CrystalBall: A Visual Analytic System for Future Event Discovery and Analysis from Social Media Data

Isaac Cho*
University of North Carolina at Charlotte

Ryan Wesslen
University of North Carolina at Charlotte

Svitlana Volkova
Pacific Northwest National Laboratory

William Ribarsky
University of North Carolina at Charlotte

Wenwen Dou†
University of North Carolina at Charlotte

ABSTRACT

Social media data bear valuable insights regarding events that occur around the world. Events are inherently temporal and spatial. Existing visual text analysis systems have focused on detecting and analyzing past and ongoing events. Few have leveraged social media information to look for events that may occur in the future. In this paper, we present an interactive visual analytic system, CrystalBall, that automatically identifies and ranks future events from Twitter streams. CrystalBall integrates new methods to discover events with interactive visualizations that permit sensemaking of the identified future events. Our computational methods integrate seven different measures to identify and characterize future events, leveraging information regarding time, location, social networks, and the informativeness of the messages. A visual interface is tightly coupled with the computational methods to present a concise summary of the possible future events. A novel connection graph and glyphs are designed to visualize the characteristics of the future events. To demonstrate the efficacy of CrystalBall in identifying future events and supporting interactive analysis, we present multiple case studies and validation studies on analyzing events derived from Twitter data.

Keywords: Social media analysis, Event detection and analysis, visual analytics

Index Terms: K.6.1 [Management of Computing and Information Systems]: Project and People Management—Life Cycle; K.7.m [The Computing Profession]: Miscellaneous—Ethics

1 INTRODUCTION

Social media provide platforms for people to respond to and communicate about events in real-time. Such events range from breaking news events that are of international, national, and regional interests, to events that are only relevant to an individual's immediate social circle. Posts on social media capture information including the scale and spread of discussions pertaining to various events, the lengths of discussion on the events, as well as people's sentiment towards different events and change of sentiment over time. Prior research has discovered that events serve as a succinct summary of large temporal text corpora [14]. And much work has been devoted to identifying past and ongoing events from social media [32].

More recently, people started to leverage social media to plan, organize, advertise, and inform others about future events. Examples include music and sports related events, social movements/campaigns such as the Bank Transfer Day [1], the Occupy Movement [2], as well as possible cancellation or change of future event due to conditions such as inclement weather. Many of these events develop spontaneously and may not be known to *city officials, public safety personnel, planners, facilities managers, or others who might be interested in participating and contributing to the events*. Furthermore, events that begin spontaneously or with a narrow focus may change and grow over time, even before a main public occurrence manifests itself. This is what happened, for example, with Occupy Wall Street [2]. Gathering, analyzing, and visually presenting information about future events will empower individuals to foresee and even be prepared for the events. On the one hand, individuals can plan ahead or decide whether to participate if there will be events of interests taking place in locations near them. On the other hand, stakeholders such as police departments or city event planners can plan and allocate resources ahead of time to encourage peaceful and orderly crowd gatherings or demonstrations.

The motivation for this paper is to develop a system that supports the discovery and characterization of future events from social media streams. Characterization in terms of time, location, topic, and social network permits much stronger identification of the future event than in terms of any one or two of these attributes. This is quite important because even significant and carefully planned future events (again, Occupy Wall Street) may remain hidden in the great flow of social media until they are launched. The topics of these discovered events need to be automatically derived, but they then can be followed and investigated in detail through user selection and queries.

A further motivation is that we have gotten feedback from stakeholders and potential users about how useful this capability would be. For example, we have talked with police in Charlotte and elsewhere who would very much like to have the capability to quickly discover future events in their jurisdiction that they should prepare for. Exploring for future events that may indicate emerging risks is of significant interest to banks and related institutions. Finally, we have worked with retailers who would be interested in an variation on the streaming analysis capabilities described here that could indicate when people will be gathering at locations, such as malls, where timely, focused online advertising would be effective. Based on our discussions with different stakeholders that have their own specific needs in identifying relevant future events, we generalized their requirements and designed CrystalBall to discover various types of future events and to allow users to focus on specific event types such as protests and marches.

