer examination. Even though the scalable event detection and our current approach share the same workbench, they can be used independently as well as complementary. The combination of LDA and STL allows for an ad-hoc analysis of a user selected set of messages regarding the topical distribution of messages and the abnormal presence of topics. Due to this characteristic, it provides an iterative analysis loop for qualitative analysis and drill down operations.

3 Spatiotemporal Social Media Analytics for Event Examination

Since several social media sources recently provide space-time indexed data, traditional techniques for spatiotemporal zooming, filtering and selection can now be applied to explore and examine the data. However, as message volumes exceed the boundaries of human evaluation capabilities, it is almost impossible to perform a straightforward qualitative analysis of the data. In order to cope with the data volumes, traditional interaction and visualization techniques have to be enhanced with automated tools for language processing and signal analysis, helping an analyst to find, isolate and examine unusual outliers and important message subsets.

To address this issue, we present an interactive analysis process that integrates advanced techniques for automated topic modeling and time series decomposition with a sophisticated analysis environment enabling large scale social media exploration. In part 3.1 of this Section we first explain how the Latent Dirichlet Allocation, a well established topic modeling technique in the information retrieval domain, can be used to extract the inherent topic structure from a set of social media messages. The output of this technique is a list of topics each given by a topic proportion and a set of keywords prominent within the topics messages. In a subsequent step, our system then re-ranks the retrieved topic list by identifying unusual and unexpected topics. This is done by employing a seasonal-trend decomposition algorithm to the historic time series data for each topic, retrieving its seasonal, trending and remainder components. Using a z-score evaluation, we locate peaks and outliers in the remainder component in order to find an indicator of unusual events. While the LDA topic extraction is done primarily for Twitter data, the abnormality estimation is also applied to different social media data sources, such as Flickr and YouTube, for each topic. This is achieved by searching matching entries for each term of a topic and applying the same STL analysis on the resulting time series. The results are available to the analyst for cross validation. The details of this step are described in Subsection 3.2 and the complete detection model is formally described in Subsection 3.3. In Section 4, we describe how powerful tools based on these techniques are used within our analysis environment, Scatterblogs, in order to iteratively find, isolate and examine relevant message sets.

3.1 Topic Extraction

Our monitoring component collects space-time indexed Twitter messages using the Twitter-API. The received messages are preprocessed and then stored in our local database. When users of these services witness or participate in unusual situations they often inform their friends, relatives or the public about their observations. If enough users participate, the communication about the situation constitutes a topic that makes up a certain proportion of all messages within the database, or some messages within a predefined area and timespan. In most cases, however, the proportion will be smaller than that of other prevalent topics, such as discussions about movies, music, sports or politics. In order to extract each of the individual topics exhibited within a collection of social media data, we employ Latent Dirichlet

Rank	Proportion	Topics
1	0.10004	day back school today
2	0.09717	lls bout dat wit
3	0.09443	people make hate wanna
4	0.08226	**earthquake thought house shaking**
5	0.05869	**earthquake felt quake washington**

Table 1: An example of extracted topics and their proportions. We extracted topics from Tweets written on August 23, 2011 around Virginia, where an earthquake occurred on this day. One can see that topics consisting of ordinary and unspecific words can have high proportion values, while the earthquake related topics have a relatively low proportion value.

Number of Iteration Steps in the LDA process		
50	300	1000
foursquare pic hall brooklyn	time back night day	time night nyc day
time night day back	york ave brooklyn btw	york ave brooklyn park
newyork nyc tweetmyjobs finance	pic bar food nyc	**foursquare occupywallstreet mayor ousted**
york brooklyn ave street	**foursquare occupywallstreet park mayor**	newyork tweetmyjobs finance citigroup
york ave park btw	newyork tweetmyjobs finance citigroup	**san gennaro street italy**

Table 2: An example of topic model results depending on the number of iteration steps in the LDA process. The topics are extracted from the Tweets posted in New York City on September 17 and 18, 2011 where the Occupy Wall Street protest movement began and a famous festival, *San Gennaro* occurred. A higher number of sampling iterations provides a better topic retrieval describing the two different events.

Allocation, a probabilistic topic model that can help organize, understand, and summarize vast amounts of information.

The LDA topic model approach, as presented by David Blei et al. [1], is a probabilistic and unsupervised machine learning model to identify latent topics and corresponding document clusters from a large document collection. Basically, it uses a "bag of words" approach and assumes that a document exhibits multiple topics distributed over words with a Dirichlet prior. In other words, the LDA assumes the following generative process for each document: First, choose a distribution over topics, choose a topic from the distribution for each word, and choose a word associated with the chosen topic. Based on this assumption one can now apply a Bayesian inference algorithm to retrieve the topic structure of the message set together with each topic's statistical proportion and a list of keywords prominent within the topic's messages. Table 1 shows an example set of extracted topics resulting from the application of LDA to Twitter data ordered by the proportion ranking. The example social media data was collected from Twitter for the Virginia area on August 23rd. On this day, the area was struck by an earthquake with a magnitude of 5.88. As seen in the table, this earthquake event was captured as a topic within the Twitter messages.

In our system, the MALLET toolkit [18] is used for the topic analysis. Prior to the topic extraction, the stemming algorithm KSTEM by Krovetz [13] is applied to every term in the messages. The results of KSTEM are more readable and introduce fewer ambiguities than the often used Porter stemmer.

