Less After-the-Fact: Investigative Visual Analysis of Events from Streaming Twitter

Thomas Kraft, Derek Xiaoyu Wang, Jeffrey Delawder, Wenwen Dou, Li Yu, and William Ribarsky

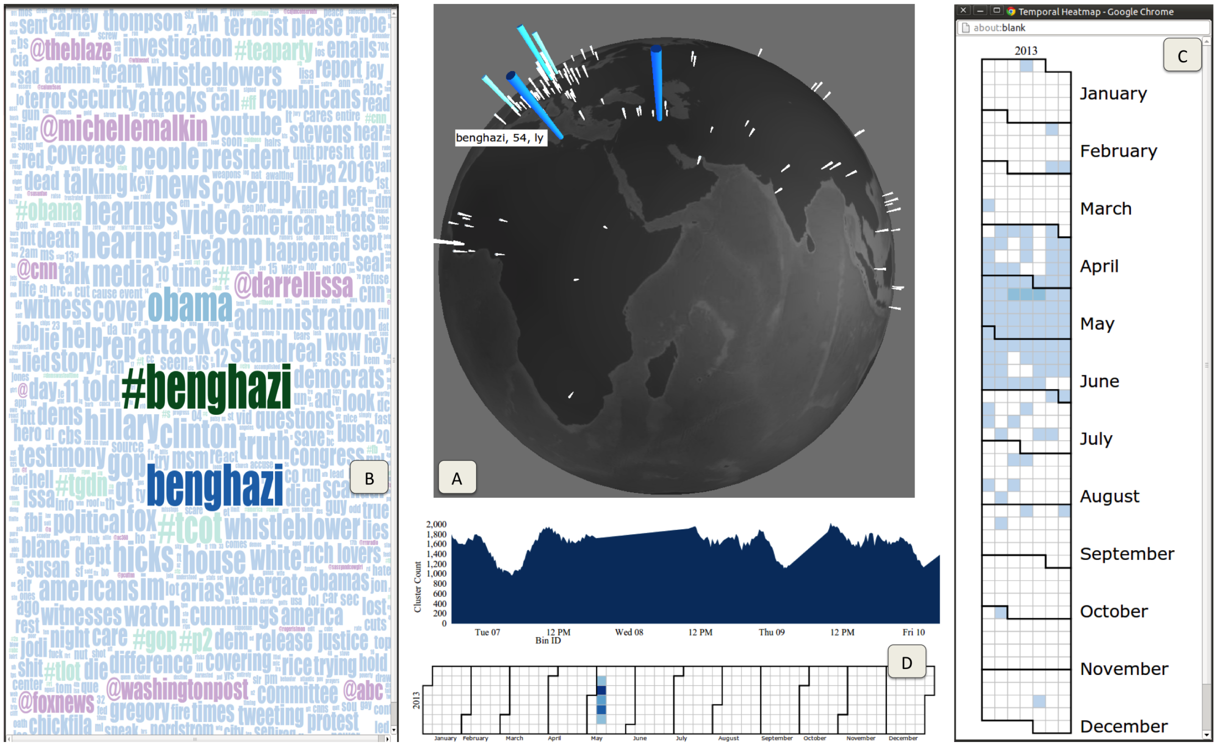

Fig. 1. Overview of the interface in our Geo and Temporal Association Creator. GTAC utilizes a set of interactive visualizations to engage users in an investigative event analysis environment. Specifically, GTAC allows users to depict events from multiple event facets, including (A) Geospatial view (*where*), (B) Future temporal heat map view (*when*), (C) on-going master timeline view (*when*), and (D) contextual word cloud view (*who and what*).

Abstract—
News and events are traditionally broadcasted in an "After-the-Fact" manner, where the masses react to news formulated by a group of professionals. However, the deluge of information and real-time online social media sites have significantly changed this information input-output cycle, allowing the masses to report real-time events around the world. Specifically, the use of Twitter has resulted in the creation of a digital wealth of knowledge that directly associates to such events. Although governments and industries acknowledge the value of extracting events from the TwitterSphere, unfortunately the sheer velocity and volume of tweets poses significant challenges to the desired event analysis. In this paper, we present our Geo and Temporal Association Creator (GTAC) which extracts structured representations of events from the Twitter stream. GTAC further supports event-level investigative analysis of social media data through interactively visualizing the event indicators (who, when, where, and what). Using GTAC, we are trying to create a near real-time analysis environment for analysts to identify event structures, geographical distributions, and key indicators of emerging events.

Index Terms—Radiosity, global illumination, constant time

1 Introduction

Events and news stories are traditionally broadcasted in an *"After-the-Fact"* fashion, where the audience are reacting to the news formulated by a group of professionals. Journalists will be the first responders in acquiring information, investigating fundamentals of a story (*"who, where, when, what"*), and structuring the news with their view points. Audience's participation in this process, given the limited access to contextual inputs, is rather limited. The audience for a long time was in a position where they didn't have effective ways of influencing the news and were disconnected from the event reporting cycle.

Now with the deluge of information and real-time online social media sites, however, the news input-output cycle has changed. At the

- *Thomas Kraft is with UNC Charlotte. E-mail: tkraft3@uncc.edu.*
- *Derek Xiaoyu Wang is with UNC Charlotte. E-mail: xiaoyu.wang@uncc.edu.*

Manuscript received 31 March 2013; accepted 1 August 2013; posted online 13 October 2013; mailed on 27 September 2013.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

input end, a simple status update from the masses could lead to the report of a real-time event and help to formulate the story. Audiences are more in the drivers' seat in terms of how information is getting to journalists and what's happening around the world. For instance, when Costa Rica was hit by a 7.6 magnitude earthquake on Sep. 5, 2012, it only took 30 seconds after the epicenter was hit for the first message to appear globally on Twitter [21]. Everyone on the Twitter-Sphere, theoretically, had the potential of knowing that a earthquake had just happened in Costa Rica.

The constant output of such activity has evidently resulted in the creation of a digital wealth of knowledge that is directly associated with real-life events from around the globe. It is generally agreed by both government and industries that organizations can benefit from being able to see events in a more organically structured way [14]. Knowing such information can help these enterprises make more informed analytical decisions, specifically the *"who, where, when, and what"* elements connected to each discussed topic. For emergency response agencies, just to give an example, sifting through massive amounts of social media data could help them monitor and track the development of and the response to natural disasters, as illustrated in the aforementioned example.