1.1 Characterizing future events

The characterization of future events impacts the identification of such events from text streams. In previous research, Dou et al. [14] defined an event as:

*"An occurrence causing **change** in the volume of text data that discusses the associated topic at a specific time. This occurrence is characterized by **topic** and **time**, and often associated with entities such as **people** and **location**."*

Such a definition allows the identification of past and ongoing

*e-mail:icho1@uncc.edu
†e-mail:wdou1@uncc.edu

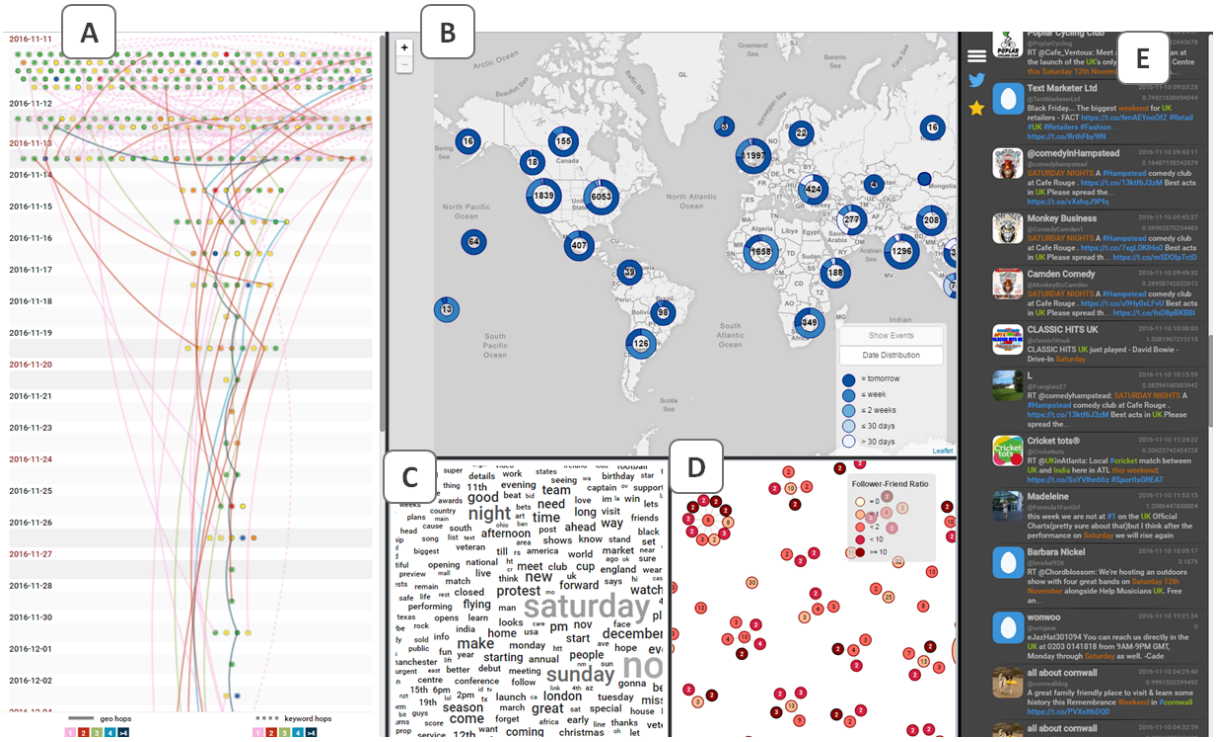

Figure 1: CrystalBall interface. The interface is comprised of four main views: A) Event Calendar View, B) Map View, C) Word Cloud View, and D) Social Network view. The E) Tweet panel is shown on demand.

events by detecting a "change" in volume. Although this definition of events is general, it is not applicable to the identification of events that may occur in the future. As the posts about future events only constitute a very small portion of the social media contents, it is unlikely that future events related posts would cause a detectable change in the volume of social media streams. Therefore, in contrast to previous research on past and real-time event detection [28, 7, 5], we can no longer rely on detecting "bursts" or "spikes" from the volume of the messages posted over time. The challenge of identifying future events lies in sifting through large amounts of social media data and identifying small signals that are buried in the overwhelming information regarding past and ongoing events, personal status updates, etc.