For the unsupervised LDA classification and topic retrieval one has to define two parameters: the number of expected topics and the number of iterations for the Gibbs sampling process[10], which is used in MALLET for the topic inference. The number of topics that should be chosen depends on the size of the document collection and the required overview level. A small number of topics (e.g., 10) will provide a broad overview of the documents, whereas a large number (e.g., 100) provides fine-grained results. The number of sampling iterations is a trade-off between computation time and the quality of discovered topics. To illustrate this, Table 2 shows the experimental results of the topic model using a varying number of sampling iterations while the number of topics was set to four. The topics were extracted from Tweets posted in New York City on September 17 and 18, 2011, where a large group of protesters occupied Wall Street in New York City[3]. A topic indicating the Occupy Wall Street protests can be seen when using at least 300 iterations. At the time of these protests, there was also a famous annual festival, the *San Gennaro*[4], occurring in Little Italy.

---

[3]http://occupywallst.org/
[4]http://www.sangennaro.org/

This can only be seen when using at least 1000 iterations. As shown in Table 2, the topics with 50 iterations do not indicate any meaningful events. The topics with 300 iterations, on the other hand, consist of more distinguishable classes. Finally, the topics with 1000 iterations obviously point out individual events which happened in the city.

3.2 Abnormality Estimation using Seasonal-Trend Decomposition

Abnormal events are those that do not happen frequently and usually they cover only a small fraction of the social media data stream. As shown in Table 1, even during an earthquake episode, highly ranked topics consist of ordinary and unspecific words. The third and fourth ranked topics include words indicating the earthquake event of August 2011: *earthquake felt quake washington*. From this observation in the distributions of ordinary and unusual topics over the social media data, it is necessary to differentiate the unusual topics from the large number of rather mundane topics. In order to identify such abnormal topics, we utilize Seasonal-Trend Decomposition based on locally-weighted regression (Loess) known as STL [5]. For each extracted topic of the LDA topic modeling, our algorithm retrieves messages associated with the topic and then generates a time series consisting of daily message counts from their timestamps. The time series can be considered as the sum of three components: a trend component, a seasonal component, and a remainder:

$$Y = T + S + R \qquad (1)$$

Here Y is the original time series of interest, T is the trend component, S is the seasonal component, and R is the remainder component. STL works as an iterative nonparametric regression procedure using a series of Loess smoothers [6]. The iterative algorithm progressively refines and improves the estimates of the trend and the seasonal components. The resulting estimates of both components are then used to compute the remainder: $R = Y - T - S$. Under normal conditions, the remainder will be identically distributed Gaussian white noise, while a large value of R indicates substantial variation in the time series. Thus, we can utilize the remainder values to implement control chart methods detecting anomalous outliers within the topic time series. We have chosen to utilize a seven day moving average of the remainder values to calculate the z-scores, $z = (R(d) - mean)/std$, where $R(d)$ is the remainder value of day d, $mean$ is the mean remainder value for the last seven days, and std is the standard deviation of the remainders, with respect to each topic. If the z-score is higher than 2, events can be considered as abnormal within a 95% confidence interval. The calculated z-scores are thus used as abnormality rating and the retrieved topics will be ranked in the analytics environment according to this estimate.
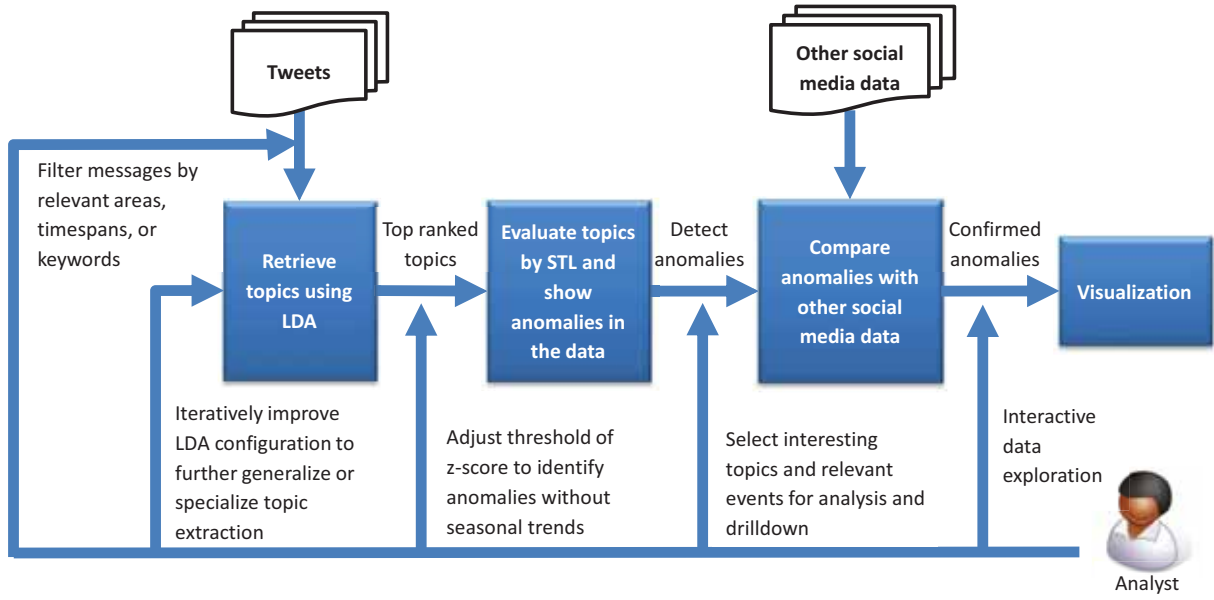
Figure 2: Overview of our iterative analysis scheme for event detection and examination.