While Twitter messages present a rich source of useful information, they are very disorganized and noisy, motivating the need for automatic event extraction, association, aggregation and categorization. On the one hand, the short and unstructured nature of tweets makes it easy to share information, but then makes interpreting semantic information difficult. Given the unique structure of a tweet (i.e. short, heavy with collegial language, limited on context), not all properties of an event may be expressed in a single message. Many of the tweets are self-contained and are therefore not composed of complex discourse structure as is the case for texts containing narratives (e.g., newswire).

On the other hand, associating useful event indicators from the online, disorganized, noisy text is another challenging problem. The event discourse between multiple tweets and canonical event fundamentals is inconsistent and sometimes conflicting, with few accurate methods of associating them. Even with the hashtags, created for the purpose of structurally initiating and propagating topics, the versatility of such meta-tags from the masses have not made the discourse coherent. The intrinsically polylingual, fragmented, and dynamic nature of hashtags is also a disadvantage in eliciting valuable information. Users can be overwhelmed with the noise of unrelated messages and conflicting information.

We consider it feasible to use social media to discover new events missed by curation, but mentioned or reported online by the masses. Our goal is therefore to induce a comprehensive event structure with a real-time visual analytics environment so that decision makers can be in the loop to investigate such events. Specifically, we aim to address the following three challenges:

- Identifying emerging events based on location and time
- Real-time event association from large-scale streaming social media data
- Human involvement in depicting and validating emerging and future events

To this aim, we have developed an analysis process that extracts structured representations of events from Twitter streaming, and supports event-level investigative analysis of the social media data. Our Geo and Temporal Association Creator (GTAC) is an analytic pipeline connecting the twitter stream to various information extraction and clustering techniques. GTAC is centered on the combination of data-driven event extraction approaches with human-centered visual analytics techniques; extraction of the event indicators (*who, when, where, and what*) is enhanced by interactive visual interfaces, providing results that can be explored, filtered, and managed by users. The resulting interface creates a real-time analysis environment for identifying events, geographical distributions, and key indicators of emerging events.

In the following sections of this paper, we will first specify the challenges in achieving our analysis process (Section 2). We then focus on describing the implementation details of our system(Section 3). Next, we will provide detailed case studies of this system in action (Section 4), and conclude our paper by discussing the strengths and weaknesses of our implementation.

2 Related Work and Domain Characterization

Event detection and structuring from texts containing narratives has been a well studied area [15, 29, 31, 32, 8], where an *Event* is commonly considered as an occurrence at a specific time and place. Recent research has demonstrated that one of the common uses of social media is reporting and discussing events users are experiencing: Sakaki et al. [24] showed that mining of relevant tweets can be used to detect earthquake events and predict the earthquake center in real-time. Becker et al. [3] proposed to identify real-world events through exploring a variety of techniques for learning multi-feature similarity metrics for social media documents. While both of these analyses were performed *retrospectively*, their evaluation results showed that events could be effectively detected from large-scale messages provided by the social media.

Online and streaming event detection and structuring is an emerging research trend. When a document comes in, the Online NED system compares it with all previous events and computes a pair-wise similarity score in real-time. During this process, single-pass clustering is widely used to process incoming news stories one-by-one to determine whether a new event has occurred [30]. To detect new events from a stream of Twitter posts, Petrovic et al. [19] presented an algorithm based on locality-sensitive hashing to deal with the large number of tweets generated every second. Online NED is shown to be powerful in detecting events in newswire (less update frequency), however, when applied to microblogs (e.g. Tweets), this process faces additional challenges due to the frequent update of the massive amount of fragmented documents. The velocity and volume of unstructured data makes extracting meaningful event information challenging. Specifically, such problems include a much higher volume of data, as well as noise.

2.1 Signal in a haystack: extracting scarce geo-temporal indicators from messy tweets

Tweets are commonly messy, short and incomplete, heavy with collegial language and very limited on contextual information [22]. Therefore, not all properties (*who, when, where, what*) of an event may be expressed in a single message. This poses a significant challenge when trying to associate geospatial and temporal information. Such information is valuable in providing context anchoring the tweet to a given time and space, and helping depict the overall discourse of an event, as shown in [9].

Many tweets are self-contained and are therefore not composed of complex event discourse, in comparison to texts containing narratives (e.g., newswire). Dates and Locations are not always mentioned together, and ambiguous information is passed back and forth making it complicated to isolate the correct time and date of a planned event. As shown in Dou et al's study [10], the geo-temporal information of the Occupy WallStreet movement, is distributed in multiple tweet threads and needs to be carefully associated to provide a more complete picture.

Exacerbating the challenge is our need for detecting and inferring future events from streaming tweets, identifying the geospatial and temporal signals related to future events. Rather than merely capturing the meta-information such as GPS or timestamp of a tweet post, which only informs us about possible ongoing events, our focused geospatial-temporal signals needs to come directly from the content of the tweets. In this regard, we need to leverage techniques from NLP [5], KDD [20] and GIR [2] communities.

Our developed GTAC provides an automated process of analyzing text, finding geographic and temporal references, and combining these references into meaningful semantic summaries (i.e. geo-temporal scopes for the tweets). For GTAC, event signals containing temporal references after the present date are labeled to be potential future events. GTAC relies on the aggregation of massive inputs from the TwitterSphere to promote and filter significant event signals; such aggregation methods will generate a similarity score that will be further used to classify new event indicators. If the score falls below a certain threshold, GTAC will mark the document as a new event; otherwise the document is labeled as old and merged into the prior corresponding topics. In the resulting geo-temporal association, strong and repetitive geo-temporal signals are pushed forward, whereas signals on a much smaller scale are only compiled into a possible event list. GTAC assesses all the available twitter content to extract the most commonly referenced times, dates, and locations.