Although posts regarding future events may not be significant in terms of volume change over time, they are likely to refer to multiple key attributes that, considered together, can be significant, these are a future time and a location where the event will occur, as well as the topic of the event and the particular social network involved. Therefore, we define a future event as:

"An occurrence that is associated with a ***location*** *and a* ***date/time (span)*** *in the future. This occurrence is characterized by* ***location*** *and* ***time*** *and usually associated with a particular topic and social network."*

The location and time are the lead attributes defining the *where* and *when* of the future event. Associated attributes including keywords from the texts and the social network strengthen the event identification by providing the *what* and *who*. These latter attributes can be embedded in measures of *informativeness* and *cohesiveness*, as can the *credibility* and *diversity* of the Twitter messages. When combines with future event identification via a pair of location, time, the overall measure provides a much stronger indicator of the event, with higher quality tweets describing the event, than measures based on the pair of location, time alone.

In this paper, we present a visual analytics system, CrystalBall, that identifies, ranks, and visually presents future events, based on the characterization presented in Section 1.1. This system provides event-oriented visualization based on automated algorithms, permitting interactive exploration of likely future events. Our work differs from past research in several aspects. First, much of prior work has focused on detecting past and ongoing events from textual sources [13, 23, 14, 21, 10, 30] while our work aims at identifying future events. Second, prior methods proposed to detect ongoing events are usually limited to certain application domains or events of interest (e.g. earthquakes and other natural disaster-related events), while our approach enforces no such limitation. Instead we first provide a general overview of all types of future events and then allow users to interactively search/filter for events of interest. As a result, we enable both the discovery of a wide range of future events as well as focused investigation of selected types of future events.

Overall, CrystalBall makes four main contributions to the Visual Analytics literature and community:

1. A general method is proposed to discover future events from streaming Twitter messages based on our definition of future events.

2. Multiple metrics are designed and included in CrystalBall to characterize and rank the identified future events.

3. A new interactive visual interface is developed to serve as the front-end of CrystalBall; the web-based interface is tightly integrated with the computational method and metrics to support the exploration and sensemaking of future events.

4. Both case studies and validation studies are presented to demonstrate the efficacy of CrystalBall.

The rest of paper is organized as follows. In Section 2, we provide a motivating scenario for identifying future events from social media data. Section 3 reviews prior related work. Section 4, 5, and 6 describe the CrystalBall visual analytics system, including the modeling and characterization of future events, and the interactive visual interface designed to facilitate the analysis of the event modeling results.

2 Motivating Scenario

Recently, social media have become popular channels to advertise and plan for future events. Some events such as sport games or concerts are well-planned and information regarding the events may also be available in other sources such as news media. In contrast, some other events are more grass-root as social media empower individuals to initiate events that may spread locally or nationally. Knowing about events that may occur in the future would benefit a wide range of stakeholders with diverse interests. On the one hand, individuals may wish to know about events (concerts, sports games, marches, celebrations, protests, strikes) that may take place in nearby locations so they could plan to attend or be prepared for event traffic. On the other hand, local police departments or other government agencies may be interested in knowing crowd gathering events prior to the event date so they can plan and allocate resources to ensure the safety of the attendees. In this section, we use organizational information of the Occupy movement found on Twitter as an example that highlights the importance and impact of detecting and analyzing future events.