3.3 Detection Model

To conclude this section, we formalize our abnormal event detection model based on the probabilistic topic extraction and time series decomposition.

An abnormal event is associated with a set of social media messages that provides its contents, location, and timestamp. To detect abnormal events for a given area and timespan, we define a set called *social spacetime* as follows:

$$S = (T, \Delta time, \Delta area, msgs) \qquad (2)$$

where T is a set of topics, $\Delta time$ is a time period (e.g., one day), $\Delta area$ is a bounded geographical region, and $msgs$ is a set of messages. The user selected parameters $\Delta area$ and $\Delta time$ define the analysis context for which all messages are loaded into the analysis system. In this context, the user selects a subset of messages ($msgs$) for which the LDA topic modeling procedure (described in Section 3.1) extracts the set of topics, $t_i \in T$. Each topic is defined as:

$$t_i = (M_i, W_i, z_i, Y_i, p_i) \qquad (3)$$

where W_i is a set of words describing the topic, M_i is a set of relevant messages, z_i is an abnormality score (z-score), Y_i is a time series, and p_i is a statistical proportion of the topic in $msgs$.

For each topic (t_i), our algorithm searches relevant messages (M_i) in the selected area ($\Delta area$) and time period ($\Delta time$) and a predefined time span of historic data preceding $\Delta time$ (e.g. one month). Messages are considered relevant if they contain at least one word in W_i. From M_i a daily message count time series (Y_i) is generated from the timestamps of the messages. The algorithm decomposes Y_i to obtain a remainder component series using the STL and calculates a z-score (z_i) from the remainder series. Lastly, it sorts the topics based on the z-scores.

For cross validation of each topic, we search for relevant entries in Flickr and YouTube by their meta-data that includes titles, descriptions, tags, and timestamps, using the respective APIs. We repeat the steps for generating a time

series from the collected timestamps, applying STL to decompose the time series, and calculating the z-score from the remainder component series.

4 Interactive Analysis Process

The complete topic extraction, abnormality estimation, and event examination are tightly integrated into a highly interactive visual analysis workbench, that allows an analyst to observe, supervise, and configure the method in each individual step. The following sections introduce the details of this system and describe how the event detection is embedded within a sophisticated analysis process as shown in Figure 2.

4.1 Social Media Retrieval and Analysis System

Our modular analysis workbench ScatterBlogs was already featured in previous works [2, 26]. It proved itself very useful for fundamental tasks like collection, exploration and examination of individual, as well as aggregated, social media messages. The UI of the system is composed of several interconnected views and the main view houses a zoomable openstreetmaps implementation showing message geolocations on a world map. The system features a text search engine and visual content selection tools that can be used to retrieve messages, show spatial and temporal distributions and display textual message contents. Additional visualizations and map overlays provide the analyst with powerful inspection tools, such as a kernel-density heatmap similar to [16], to show aggregated and normalized message distributions and a movable lens-like exploration tool (called 'content lens') that aggregates keyterm frequencies in selected map areas [2]. To indicate spatiotemporal anomalies in the message set, the system features a mechanism to detect spatiotemporal clusters of similar term usage, and suspicious message clusters can be represented as Tag Clouds on the map [26]. For the real-time collection of messages using the Twitter Streaming API the system features a scalable extraction and preprocessing component. This component was used to collect Twitter messages since August 2011 and it currently

147

processes up to 20 Million messages per day[5], including the almost complete volume of up to 4 million messages that come with precise geolocation information.

4.2 Visual Topic Exploration and Event Evaluation

Results from the topic retrieval and event detection as described in Section 3 can be iteratively refined by means of visual result presentation and interactive parameter steering. Both, the final result of event detection as well as intermediary findings during data filtering and topic extraction can be used by the analyst to adjust the process in order to identify interesting topics and keyterms as well as relevant map areas and timespans for a given analysis task. New insights can be generated on each of four individual analysis layers which, in conclusion form an iterative analysis loop from data filtering to result visualization:

- **Spatiotemporal Data Filtering:** The analyst selects an initial spatiotemporal context of Twitter messages to be represented in the visualization and to serve as a basis for analysis. He can do so by using textual as well as spatiotemporal query and filter mechanisms that load the relevant base message set from a larger database into active memory. The analyst can further filter the base set and remove unimportant parts by using a time-slider, depicting temporal message densities, or polygon and brush selection tools. Using these tools the analyst can gain an initial impression of the spatial and temporal distribution and location of messages that could be relevant for his analysis task.

- **LDA Topic Examination:** In the subsequent step the analyst can choose to start the topic extraction either on the whole analysis context or on some subset of selected messages. At this stage he can utilize the configuration parameters of LDA extraction to interactively explore available topics by generalization and specialization. In this regard the most important parameter is the number of topics that have to be defined for the topic model inference. If the analyst decreases the number using the provided tools, the extracted topics will be more general. If he increases it, they will be more specific and thus candidates for small but possible important events. Once topics are generated from the data they will be presented to the analyst through a list of small tag clouds for each topic. He can now select the topics from the list to see their individual message distribution on the map and the temporal distribution in the time-slider.