2.2 Real-time event association from large-scale streaming social media data

Real-time event detection is another goal of our designed system. Our primary goal is to create a system that can alert users when and where events are on-going or will occur, using the content streamed from Twitter. While events can be planned using Twitter, the amplitude of the signal prior to the event could be significantly smaller than the signal produced once the event has begun. For this reason, large events sometimes have relatively small precursors. Using The Occupy Wall Street movement case study as an example [10], it is illustrated that this movement evolved over the course of a month prior to the official start date of the protest, growing more interests on Twitter as the start date approached.

Retrospectively depicting what has happened and recognizing past events on twitter, within an established timeframe, is certainly of great value. The challenging task, however, is detecting future events and catching the next influential movement before it happens; one not only has to isolate the the event signals, but also identify them in real-time.

In order to fully understand the emerging events on twitter, we need to collect as much information as possible to have the broadest range of data. There are a variety of social media tools that provide awareness to streaming and current twitter feeds [26, 17, 4] However, the majority of the work has been done on targeted streams (e.g, pre-filtered to specific keywords) but not on a broader analysis scope, limiting the possibility of finding new and pertinent information. Yet, it is challenging to find new events in a larger scale because the vagueness of tweets leads to fewer comparable features; thus grouping related messages, let alone associating metadata, becomes nontrivial.

Our GTAC is therefore developed to collect, analyze and organize event information using the 1% public sample stream from Twitter. [1] Currently, GTAC incrementally sifts through 4.5 million tweets daily (approx. 190,000/hr or 3,000/min) using a binning strategy to compare and associate event information. Based on a multi-core architecture, GTAC parallelizes the processes of event indicator extraction, semantic association, and content comparisons. It further achieves a client-server architecture that enables the real-time analysis of twitter events for domain users.

2.3 Human involvement in depicting and validating emerging and future events

Once a new event (e.g., an emergency or humanitarian crisis) has been detected, the logical next step is for people to track the development of such an event. While GTAC can provide structural representation of an event, it is ultimately domain users who will make a decision on the validity and usefulness of that information. Therefore, it is important to design a visual analytics interface that involves and engages end-users.

Studying how hashtags (represents ideas and sometimes events) spread within a Twitter user network, Romero et al. [23] found significant variation in the ways that widely used hashtags on differing topics spread. In addition, Marcus et al. [17] proposed a system-TwitInfo-for visualizing and summarizing events on Twitter. TwitInfo allows users to browse a large collection of tweets using a timeline-based display that highlights peaks of high tweet activity. Users can further drill down to subevents, and explore via geolocation, sentiment, and popular URLs.

To better facilitate the human involvement in this process, GTAC combines the real-time event extraction results with an interactive visual analytics interface. Instead of requiring users to specify what event they want to explore, as seen in TwitInfo, GTAC is assessing a broader scope of events using the automated event extractions. As shown in Figure 1, it further visualizes extracted events using a set of interactive visualizations to represent the geographic and temporal patterns being mentioned on twitter in near real-time. The resulting interface creates an effective analysis environment for users to interactively depict the key indicators of emerging events.

3 System Architecture and Implementation

To address the aforementioned three challenges, we designed and developed the Geo and Temporal Association Creator (GTAC) to facilitate the real-time analysis of the social media data. As shown in Figure 2, there are three main analysis components in GTAC. First, it provides an analytics pipeline that can sift through messy live twitter streams for event indicators. Secondly, it formulates events by utilizing graph algorithms to semantically associate and cluster multiple event indicators together. Finally, it keeps the human decision maker in the analysis loop through multiple coordinated geospatial-temporal visualizations. Utilizing analytic output, the visualization system provides interactive coordinated views of temporal and geospatial heatmaps, alongside word clouds, to enable the near real-time analysis of events on Twitter.

In the following sections, we will detail the algorithms and processes used in event analysis, alongside the visual designs that are tailored for identification and analysis of the events.

3.1 Architecture for Collecting and Analyzing Streaming Twitter Data

A key technical contribution of this paper is the near real-time analysis of events on Twitter. It is achieved based on a multi-core processing architecture, where GTAC parallelizes the processes of event indicator extraction, semantic association, and content comparisons.

On the data level, it is obvious that the speed at which the data storage operates is an important factor because both the analytical level and visualization level of our application need to access the database. Working with Twitter Streaming data poses a problem for traditional RDBMS systems due to the unstructured information as well as the constant growth of the database. Therefore, we have developed a parallel data crawler to interface with the Twitter stream to constantly collecting tweets from the Garden-hose. Instead of using relational database management systems, GTAC leverages Mongo [1], a NoSQL data storage solution, to effectively store and retrieve extracted event analysis results across a distributed system. Both the original tweets and its associated event information are stored in the NoSQL data repository and further accessed by our interactive visualizations.

On the analysis level, GTAC utilizes a parallel binning strategy to identify meaningful event information. Tweets are streamed and processed one at a time, and then saved in *bins*, which holds a subset of tweets for a certain interval of time (typically 5 minutes). As the event information grows, binning is a useful preprocessing procedure to classify such constantly evolving information accumulating in a continuous manner. Specifically, binning tweets into chunks allows our event analysis to be performed on distributed systems, thus making it possible to scale up the process and enable concurrent analyses. Information fragments are delegated to multiple processors and analyzed in a networked computing environment. As shown in Figure 2, GTAC utilizes binning strategy to organize tweets, creating snapshots illustrating what was happening at certain points of time. It further performs analyses between bins to identify the evolution of event information over time.

[1] As shown in [18], this 1% sample stream is a good representation of the entirety of tweets.