The Occupy Movement is a large-scale social movement in terms of participation, the length of the movement, and geographic spread of the related protest events [2]. Based on our data-driven analysis of the rise and fall of the movement [12], we discovered a **precursor** signaling organizing information *well ahead* of the official start date of the Occupy Movement. The volume of messages regarding the organizing activities is an order of magnitude smaller than the messages about the actual protest launched in New York City on September 17, 2011. Therefore, without knowing what keyword to search for, it is highly unlikely that one would detect the organizing activities on Twitter prior to September 17. The small precursors to the Occupy Movement or similar organizing activities are likely buried in large amounts of tweets about ongoing events and personal status updates. From the stakeholder's perspective, knowing about the social movement ahead of time would benefit planning activities. Seeing the signal of the event, individuals can search for additional information regarding the upcoming event and decide whether to contribute or participate. It would also be extremely valuable for the local police force to know about the events ahead of time to allocate proper resources to ensure the safety of the protesters and non-participating citizens, to anticipate and direct traffic in areas near the protest site, as well as prepare the policemen on interacting with protesters. This and related situations are the ones that CrystalBall is designed to address.

3 Related Work

Two areas of research, namely event detection and visual event analysis, are the main inspirations for the design of CrystalBall.

3.1 Retrospective and online event detection

Yang et al. [34] surveyed event detection related papers in the Information Retrieval community and categorized the research into Retrospective and Online detection. Much of the recent work has focused on realtime event detection from social media. Abdelhaq et al. [5] proposed methods to detect "local" events in real-time from tweets by measuring the entropy of the tweets' spatial signature. Becker et al. [7] also extracted events from social media by learning similarity metrics that enable online clustering of events. Along the same line, Ritter et al. [28] presented TWICAL, an open-domain

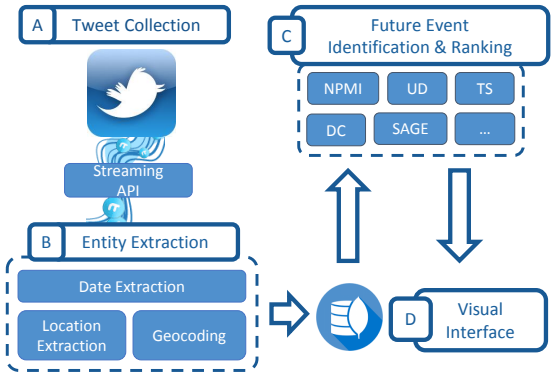

Figure 2: System pipeline of CrystalBall: A) streaming Twitter data collection, B) named entity recognition and geocoding, C) future event identification and ranking, and D) interactive visual interface.

event extraction and categorization system for Twitter. Compared to prior work, Ritter et al. pioneered open-domain event extraction instead of focusing on a certain domain or a specific type of events. Our proposed work shares the same mentality on the open-domain event extraction aspect but focuses on identifying "future" events as opposed to ongoing events.

With respect to future event detection, Radinsky et al. [26] presented methods for learning to forecast forthcoming events of interest from news stories. The events of interest included disease outbreaks, deaths, and riots. Their automated abstraction technique moves the level of analysis from specific entities to consideration of broader classes of observations and events. Sadilek et al. focused on the task of fine-grained prediction of the health of specific people from noisy and incomplete data [29]. They proposed a probabilistic model based on a person's social ties and co-location with sick individuals. Our approach differs from both research in several aspects, including not restricting to a predefined set of events or application domains, a new way to identify future events based on space-time relationship, and providing an interactive visual interface for event exploration and analysis.

3.2 Visual event analysis

Event detection and analysis have been an area of active research in the field of visual analytics. A few survey papers have been published on the topic of event detection. Dou et al. summarized work related to event detection from social media data [13]. Four tasks, including new event detection, event tracking, event summarization, and event association were derived based on literature from the data mining community. Wanner et al. [32] summarized the evolution of event detection in combination with visual analysis and provided an overview of the state-of-the-art methods. The ultimate goal of the survey was to lay ground work towards building guidelines on how to construct successful visual analysis tools that can handle specific event types and diverse textual data sources.