- **STL Evaluation:** Depending on the analyst's choice, the topics can be evaluated and ordered based either on absolute topic frequency or based on abnormality estimates that have been computed using STL. As described in Section 3.2, a valid estimate of abnormality depends on the computation of z-scores from data seven days prior to the observed time frame. Therefore, the STL evaluation will extend the data examination to a range prior to the selected spatiotemporal context, if data is available. Once abnormality is computed for each topic, the topic list will be ordered according to the values and the topics with most outstanding abnormality are highlighted.

- **Crosscheck Validation:** Each selection of messages is accompanied by charts showing the total time series and the remainder components for the selected message set using STL. This is true for spatiotemporal selections as well as for selections using the LDA topic list. In addition to the geolocated Twitter messages this STL is at the same time performed for data that has been extracted from supplemental services like Flickr and YouTube. Based on the multiple charts the analyst can crosscheck the importance and abnormality of examined events and topics.

In our system, the analyst is supposed to iteratively use these means of semi-automated processing, visualization and interaction to refine the selection of messages up to a point where he can begin to examine individual message details. For this task, he can then utilize tools like the content lens for small scale aggregation or the table view to read the messages textual content. The application of these tools is shown in Figure 3. Usually the most valuable messages will be reports from local eyewitnesses of an important event or from insiders for a given topic. Thus, to retrieve large quantities of such messages helping to understand an ongoing event or situation will be the final goal of the iterative process. Unusual topics, suspicious keyword distributions and events with high STL abnormality discovered on the repeatedly traversed analysis layers can guide the analysis from a very broad and general overview to very specific topics and a relatively small message set suitable for detailed examination.

5 Case Study

In this section, we present three case studies for our system covering different types of events including the Chardon High School Shooting, the Occupy Wall Street protests in New York, and the 2011 Virginia Earthquake. The first case shows how analysts can use our system efficiently to find and
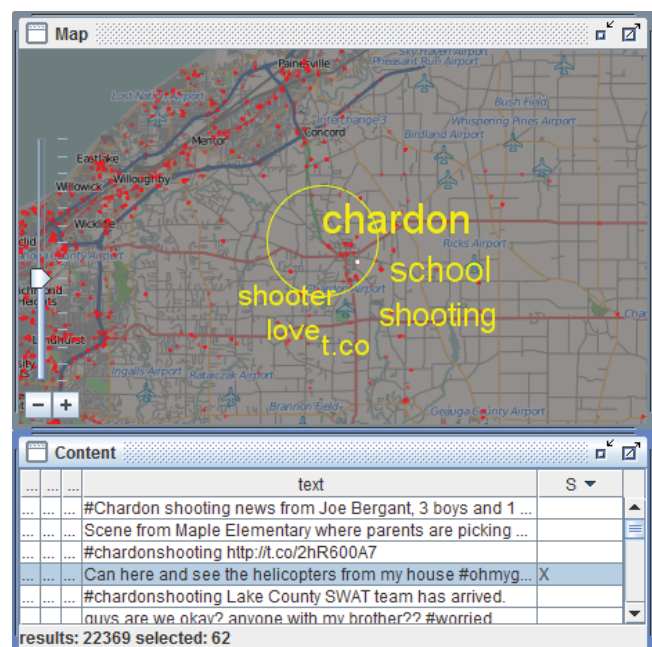

Figure 3: Examining the location of the Chardon high school shooting with a text aggregating content lens.

---

[5]This is just a 10% sample of the total 200 million Twitter messages due to API rate limitations.
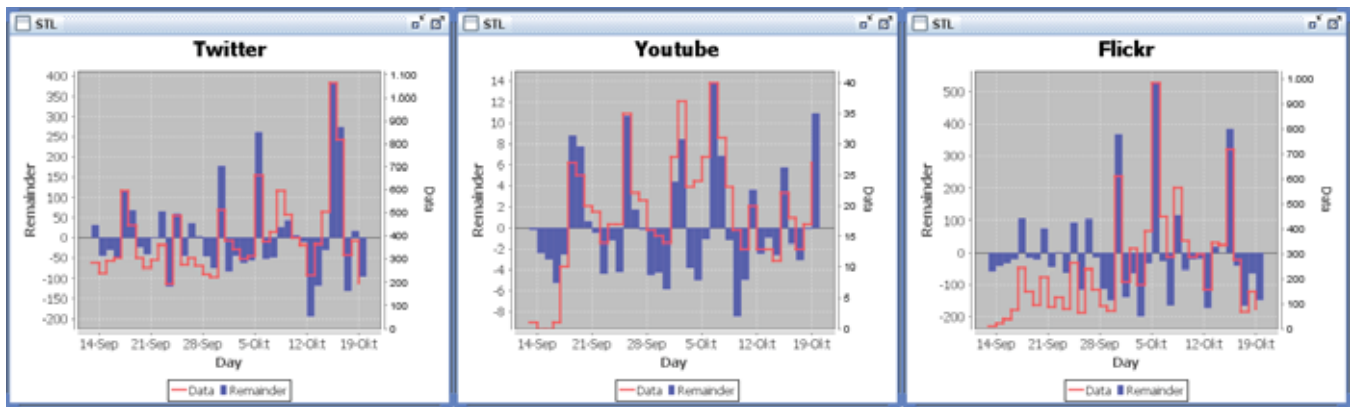
Figure 4: Cross validation of an event using Twitter, Flickr, and YouTube data for the Occupy Wall Street Protests. The protests occurred on Sep. 17 and 30, Oct. 5 and 15. The line charts show the remainder components R (blue) and the original data volumes Y (red) for the STL evaluation. The scales on the right and left side of each chart view are adapted to the maximum values.

explore an abnormal event. The second case highlights the differences between social media types by cross validation of a planned event. Finally, the last example showcases the effects of an abrupt, unexpected, natural disaster.