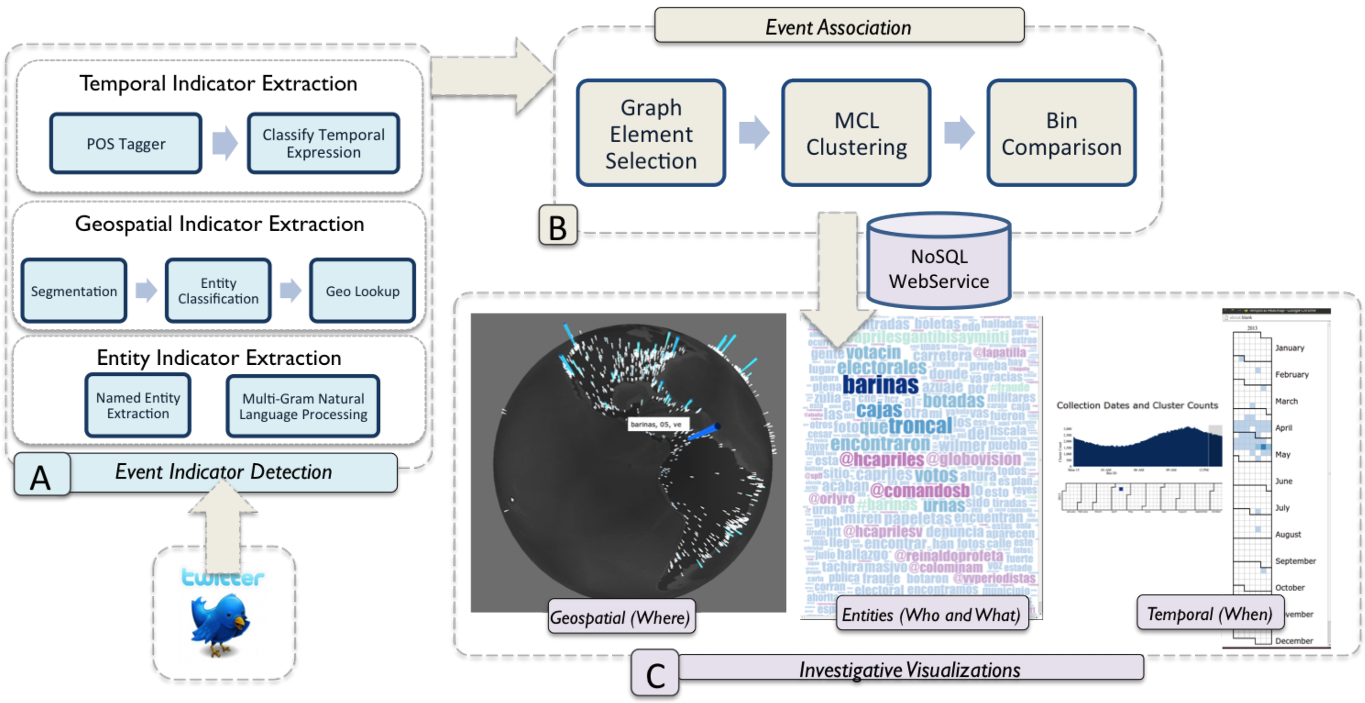

Fig. 2. System Architecture of GTAC. Starting from bottom left, Twitter data is first received and stored into our distributed storage system. The data then goes through an Event Indicator Detection stage (B) before entering the Event Association Stage (C). These two stages are computed online as the tweets streaming in. The extracted event indicators are further pipelined to the visualizations, where the information is then visualized in the interactive visual interface (C) for users to analyze the geospatial and temporal trends in an investigative environment to derive understanding of the on-going and future events.

3.2 Extraction of Event Indicators

To identify the event signal, GTAC relies on Natural Language Processing [12] and Named Entity Recognition [16] methods to classify and extract event indicators. Specifically, GTAC is designed to automatically extract key information on temporal indicators (when), geospatial entities (where), people or organization information (who), and general terms (what), from the text.

3.2.1 Temporal Extraction

To provide less “After-the-Fact” events and monitor events in real-time, our primary goal is to extract temporal information regarding to an on-going or future time frame. While all tweets contain timestamps (i.e., at the moment when the tweet is posted), this data only provides arbitrary information relevant to the past timeframe; thereby it is less informative in providing a sense of what’s happening now and what’s going to happen next. Our primary source for temporal extraction is, therefore, the actual content of the tweets.

However, extracting such desired time information is a non-trivial task. Such a process is challenging in two ways. On the one hand, dates and time on Twitter can be referred in a versatile manner given the flexibility in content creations. For example, the starting data of Occupy Wallstreet Movement was referred in various forms like #sep17, #917 and sep 17. While standard Regular Expression can be applied here to narrow down the terms, the creation of proper expression will also be lagging behind the real-time streams. On the other hand, temporal information can be difficult to interpret without proper context. For example, depending on the relative timeframes, the date of May 15th could be referred to a time that happened in the past (e.g., last week/month/year), or current (e.g., today), or future time (e.g., next Wednesday the 15th), etc. For this reason, we turned to TARSQI, which implements NLP techniques, to find and classify these temporal expressions [28]. Using TARSQI we were able to extract absolute times, relative times, and lexical triggers to anchor the contextual references documented timestamp.

While TARSQI is effective at picking up temporal indicators in the content of tweets, the speed of the classification process was not able to handle the stream of incoming data in near real-time. Like other NLP methods, the speed of the classification was mainly limited due to the Part of Speech tagger need for proper classification.

To overcome this issue we distributed the Temporal Extraction process using a producer/consumer model. Using this model we are able to parallelize the TARSQI process in a high-performance computing environment. We created a pool of background workers, each containing a single TARSQI instance and pulling tweets one at a time looking for the temporal expressions. As indicators are found, the workers pass back the tweet id, along with other data extracted back to the main thread which then puts the information inside our NoSQL database. Our parallelization of the TARSQI algorithm makes the overall temporal extraction process scalable; as the velocity of data increases, we can simply add more workers to handle the data increase.

3.2.2 Geospatial Extraction

As it’s difficult to extract temporal data from the text due to the ambiguous nature of tweets, it is just as challenging to identify geospatial indicators. While standard NER and Part-of-Speech tagging approaches could be applied here, their extraction would result in significant false positives. Therefore, we divided our geospatial extraction process into two steps: segmentation, and entity and location lookup.

Segmentation: The primary goal of our geospatial extraction is to balance the number of entities extracted while reducing the false positive. Traditionally, single token based geospatial extractionwould only result in limited number of locations, while N-Gram based extraction [12], if without proper validation, will lead to significant false positives. Our segmentation technique combines both single token lookup and n-gram extraction; it associates words based on their co-occurance in tweet stream. It depends on a greedy algorithms that will automatically group one-or-more words into a segment. Our algorithm further ranks the existing segments based on their prior usage in all the tweets.