Other research focuses on proposing visual analytic systems that enable the discovery and analysis of events from textual data. Krstajic et al. introduced CloudLines [21], a visualization system that effectively detects important events from large collection of news stories and allows users to interact with the event series. CloudLines enables users to inspect recent documents in the time series in the context of previous ones. Extending from news to social media sources, Dou et al. developed LeadLine [14], a visual analytics system for event identification and exploration. Events are presented in an interactive visualization based on the 4Ws, namely who, what, when and where. Both CloudLines and LeadLines focus on identifying retrospective events from textual sources.

Other work has performed event detection from streaming data. Luo et al. presented EventRiver [23], a visualization system that supports event browsing, tracking, association and investigation tasks on live-stream documents. The system models events taking into account temporal-locality and similar semantics between the documents. Incremental mechanisms were applied to process live-stream documents.

4 System overview and pipeline

In this section, we describe the system pipeline for CrystalBall, shown in Figure 2. CrystalBall integrates multiple components, including data collection from the Twitter Streaming API [3] (Figure 2A), entity extraction (Figure 2B), future event identification and ranking (Figure 2C), and an interactive visual interface (Figure 2D). All the data collection and analysis are done online while the tweets are streaming in. The interface refreshes daily to present results on future events that may occur in the coming days or weeks.

The system pipeline starts with the streaming data collection, with the tweets stored in a MongoDB collection (Figure 2A). Batches of tweets are then sent to the entity extraction component (Figure 2B), which involves date extraction and mapping, location extraction, and geocoding. We employ Stanford SUTIME [9] for the date extraction and TweetNLP [25] for the location extraction. For the geocoding, we use OpenStreetMap Nominatim [24]. Based on the formulation of future events in Section 1.1, our criterion on tweets contributing to a certain future event is that they include a location and a future time reference. The entity extraction component lays the groundwork for the event identification and ranking. Specifically, given a tweet, we first identify if there is a location and a time reference mentioned within the tweet content. We then map the detected time reference to a calendar date and compare the mapped date against the date the tweet was posted to determine if the tweet mentions a future time. If the tweet indeed refers to a future time and also mentions a location, it has met our criterion for further analysis. On average, we can collect around 150,000 qualified tweets per day for future event extraction given our criteria. These tweets then go to the future event identification and ranking process which takes approximately an hour and half. In the next two sections, we will describe in detail the future event modeling component and the visual interface.

5 CrystalBall: Future Event Identification and Characterization

In this section, we describe the "future event analysis and ranking" component of CrystalBall (Figure 2C). The main focus of this component is on identifying future events from tweets and performing further analysis to rank the events in order to present users with events of high quality. The ranking of the events are based on measures we propose in Section 5.2.

5.1 A novel method to identify future events

One important question that is yet to be solved is how to detect small future event signals from tweets? As mentioned before, the tweets about future events only constitute a tiny percentage of the tweet stream, with the number of tweets related to planning activities of an upcoming event being in the teens or fewer. Therefore we need to find a way to extract such small signals from noisy data.

Normalized Pointwise Mutual Information Based on our formulation of future events (Section 1.1), the deterministic factors of a future event are a location and a reference to future time. Therefore, we propose to identify future events by modeling the correlation between the mentioned locations and future dates. To this aim, we measure the Normalized Pointwise Mutual Information (NPMI)[8] of location-time pairs identified from the tweets. More specifically, when both a location and a future time are mentioned in a tweet, the location-time pair is stored in our database. As more tweets with the same day are processed, the counts of the location-time pairs are updated. After identifying the location-time pairs from daily tweets, we calculate the NPMI as follows:

$$NPMI(loc,t) = \frac{PMI(loc,t)}{-\log p(loc,t)}$$

$$PMI(loc,t) = \log \frac{p(loc,t)}{p(loc)p(t)}$$

p(loc) and *p(t)* are marginal probabilities of each location and future time extracted from the tweets while *p(loc,t)* is the joint probability of a location-time pair. The PMI of a location-time pair is a measure of how much the actual probability of a particular location-time pair differs from what would be expected based on the probabilities of the individual occurrences and the assumption of independence. A completely uncorrelated location-time pair would receive a PMI of 0 [8]. To enable the comparison of the correlation between different location-time pairs, we convert PMI to NPMI to give the measure a fixed upper bound of 1 (with lower bound being -1). We calculate the NPMI score for all location-time pairs found in the tweets. Positive NPMI indicates a correlation compared to being independent, with the pairs with higher NPMI scores indicating a greater correlation between the when and where aspects of a future event. In CrystalBall, we save location-time pairs with positive NPMI scores for further analysis, which enables us to detect weak signals from the tweets.