5.1 Ohio High School Shooting

On February 27, 2012, a student opened fire inside the Chardon High School cafeteria in the early morning. The gunman killed one student and injured four, from which two eventually died after the incident.

To examine this incident we first locate and select the broader Cleveland area on the map and select a time frame covering three days from February 26 to February 28. Using the text search engine and a wildcard query ('*') we can establish an exploration context showing all messages plotted on the map with their respective contents and meta data listed in a separate table view. First, we want to get a broad overview of the topics discussed in the region and thus we select all messages in the area and apply the LDA extraction tool to the current selection. In order to see the most general topics, we chose a low parameter value for the number of topics and a high iteration count to achieve good separation. At this level of semantic detail, the extracted topics indicate messages about the NBA all-star game (February 26 in Orlando) with keywords like *kobe, game, dunk and lebron* as well as the showing of the movie 'The Lion King'on TV with keywords *king, lion, tv*. If we look at the STL-Diagrams of these topics and the computed z-scores, we also see a peak for these events. By clicking on the retrieved topic representations the associated messages are highlighted in each view. By reading some of the message contents (e.g. *'Watching my fav. Movie on ABC family..... Lion King!!!!', 'Can't wait till the dunk contest starts!'*), the analyst can easily disqualify these from further analysis.

To get a higher semantic resolution we can now increase the number of topics and slightly decrease the iteration count in order to achieve a fast computation. By selecting 20 topics, the topic indicating the shooting event is extracted and indicated by keyterms like *shooting, chardon and school*, alongside the other topics. Although the proportion of the topic is not very high compared to the others, the topic receives a very high z-score (i.e., 3.77) and is ranked among the top five topics (highlighted in orange). Figure 1 demonstrates the system view of this observation. An analyst can now select the incident topic to see the spatial distribution

of associated messages on the maps as well as the temporal distribution in the timeslider histogram. By examining messages using the content lens to aggregate topics over map areas as well as the tools for reading individual message contents, we can easily distinguish between messages informed by media reaction and messages of actual observers in the Chardon High School area. In this case, after isolating the messages from local observers, we find messages like *'Omg shooting at Chardon High School?!?!'* and *'Helicopter overhead. We are on scene. Message from school says students moved to middle school'*.

5.2 Occupy Wall Sreet

Starting on September 17, 2011 in the Wall Street financial district in New York City, people have been gathering for the Occupy Wall Street protest movement. The movement against economic inequality has since spread to other major cities throughout the world. Various social media services including Twitter, Facebook, Flickr and Youtube have been utilized both by the participants and the global media for communication and reports about the movement in forms
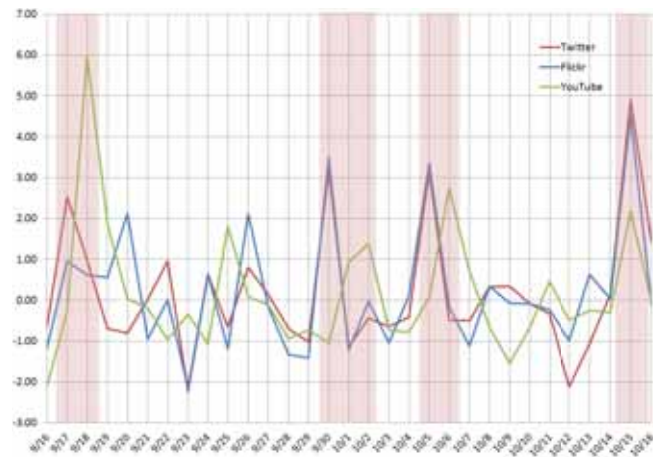