This technique is especially useful for the short and noisy tweets, since a subset of location are also part of commonly used phrases (e.g., New York could be correctly recognized, instead of just being "New" or "York").

Entity and Location Lookup: GTAC leverages Freebase's [13] massive wealth of information to create a hierarchical fuzzy lookup dictionary. This dictionary is used to compare with the tweet segments and determine their entity types (e.g., geo location or people/organization). The results from this matching process are further passed to a location-to-geocoordinate table to verify and validate the location information. In this step, entities with a corresponding geocoordinate will be stored for spatial visualization, whereas results without geocoordinate will be considered a named entity for contextual information visualization, as described in the following section.

3.2.3 People, Organization and Entities Extraction

Beside geo-information, the aforementioned fuzzy lookup dictionary also contains a variety of known named entities, including venues, brands, companies, government agencies, and sports teams. It serves as a valuable source for contextual information extraction regarding people, organization and topics. Essentially, GTAC utilizes this process to acquire context as to "who" and "what" has been mentioned. In addition, extracting named users becomes straightforward by matching tokens with the "@" symbol. This extracted user information provides the "who", while the remaining extracted tokens in the cluster can be considered the "what" on the TweeterShpere.

3.2.4 Tallies

To illustrate the scale of GTAC, we report the following tallies that are generated within one processing interval. Incrementally, GTAC groups tweets into an average of 288 bins every day with each bin looking at processing Tweets every 5 minutes for event extraction process. Typically, a single bin will contain approximately 1,500 clusters, which is introduced with more details in Section 3.3. The size of the clusters are not uniform, with top 5 percent clusters storing most of the tweets. Out of all these clusters, around 700 contain a location entity, 70 clusters contain a time entity, and around 60 contain both. These are just the number of clusters containing geospatial and/or temporal entities; however, of all the tokens extracted, we collect around 3,000 unique location entities and near 100 unique temporal entities within this 5 minute period. During this interval, we also collect around 11,000 total location entities with near 600 total temporal entities. The computing environment GTAC is currently running on is a 48 core Xeon Processor @2.00GHz with 125 GB of RAM while the database is being distributively stored on a 24 core Xeon Processor @2.30 GHz with 62 GB of RAM.

3.3 Semantic Event Indicators Association using Graph Structure

In terms of event detection, times and locations mentioned in the context of a tweet are much more relevant. GTAC uses a graph data structure to model such relationship between event indicators isolated from incoming tweets. The purpose of this graph is to store the tweets in a meaningful manner, which allows for later clustering by a Markov Clustering Algorithm (MCL).

We start building our graph by receiving streamed tweets one at a time, breaking them down into tokens, and adding the counts to graph incrementally. We defined the graph G = <T,E>, where T is the set of tokens and E is the edge set which contains undirected links between these tokens. Each token is a unique phrase or word, while an edge represents the co-occurrence between these words. The edge between nodes contains a weight which signifies the frequency of the two terms have been simultaneously used.

To find semantic associations between the extracted temporal and geospatial indicators we ran the previously mentioned graph structure through an MCL algorithm. The MCL analysis is a key step in GTAC as it associates the extracted event indicators into events. These associated structures represent the investigative 5W's and are later used in our visualization to facilitate domain users' event investigations. This algorithm locates natural groupings of tokens and outputs them in the form of clusters. Clusters are tokens that have been deemed related to each other based on the frequency of their conjunctive use. Although it is usually the case that not all the aspects of an event are contained in a single tweet, the MCL algorithm can connect geospatial and temporal tokens extracted from multiple tweets and group them into clusters to show possible events.

To provide a comprehensive event structure, data of the current MCL cluster must be compared and related back to previously clustered data. To achieve this, GTAC uses the sorensen similarity coefficient [27], a statistic measure developed to determine the similarity between two separate samples (i.e., MCL clusters). Using this index, we compare each cluster from the MCL process against previous cluster. This process allows the tracking of the evolution of clusters over time, essentially showing how events change as time proceeds.

The graph clustering and comparison process is able to keep up with the streaming data mainly due to the binning strategy mentioned in Section 3.1. Binning allows us to keep the tweet graph at a manageable size, minimizing the clustering time. Binning also allows us to distribute the comparison process among a pool of workers to further increase scalability.

3.4 Visually informing and involving humans in the Event Analysis Process

To support event-level investigative analysis on the Twitter Streaming data, GTAC implements coordinated interactive displays to show patterns in near real-time. The main goal for the visualization system is to allow users to explore, interact, and probe the event structure to discover new insights. It is important that a human is involved in the analysis process because it allows them to inject their own domain knowledge to help classify and confirm events, further increasing the applications practical use in multiple scenarios. Inspired by Segel and Heer's narrative genres [25], we extend our previous work [11] and have developed a Partitioned Poster style interface designed to assist in summarizing events into narratives, as shown in Figure 1.

Our visualizations is rendered on client-side using web-based rendering techniques (e.g, D3 [6] and WebGL) to transform and view information that has been aggregated and stored in the NoSQL database. The web-based design aims to increase accessibility and to allow further exploration of the data by remote users. To reduce the communication cost between the client and server, we created a web REST API that allows information to be pushed only when requested by users.

On the high-level, GTAC presents the users with multiple event facets, including geospatial, temporal and content visualizations. It allows the users to interactively navigate through event information, browsing and selecting specific temporal range, and finally narrow down emerging patterns. On the detailed view, GTAC shows a group of tweets based on users' selections and highlights the association of those tweets on an interactive timeline. In the following section, we will focus on introducing the high-level views in GTAC and their intended analysis aspects.

3.4.1 Geospatial Visualization

As shown in Figure 1A, we present the extracted geospatial information by using a interactive 3D spatial visualization. Built upon a customized WebGL process, our geospatial view provides an interactive map allowing users to zoom and pan to different regions. To reduce visual clutter, the extracted event indicators are aggregated in this view in a hierarchical form. Specifically, we keep tallies for cities, states and countries, then plot their magnitudes on the globe in the form of a three dimensional cone. Cities or areas with more frequent mentions appear as taller cones with wider top-radius and a darker blue tone; whereas less mentioned location entities are rendered in shorter and white cones. Our visualized event cones are normalized based on all the mentioned cities in that batch.