After calculating the NPMI, we now have a list of location-time pairs that may serve as indicators of future events. The NPMI alone only captures the correlation of a time and a location. There is other information we can leverage to determine the quality of the future events. In the next section, we describe additional measures to characterize and rank the possible future events.

Table 1: Measures of a future event

Event Identification	
NPMI	Normalized Pointwise Mutual Information
Event Tweet Informativeness	
LR	Link Ratio
HR	Hashtag Ratio
UC	User Credibility
UD	User Diversity
Event Tweet Cohesiveness	
DC	Degree Centrality
TS	Tweet Similarity

5.2 Characterizing and ranking future events

As crucial as the NPMI is in determining future events, we need other metrics to rank the discovered events in order to visually represent the most relevant ones to end users. In this section, we describe six additional measures that characterize the future events. From these RankSVM model is built to leverage all measures for ranking events. Previous work has ranked tweets based on informativeness and trustworthiness of the content [18]. Our six additional measures were developed following this train of thought. Table 1 provides an overview of the measures.

5.2.1 Measures on the informativeness of tweets related to future events

To speak to the quality of the detected possible future events, we present four measures that are related to the informativeness of the

tweets regarding individual events. We build on findings from prior work [18] and propose measures including link ratio, hashtag ratio, user credibility, and user diversity.

Link and hashtag ratio: As found in previous work [18], tweets that contain hashtags and links to external sources are more likely to be informative. The linked sources include web blogs, images, news articles as well as Facebook pages for a future event. We measure the ratio of tweets containing links (LR) over all tweets that are related to a possible future event. Similarly, we measure the hashtag ratio (HR) of tweets related to one possible future event.

User credibility: Previous work has found that "informative tweets are more likely to be posted by credible users" [18]. Extending the logic, a future event is likely to be valid if tweets related to this event are posted by credible users. There are multiple ways to measure the "credibility" of users based on the number of followers, mentions, and retweets. We choose a simple measure, the Twitter Follower-Friend (TFF) ratio, to represent the user credibility [33]. The TFF is the ratio of followers to friends. A ratio of between 1.0 to 2.0 indicates that the user has a balanced following/follower relationship. A ratio of less than 1.0 means not as many accounts follow the user, while a TFF ratio of 2 and above indicates progressively higher popularity. For a future event, we calculate the user credibility as the average of the TFF ratios across all users tweeting about the event.

User diversity: Another measure that speaks of informativeness is related to the sources of the tweets. If all tweets regarding one potential future event all came from one account, it is likely that these tweets are from a bot that is programmed to send out certain tweets periodically. To be able to demote the possible events related to such tweets, we measure the diversity of the sources of tweets linking to one future event. The User Diversity (UD) is calculated as the number of unique users divided by the total number of tweets related to one future event.

5.2.2 Measures on cohesiveness of tweets related to possible future events

The aforementioned measures contribute to evaluating the informativeness of the tweets regarding future events. However, there are times that individual tweets related to one location-time pair appear informative but a subset of the tweets may not be related to the event that is going to occur in the location. For instance, a group of tweets may all refer to NYC and April 3 (a future date). A majority of the tweets could be referring to an upcoming concert. But some tweets may not be related to the concert event, such as personal status updates "I am going to visit my friends in NYC on April 3", which happens to mention the same location-time pair. To measure the cohesiveness of the extracted events based on the tweets, we propose two additional measures on network centrality and tweet similarity.