Figure 5: Abnormality and correlation on multiple social media sources. As a result of high z-scores around the same time periods, we found a strong correlation between the three social media sources. Marked regions correspond to periods where at least 2 providers received scores over 2.0.
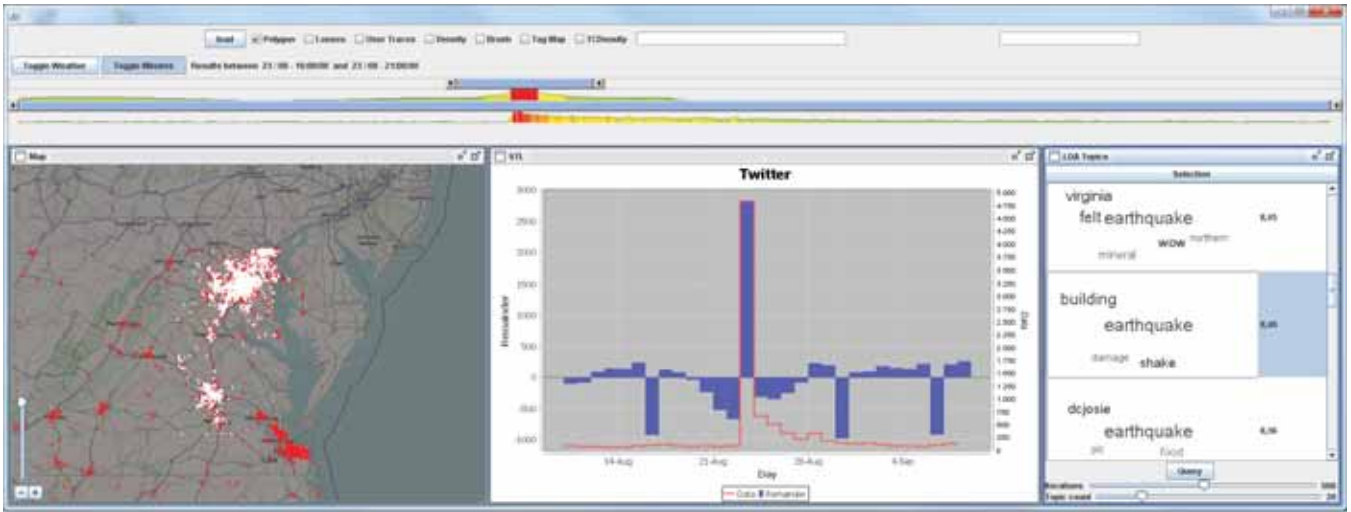
Figure 6: Virginia earthquake on August 23rd, 2011. Our abnormal event detection system detects the earthquake event using our STL based anomaly detection algorithm. The abnormality degree is extremely high on August 23rd, 2011 (times are given in UTC).

of text, images and videos. For the related extracted topic (*occupywallstreet, wall, takewallstreet, takewallst, park*), Figure 4 shows the results of our abnormality estimation for the three social media services Twitter, Flickr, and YouTube over the course of one month. As shown in Figure 5, in each of the marked regions, at least two of the services show z-scores over 2.0 and they correspond to actual events during the Occupy Wall Street protests. From this experimental result, one can derive a strong correlation between the three social media data sources. The related data volumes and remainder (R) are shown in Figure 4 for all three providers.

As shown in Figure 5, on September 17 (the first day of the protests with approximately 1,000 participants [30]), only the Twitter stream received an abnormal score while the Flickr and YouTube data artifacts are delayed by 1-3 days. We attribute this initial delay to the simple nature of Twitter usage compared to Flickr and YouTube where the data potentially has to be recorded, edited, and uploaded and is thus more labor intensive. Additionally, eighty protesters where arrested while marching uptown on September 24, but even though Flickr and YouTube reaction on this event created higher z-scores in the following days, they were not significant enough to register an event. The following spikes of high z-scores overlap with a march across the Brooklyn Bridge (Oct. 1 [29]), a large demonstration (Oct. 5 [27]), and globally coordinated protests (Oct. 15 [28]).

5.3 2011 Virginia Earthquake

For the last use case we examine a magnitude 5.8 earthquake that occurred on the afternoon of August 23rd 2011 in Mineral, Virginia [32]. Starting with the minute of the earthquakes occurrence, Twitter users posted more than 40.000 earthquake-related Tweets reporting tremors they felt along the East Coast [11]. Among these were messages like: *'EARTHQUAKE!!!!!!!'*; *'Whoa!!!! Just experienced an earthquake here in Virginia!!!!'*; and *'Omg I just felt an earthquake'*. Figure 6 gives an impression how our system is applied to examine this event.

For the analysis we begin with selecting the Virginia area from Baltimore to Virginia Beach and three days around the 23th. A topic extraction with 5 topics and just 100 iterations already retrieves two earthquake related topics showing

that this event is very prominent within the selection. By clicking these topics one can observe that the highest density of earthquake messages can be found in the Washington, Baltimore and Richmond areas.

To observe the areas in more detail we combine the topic selection with a spatial selection of the three cities and reapply the topic extraction. This time we use 20 topics with 500 iterations. Since we are now operating only on earthquake related messages, the retrieved topics all contain earthquake as a dominant keyword. On this level of detail we can see topics indicating that buildings have been evacuated due to the earthquake (*earthquake, people, evacuated, early, building*) and that damage has been caused (*earthquake, building, shake, damage*). The z-scores for all top ranked topics are now very high (often above 8.0) and thus indicate the high abnormality of this event.

Finally, when going into even higher detail with 100 topics and 1000 iterations we can see smaller events within the big earthquake event. For example, one topic indicates that damage was caused to the Washington Monument and by clicking on the topic we can see messages like *'damage to Washington Monument'*; *'Washington Monument is tilting?!?'*; and *'Helicopter just landed next to Washington Moniment, west side. #DCearthquake'*. There are also misleading messages, indicating that the damage to the Washington Monument was just false rumors: *'the Washington monument was not damaged in any way from the earthquake. #rumor'*. However, media crosschecks show that visible damages did in fact happen and will probably cost the city 15 million dollars to repair[6].