By utilizing space and color, our view immediately draw users' attentions to geospatial areas that are frequently mentioned from the collected tweets. This is especially useful in informing the location based events, especially in times of natural disasters and diseases. As shown in Section 4, our geospatial view was a key portal for users to effectively identify the breaking news of Boston marathon bombing.

3.4.2 Temporal Visualizations

To facilitate multi-scale temporal analysis, we developed two interactive visualizations that allows users to monitor events in both retrospective and future time frame. In particular, a master timeline view (Figure 1D) is presented for the users to depict and compare event cluster counts at time of tweet collections. This view provides the ability to examine temporal indicators extracted from tweets. Using this view, the user can highlight a time range to select a subset of data (e.g, a few minutes to single day or months) to be analyzed retrospectively.

To illustrate events that could happen in the future, we also designed a temporal heatmap view that visualizes the event indicators that are deemed to be in a future timeframe. This interactive visualization takes the dates extracted from twitter data and plots them on a interactive calendar. Each cell in Figure 1C represents an aggregated mentions of that date in all the existing tweets; dates that have been mentioned more frequently appear in a darker shade of blue.

3.4.3 Content Representation

To provide the users with a contextual sense of what is being tweeted, we used a word cloud representation to model frequency of specific words in the current scope. After selecting a specific location and time, the user is presented with aggregate wordcloud view that displays entity terms of all shown events. Based on users' selection in temporal and geospatial views, the word cloud is generated based upon the results from the sentence splitting, tokenization, and lemmatization. Hovering over an event in the timeline view causes the map and wordcloud views to display only keywords associated with that event rather than the aggregate for the entire entity. Hovering over each term, on the other hand, our system will highlight corresponding geospatial-temporal indictors that are associated with the specific keywords. In addition, to emphasize the most significant terms, as shown in Figure 1B, words used more frequently appear larger and in a deeper shade of blue.

3.4.4 Interactive View Coordination

Each designed visualization in itself is informative, as each provides the user one dimension of the extracted event indicators (*where, when, who and what*). To provide a more comprehensive understanding of the events, our interface also supports a tight integration and coordination between the visualizations. This aims to provide users a coherent analysis environment. Specifically, if data in one visualization module is selected, its corresponding data points in all the other views will be highlighted. It enables users to visually investigate events by connecting the dots between different event indicators; so that they can better explore, probe, and validate events proposed by the system. As shown in the following case study (Section 4), such view coordination presents an intuitive way for depicting events and provides users a succinct event summary of the streaming tweets, as well as details about events of interest upon user's request.

4 Case Study

To evaluate the efficacy of GTAC in detecting events from streaming Twitter data, we applied the visual analytics environment to monitor and explore on-going and future events. The goal of our case study is to assess whether GTAC can efficiently provide spatial-temporal awareness and direct users attention to new emerging events. In this section we will first describe the dataset, then follow up with the analysis performed using GTAC, and finally show how they compare to news data sources.

4.1 Dataset

As described earlier, GTAC receives streaming data from the Twitter Garden-Hose stream in an unfiltered manner. Our analysis is based on the 1% sample of tweets that has been published on Twitter, and is conducted in near real-time. In this process, GTAC analyzed and aggregated the twitter data every 5 minutes, and further pushed the extracted event information to the multiple coordinated visualizations. For the purpose of concisely showcasing GTAC, this case study focused on describing our analysis process on the day of the Boston Bombing using historical data collected on April 15th, 2013 (Figure 4); however, we explored the data in chronological order, essentially analyzing the data as if it were streaming. This allowed us to replicate the streaming nature of Twitter data and assess GTAC's capability in facilitating analysis of detecting on-going events.

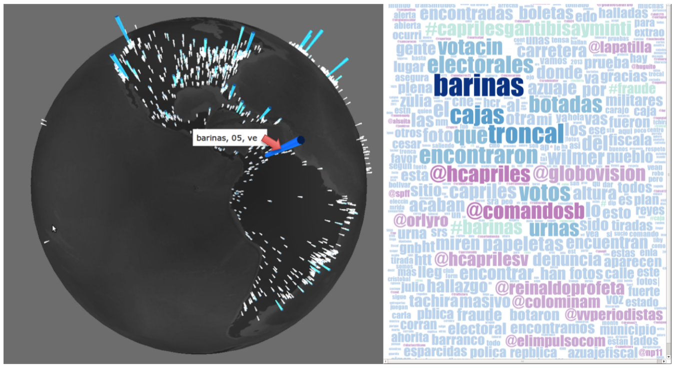
barinas, 05, ve
encontradas boletas edo halladas
#caprilesgantibisayminti
votacin carretera @lapatilla
electorales prueba
barinas azuaje por #fraude
cajas botadas militares
quetroncal fiscala
encontraron wilmer pueblo
@hcapriles @globovision
sitio capriles votos altura todos
acaban @comandosb
@orlyro #barinas urnas sido tiradas
miren papeletas encuentran
@hcaprilesv denuncia aparecen
encontrar han fotos calle
hallazgo @reinaldoprofeta
tachira masivo @colominam
pblica fraude botaron @vvperiodistas
electoral encontramos municipio
ahorita barranco @elimpulsocom
esparcidas polica repblica azuajefiscal

Fig. 3. Case Study: Making sense of on-going events. Geospatial view (left) indicated a significant surge of spatial indicators over Barinas, VE. (with a blue hue). A quick examination of the Word cloud view (right) indicates many discussions on “riot” in Venezuela.

4.2 Use Case

To fairly assess the capability of GTAC, we chose an undergraduate student who was not involved in the development of the system. Since the case study was performed after the Bombing event, the user was aware of such news. However, we purposely offset the temporal indicator by a month, so that at the time of the study he was not aware of the date in which the data was collected. During the study, we updated the evolving information and refreshed the visualizations in the same way that our online system did. Specifically, the data presentation started around 4:00 am and was streamed until 5:00 pm in the afternoon at that day. This setup was intended to keep a more realistic analysis environment for the participant.

The user began his analysis by browsing the geospatial heat map. By hovering over different locations of various countries, he tried to examine events that were presented around the globe from the morning. After a quick scanning on the geospatial view, the user was able to identify top city mentions and use this as a start for further exploration. As shown in Figure 3, the user identified an area of particular interest, located in “Barinas, Venezuela”.