Degree Centrality: In CrystalBall, each identified future event is associated with a set of tweets that mention the same location and future date. We hypothesize that if the tweets are connected to each other, they may be more likely to refer to the same upcoming event. To measure the connections among these tweets, we construct a social network based on retweet (RT) and mention (@) information from the set of tweets regarding one possible future event. We then calculate the degree centrality (DC) using Freeman's general formula for centralization [16]. A highly connected tweet network would have a degree centrality close to 1 while a scattered tweet network yields centrality close to 0.

Tweet Similarity: When we analyzed the tweeting behavior centered around events, we discovered that Twitter users do not always credit the original tweet source using mentions (@) or retweet (RT). Therefore, there are situations where the content of all tweets related to one possible event is very similar but the tweets are not connected to each other The degree centrality in this case would fail to capture the similarity among the tweets. To address this issue, we propose tweet similarity, a measure to evaluate the content similarity of the set of tweets linked to one future event. Given the set of tweets, we first compute the similarity between every pair of tweets using the Levenshtein distance [22]. We then average the similarity across all pairs to derive a final score. The Tweet Similarity is a measure ranging from 0 to 1, with 1 denoting that all tweets are identical in the collection.

So far, we have presented a measure to identify possible future events (NPMI), and six additional measures to characterize the events. The next step is to combine these measures to evaluate the quality of the identified future event. Given the exploratory nature of CrystalBall, we do not want to disregard any types of events before users have a chance to see them since different users may be interested in different types of events. Instead we want to rank the events so that CrystalBall visually represent events of high quality first. As observed during our analysis, lower quality events are mainly comprised of advertisement messages or tweets from bots.

Table 2: 5 categories for the RankSVM training dataset

Category	Description
4	geopolitical events (e.g. protests)
3	non-geopolitical events (e.g. seminars, conferences)
2	social events (e.g. concerts, sports)
1	periodic events (e.g. weather forecast, traffic updates)
0	false positives

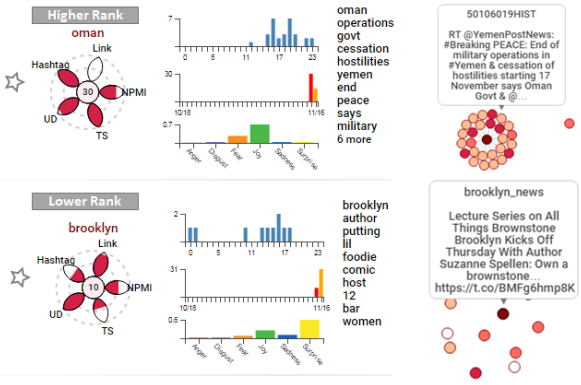

Figure 3: Examples of two events from the same day that received high and low rankings from the RankedSVM model. The top (higher-ranked) event is seen with a tightly connected social network while the bottom (lower-ranked) event has a more scattered social network.

5.2.3 Ranking future events

The seven measures described in the previous section aim to identify and characterize events that are likely to occur in the future from Twitter streams. The measures provide an opportunity to rank the events based on the informativeness and cohesiveness of the tweets related to a certain future event. We adopted RankSVM [19] to evaluate the relationship between the explanatory variables (the seven measures) and a response variable denoting the overall quality/rank of a future event. To train the RankSVM, we developed a labeled dataset comprised of extracted future events from three days (approx. 1000 events). We defined 5 categories for event labeling as listed in Table 2. The labeling decision indicates that we value events that are geopolitical and of grassroots nature.

Five coders independently went through the 1000 extracted future events and ranked the events using the above categorization.We then used the labeled dataset to train the RankSVM and developed a model that can be applied to rank events without labels. In the

CrystalBall visual interface, the **order** of the events shown within each day in the Event List view (as seen in Figure 3) reflects the RankSVM results.

In summary, we described the analytic component of CrystalBall in this section. The analytic component focuses on identifying, characterizing, and ranking future events from streaming tweets. In the next section, we present the visual interface that deliver the analytic results regarding the future events to end users.