At this point, it is important to note, that while several earthquake topics produced significant z-values in Twitter, the event did not produce high z-scores in Flickr and YouTube. This is probably due to fact, that many people will write a quick message after a shock has been felt by themselves, but it takes quite some time until images or videos are uploaded from cameras to Flickr and YouTube. The event also demonstrates that large and unexpected events will produce immediate and significant reactions in

---

[6]http://www.huffingtonpost.com/2012/03/14/washington-monument-did-e_n_1344422.html

services like Twitter and they can thus easily be detected by using our system.

6 Discussion

In this section we want to discuss four important notes and observations relevant to the presented approach.

Event Types: As was demonstrated with the three case studies, events in social media can be categorized into two different types. The 2011 Virginia Earthquake and the Ohio High School Shooting can be categorized as abrupt or disaster events, while Occupy Wall Street can be considered a social and planned event. The two types of events have quite distinguishable features. For the abrupt events, there is a strong change in daily counts mainly in the text based Twitter messages. For the planned event, the Twitter signal may still be faster, but due to the gradual increase and decrease, it is less pronounced. In contrast, Flickr and YouTube have delayed, but very prominent changes, for planned events; however, we could not find significant signals for abrupt events. This reflects that video and photo recording happen rarely during abrupt events. Social events, e.g., Occupy Wall Street or election debates, however, have a high impact on such multimedia based social media; Relevant videos, photos, and even meta-data (e.g., descriptions, tags) allow analysts to find additional information about them. We, therefore, think that cross validating events among multiple social media types is important in order to establish situational awareness.

Base Data: Regarding the base data, it is important to note, that our approach depends on geo-located Twitter messages with precise coordinates, which are only a fraction of the whole Twitter stream. While this fraction still consists of several million messages per day, it is not a representative sample of the population, because it mainly covers mobile users equipped with GPS enabled devices. We think, however, that mobile users, who share their daily experiences freely, are the most relevant group for situational awareness scenarios. Some studies [4, 17] tried to overcome the problem of location information scarcity in Twitter messages, which adds another source of uncertainty. First, the user's self reported locations can be outdated. Second, the geo-coding of the location can be considerably wrong due to place name ambiguities. Furthermore, we have just shown the feasibility of the approach for Twitter, Flickr, and YouTube data, but it can easily be adapted to other social media providers like Facebook or Forsquare as well, in order to widen the sample of the population.

Probabilistic Models: In this work, we use STL to decompose time series of topic streams. There are many alternative statistical models for this task, such as DHR (Dynamic Harmonics Regression) [38] and SARIMA (Seasonal AutoRegressive Intergrated Moving Average) [3]. DHR and SARIMA models are particularly useful for forecasting and STL can also be used for prediction based on seasonal (periodic) time series [12]. Our main reasons for choosing STL was the fact that it is non-parametric, can be computed faster than SARIMA [12] and needs less training data for equally good results.

End User Feedback: We requested informal feedback from users within our institutes and received comments and suggestions. To compare the LDA topic modeling plus the seasonal-decomposition based abnormality analysis versus only the LDA topic modeling, we enabled our system to switch between these modes. The users were impressed by the fact that both results (two lists of topics) from two different modes were quite different. Highly ranked topics by LDA

topic modeling consisted of ordinary words, while the combined analysis was indicating unusual events. They noted that the tightly integrated visual analysis workbench was useful to apply the automated methods. Furthermore, they suggested a function allowing people to see a pattern of abnormality for a user-defined topic.

7 Conclusion

In this paper, we presented an interactive abnormal event detection and examination system for the analysis of multiple social media data sources. The system uses an abnormality estimation scheme based on probabilistic topic modeling and seasonal-trend decomposition to find and examine relevant message subsets. This scheme is tightly integrated into an highly interactive visual analytics system, which supplements tools based on automated message evaluation with sophisticated means for parameter steering, filtering and aggregated result set exploration. Three use cases demonstrated the visualization and user interaction within the system and its capabilities to detect and examine several different event types from social media data. The ability to crosscheck findings based on three distinct social media sources revealed the kinds of correlations that can be expected from various event types.

For future work, we will further investigate context-based analysis and improve the current detection algorithm to allow for a faster analysis. Due to the fast-paced and low quality nature of microblogging, we will also investigate the effects of additional preprocessing options like automated spell-checking or synonym recognition under the constraint of preventing ambiguities. Furthermore, we want to supplement the system with real-time monitoring features, demanding additional means for adaptive attention guiding as well as interaction techniques for use in high pressure environments. For the final system we are currently preparing a thorough evaluation to test it in cooperation with crisis management personnel and other domain experts.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[2] H. Bosch, D. Thom, M. Worner, S. Koch, E. Puttmann, D. Jackle, and T. Ertl. Scatterblogs: Geo-spatial document analysis. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 309 –310, 2011.

[3] G. E. P. Box and G. Jenkins. *Time Series Analysis, Forecasting and Control*. Holden-Day, Incorporated, 1990.

[4] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 759–768, New York, NY, USA, 2010. ACM.

[5] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning. Stl: A seasonal-trend decomposition procedure based on loess (with discussion). *Journal of Official Statistics*, 6:3–73, 1990.

[6] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

[7] M. Dörk, S. Carpendale, C. Collins, and C. Williamson. VisGets: Coordinated visualizations for web-based information exploration and discovery. *IEEE Transactions on Visualization and Computer Graphics (Proceedings Information Visualization 2008)*, 14(6):1205–1212, 2008.