By selecting these geo-entities, the user was able to have a more focused analysis of possible events at this location. With further exploration of the content view, other terms relevant to the location and event began to emerge. Although the majority of terms were in foreign languages, making it difficult to make sense of what was being said, the user was able to pick several interesting key terms that were marked with a high-frequency. Further exploration of the “Barinas” area revealed that an incident involving the burning of voting ballots was taking place during a political election. In response, a riot broke out and the police and troops began attacking villagers, as can be seen from the following translated tweet: “help the police and army attacked the village of barinas with buckshot and tear gas need help urgently” [2]. The user found this information very interesting since he was aware of the regime change in Venezuela in 2012.

The user expressed that he wasn't aware of the continuation of the political fallout there and was interested in monitoring the development of this riot. As he continued watching global attentions

[2] Original Tweet was: “RT @AZUAJE_WILMER: ayuda la policia y el ejercito arremete contra el pueblo de barinas con perdigones y bombas lacrimogenas UREGENT”

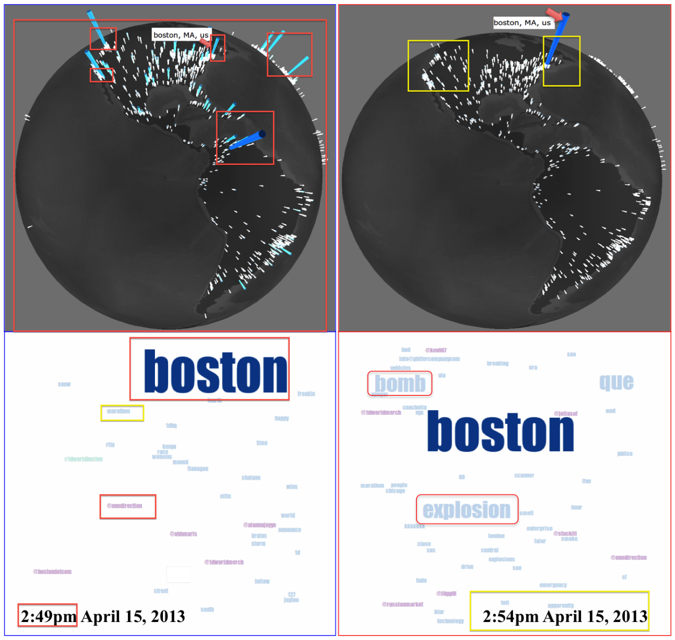

Fig. 4. Snapshot of GTAC's visualizations before (left) and after (right) the Boston bombing. One can identify the big surge of tweets that discussing about Boston as well as the emergence of keywords like "explosion" and "bomb".

to Venezuela until 2:40pm, his attention shifted as GTAC indicated a new high-frequency geolocation (Boston) becoming the center of focus around the globe. The user noticed that there was a tremendous shift in twitter discourse; the Boston mention bar dwarfed all other location mentions.

As shown in Figure 4, Boston had been significantly highlighted on the geospatial view due to the extracted geo-entities towards that city. The user quickly hovered over to Boston and checked the information by using the content view. By shifting time windows on the master timeline view, he observed that, prior to 2:55pm, the mentions of Boston were typically referencing the Red Sox game and the Boston Marathon, which were co-occurring; however this content was overwhelmingly changed to discussions of "Explosion" and "Bomb" in the content visualization, which coincide with the later-reported 2:50 pm bombing at the Boston Marathon. Within minutes, the user noticed that tweets began to flood in, spreading news and showing support for those in Boston. Twitter users shared insights by posting information directed at Twitter, uploading images and videos of the event, allowing those in charge to crowd source the information from many different perspectives. As a result, GTAC was able to detected the booming event around 2:54pm (3 minutes after the first tweet).

This case study shows the speed and efficacy of GTAC at discovering events. It was able to find the Boston Marathon Bombing literally within minutes after the event broke out. While the earliest news sites published articles covering the story as early as tens-of-minutes after-the-fact, GTAC was able to inform the active user of this new and ongoing event as it was happening. Interactive visual investigations of events highlighted by the geospatial-temporal heat map allowed the user to gain further insights on the highlighted areas, showing interesting events from across the globe.

5 Discussion, Limitation, and Future Work

We undertook this research to design a scalable visual analytics pipeline that extracts structured representations of events from Twitter stream, and supports event-level investigative analysis of the social media data. To this end, we presented GTAC, an analytics pipeline that demonstrated unique and effective capabilities for addressing a class of problems that involve streaming event detection, association, and representation in real-time.

The design of GTAC is grounded in the characterization of challenges in the event analysis process, and its algorithm is further materialized through parallel natural language processing and distributed data storage. By combining data-driven analytics and interactive visualizations, GTAC provides an investigative environment for decision makers to access and depict the events that are reported from social media.

There are limitations to our current research that need to be further addressed. On the one hand, GTAC will enrich the streaming event extraction process. Our approach relies on automated algorithms to discover information regarding who, what, when, and where in order to characterize and structure an event. Inevitably, the final event representations are influenced by the performance of each algorithm. In particular, location extraction from social media content is a nontrivial task; while our current method works for extracting cities, states and countries, there are sometimes false positives due to the ambiguous nature of location names and the way they can be expressed on Twitter. In addition, locations from social media are not always tied to a physical address. Thus, the accuracy of detected entities relies on the performance of the named entity recognition (NER) algorithm.

Solving the issues in the short term is challenging, but we think it is useful to make users aware of these issues and further involve them in the process of refining and reducing the false positives. In order to do so, we plan to borrow methods from uncertainty visualization and crowdsourcing methods to annotate different layers of uncertainty so that users can make more informed decisions during investigation and analysis. Moreover, we hope to utilize the words surrounding the possible location to see if there are further matches and to improve our confidence when extracting. By focusing on a specific analysis domain, we believe that will allow us to extend our extraction capabilities to not only location names, but also to non-physical locations, such as cyber gather and forums (e.g., Reddit).