[8] W. Dou, X. Wang, R. Chang, and W. Ribarsky. Paralleltopics: A probabilistic approach to exploring document collections. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 231 –240, oct. 2011.

[9] K. Field and J. O'Brien. Cartoblography: Experiments in using and organising the spatial context of micro-blogging. *Transactions in GIS*, 14:5–23, 2010.

[10] T. Griffiths and M. Steyvers. Finding scientific topics. In *Proceedings of the National Academy of Sciences*, volume 101, pages 5228–5235, 2004.

[11] L. Indvik. East coasters turn to twitter during virginia earthquake. Retrieved March 30, 2012, `http://mashable.com/2011/08/23/virginia-earthquake/`, 2011.

[12] B. Jiang, S. Liang, J. Wang, and Z. Xiao. Modeling modis lai time series using three statistical methods. *Remote Sensing of Environment*, 114(7):1432–1444, 2010.

[13] R. Krovetz. Viewing morphology as an inference process. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, pages 191–202, New York, NY, USA, 1993. ACM.

[14] R. Lee and K. Sumiya. Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, LBSN '10, pages 1–10, New York, NY, USA, 2010. ACM.

[15] A. MacEachren, A. Jaiswal, A. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford. Senseplace2: Geotwitter analytics support for situational awareness. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 181 –190, oct. 2011.

[16] R. Maciejewski, S. Rudolph, R. Hafen, A. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis, and D. S. Ebert. A visual analytics approach to understanding spatiotemporal hotspots. *IEEE Transactions on Visualization and Computer Graphics*, 16(2):205–220, Mar. 2010.

[17] J. Mahmud, J. Nichols, and C. Drews. Where is this tweet from? Inferring home locations of twitter users. In *International AAAI Conference on Weblogs and Social Media*, 2012.

[18] A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[19] A. Pozdnoukhov and C. Kaiser. Space-time dynamics of topics in streaming text. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, LBSN '11, pages 1–8, New York, NY, USA, 2011. ACM.

[20] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In *ICWSM*, 2010.

[21] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[22] E. Roth and J. White. Twitterhitter: Geovisual analytics for harvesting insight from volunteered geographic information. In *Proceedings of GIScience*, 2010.

[23] T. Sakaki, M. Okazaki, and Y. Matsuo. 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, New York, NY, USA, 2010. ACM.

[24] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336 –343, sep 1996.

[25] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 306–315, New York, NY, USA, 2004. ACM.

[26] D. Thom, H. Bosch, S. Koch, M. Woerner, and T. Ertl. Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages. In *IEEE Pacific Visualization Symposium (PacificVis)*, 2012.

[27] N. Times. Major unions join occupy wall street protest. Retrieved June 25, 2012, `http://www.nytimes.com/2011/10/06/nyregion/major-unions-join-occupy-wall-street-protest.html`, 2011.

[28] N. Times. Occupy wall street protests worldwide. Retrieved June 25, 2012, `http://www.nytimes.com/2011/10/16/world/occupy-wall-street-protests-worldwide.html`, 2011.

[29] N. Times. Police arresting protesters on brooklyn bridge. Retrieved June 25, 2012, `http://cityroom.blogs.nytimes.com/2011/10/01/police-arresting-protesters-on-brooklyn-bridge`, 2011.

[30] N. Times. Wall street protest begins with demonstrators blocked. Retrieved June 25, 2012, `http://cityroom.blogs.nytimes.com/2011/09/17/wall-street-protest-begins-with-demonstrators-blocked`, 2011.

[31] Twitter. 200 million tweets per day. Retrieved March 1, 2012, `http://blog.twitter.com/2011/06/200-million-tweets-per-day.html`, 2011.

[32] United States Geological Survey (USGS). Magnitude 5.8 - virginia. Retrieved March 30, 2012, `http://earthquake.usgs.gov/earthquakes/recenteqsww/Quakes/se082311a.php`, 2011.

[33] S. Wakamiya, R. Lee, and K. Sumiya. Crowd-based urban characterization: extracting crowd behavioral patterns in urban areas from twitter. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, LBSN '11, pages 77–84, New York, NY, USA, 2011. ACM.

[34] J. Weng and B.-S. Lee. Event detection in twitter. In *International AAAI Conference on Weblogs and Social Media*, 2011.

[35] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 261–270, New York, NY, USA, 2010. ACM.

[36] P. C. Wong, B. Hetzler, C. Posse, M. Whiting, S. Havre, N. Cramer, A. Shah, M. Singhal, A. Turner, and J. Thomas. IN-SPIRE Infovis 2004 contest entry. In *IEEE Symposium on Information Visualization*, Oct. 2004.

[37] J. J.-C. Ying, W.-C. Lee, M. Ye, C.-Y. Chen, and V. S. Tseng. User association analysis of locales on location based social networks. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, LBSN '11, pages 69–76, New York, NY, USA, 2011. ACM.

[38] P. C. Young, D. J. Pedregal, and W. Tych. Dynamic harmonic regression. *Journal of Forecasting*, 18(6):369–394, 1999.

[39] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European conference on Advances in information retrieval*, ECIR'11, pages 338–349. Springer-Verlag, Berlin, Heidelberg, 2011.