On the other hand, we will conduct further user study and implement crowdsourcing to verify and validate the extracted events. Beside conducting a formal user study and gathering feedback based on the design and efficacy of the pipeline, we also aim to validate the assumed coherence and relevance of event structuring through crowd sourcing and summarizing the results based on their statistical significance. This will combine machine intelligence with human experience. Following the suggestion from Chang et al's [7] work, we will focus on identifying metrics that fit the real-work tasks in a specific domain. The metrics will be statistical but will be shaped by domain-specific constraints. In addition, we will improve our spatial-temporal heat map to better show anomalies in the identified graphs. This will help the users better understand interesting outliers within the dataset by allowing them to focus on/detect areas that are behaving out of character.

While there is still future work needed on our presented architecture, it illuminates the strong role that a combined approach of data-driven modeling and user-centered visual analytics can play in revealing the real-time events within complex and noisy social media streams. We have demonstrated its efficacy through a detailed case analysis. It is our hope that by identifying these system limitations, the research domain of visual analytics, event detection and association, parallel computing, distributed data management could be brought together to provide scalable solutions for streaming event analysis in social media and new techniques for revolutionizing the analysis environments.

6 Conclusion

While computation algorithms can be similar, the perspective of how to apply them changes when considering streaming and real-time data. The key end-goal, however, should still be focused on how to facilitate the human decision making process. In this paper we present a visual analytic pipeline that combines automated event detection and association processes and interactive investigative visualizations to facilitate the analysis of streaming Twitter data. Our approach focuses on the extraction of the event indicators (*who, when, where, and what*), which are structured into events, by integrating MCL clustering, NER, and NLP techniques. The analysis of such event structures is further enhanced by interactive visual interfaces, providing results that can be explored, filtered, and managed by users. The resulting interface creates a real-time analysis environment for identifying event structures, geographical distributions, and key indicators of emerging events. To demonstrate its efficacy, we performed a case study in which a domain user monitored the development of the Boston Bombings in near realtime. The results illustrate that the GTAC can not only help depict the emergence of an event, but also provide context and information on what is happening from streaming tweets.

7 Acknowledgment

The authors greatly acknowledge the support from the Army Research Office (ARO) under Grant #A13-0154-001. The authors would also like to thank Nora Bachman for her helps in editing the paper.

References

[1] I. 10gen. Mongo db. online.

[2] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of new topics. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 10–18, New York, NY, USA, 2001. ACM.

[3] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 291–300, New York, NY, USA, 2010. ACM.

[4] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi. Eddi: interactive topic-based browsing of social status streams. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, UIST '10, pages 303–312, New York, NY, USA, 2010. ACM.

[5] F. Bilhaut, T. Charnois, P. Enjalbert, and Y. Mathet. Geographic reference analysis for geographic document querying. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Volume 1*, HLT-NAACL-GEOREF '03, pages 55–62, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[6] M. Bostock, V. Ogievetsky, and J. Heer. D¡sup¿3¡/sup¿ data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, Dec. 2011.

[7] J. Boyd-Graber, J. Chang, S. Gerrish, C. Wang, and D. Blei. Reading Tea Leaves: How Humans Interpret Topic Models. In *Neural Information Processing Systems (NIPS)*, 2009.

[8] T. Brants, F. Chen, and A. Farahat. A system for new event detection. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 330–337, New York, NY, USA, 2003. ACM.

[9] Y.-F. R. Chen, G. Di Fabbrizio, D. Gibbon, S. Jora, B. Renger, and B. Wei. Geotracker: geospatial and temporal rss navigation. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 41–50, New York, NY, USA, 2007. ACM.

[10] W. Dou, D. X. Wang, Z. Ma, and W. Ribarsky. Discover diamonds-in-the-rough using interactive visual analytics system: Tweets as a collective diary of the occupy movement. *AAAI International Conference on Weblogs and Social Media*, 2013.

[11] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou. Leadline: Interactive visual analysis of text data through event identification and exploration. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, Oct. 2012.

[12] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. Part-of-speech tagging for twitter: annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[13] I. "Google. Freebase.

[14] I. G. Hunt. Speaks on structured data conference, March 2013.

[15] G. Kumaran, J. Allan, and A. Mccallum. Classification models for new event detection. *Applied Optics*, 15:2513–2519, August 1980.

[16] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee. Twiner: named entity recognition in targeted twitter stream. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 721–730, New York, NY, USA, 2012. ACM.

[17] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. 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, New York, NY, USA, 2011. ACM.

[18] F. Morstatter, J. ürgen Pfeffer, H. Liu, and K. M. Carley. Is the sample good enough? comparing data from twitters streaming api with twitters firehose. 2013.

[19] S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 181–189, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[20] D. Ramage, C. D. Manning, and S. Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 457–465, New York, NY, USA, 2011. ACM.

[21] T. Reuters. Costa rica earthquake. online, Sep 2012.

[22] A. Ritter, Mausam, O. Etzioni, and S. Clark. 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.

[23] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 695–704, New York, NY, USA, 2011. ACM.

[24] 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.

[25] E. Segel and J. Heer. Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1139–1148, Nov. 2010.

[26] L. Shi, F. Wei, S. Liu, L. Tan, X. Lian, and M. Zhou. Understanding text corpora with multiple facets. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 99–106, 2010.

[27] T. Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol. Skr.*, 5:1–34, 1948.

[28] M. Verhagen, I. Mani, R. Sauri, R. Knippen, S. B. Jang, J. Littman, A. Rumshisky, J. Phillips, and J. Pustejovsky. Automating temporal annotation with tarsqi. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, ACLdemo '05, pages 81–84, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[29] C.-P. Wei and Y.-H. Chang. Discovering event evolution patterns from document sequences. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 37(2):273–283, 2007.

[30] R.-F. Xu, W.-H. Peng, J. Xu, and X. Long. On-line new event detection using time window strategy. In *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on*, volume 4, pages 1932–1937, 2011.

[31] C. Yang, X. Shi, and C.-P. Wei. Discovering event evolution graphs from news corpora. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 39(4):850–863, 2009.

[32] Y. Yang, J. Zhang, J. Carbonell, and C. Jin. Topic-conditioned novelty detection. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 688–693, New York, NY, USA, 2002. ACM.