6 CrystalBall: Visual Interface

CrystalBall includes web-based interface that is tightly integrated with the analytic component in order to present the most up-to-date results regarding future events. The interface is designed to permit the discovery and analysis of future events in an intuitive manner. Leveraging prior work that has successfully modeled events based on the 4Ws (who, what, when, where) from investigative journalism [14], our visual interface includes 4 main views, with each centered around one of the 4Ws.

6.1 Event Calendar: When will the Events Occur?

The Event Calendar presents an overview of events based on when the events will occur. More details about the selected subset of events (event list) are shown upon user interaction.

6.1.1 Future Events Overview

Figure 4: Future events overview. A) Solid lines indicate events that share the same locations, while B) dotted lines indicate events sharing the same keywords.

The future events overview (Figure 1A and Figure 4) is designed to present the identified events and the connections among events. The view is divided into rows with each row denoting a day. Within each day, a number of future events that will occur on this date is listed, with each event drawn as a circle. The date information is displayed on the top left of each row with Sundays showing in dark red[1] (Figure 4a). To present the connection among events, we examine whether multiple events share either keywords or location. To visually represent the relationship among events, links are drawn to connect events that may be related as follows:

A solid line is drawn to connect two events that share the *same location*. Hovering the mouse over the link will invoke a pop-up window showing the future events occurring in the same location (Figure 4A).

A dotted line is drawn to connect two events that share *2 or more keywords*. Hovering the mouse over the link will trigger a tooltip displaying the shared keywords (Figure 4B).

[1]Note that the days in the Event Calendar are not continuous as in a regular calendar since there are future dates with no events identified.

Figure 5: Colors for emotions [11] and fuzziness for uncertainty [17].

The color of the links is determined by the number of events one link encompasses. As seen in the legend on the bottom of Figure 4b, a distinct color is assigned based on the number of hops. The color of event circles encode emotion towards the event. We employ an emotion analysis classifier trained on Twitter data [31] to infer six emotions (anger, disgust, fear, joy, sadness and surprise) from tweets related to each event. The result of the emotion analysis is in the form of probabilistic distributions over the six dimensions. For some events, the distribution seems to be concentrated on a particular emotion dimension; while the distribution is more even for other events. To encode the uncertainty of the emotion analysis results, we used fuzziness which has been shown to be an effective representation of uncertainty [17] to present the emotion analysis results in three levels based on the entropy ($<$0.50, $<$0.75 and $\leq$1.0). The events with emotion analysis results of higher uncertainty are fuzzier as shown in Figure 5.

Overall, the Future Event Overview (Figure 4) is designed to show the identified events by date and the relationship among these events. To get more details on the events, the user can click on either a date or a link to go into the Event List view.

6.1.2 Event List View

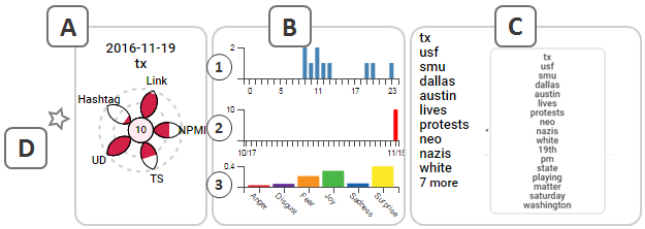

Figure 6: Event List view. A) The flower glyph visualizes five measures of the future event (LR, HR, NPMI, UD, and TS) with the number of tweets in the center. B) The top bar chart shows distribution of the tweet posting time, the middle chart shows distribution of the number of tweets in 30 days, and the bottom chart shows the predicted emotion values of the event. And C) Keyword summary of a future event.

The Event List view occupies the same screen space as the Future Event Overview and is designed to present the characteristics of the selected events. As seen in Figure 6, information regarding an event consists of five elements: a flower glyph on the left, two timeline bar charts in the middle, an emotion analysis result, and an event keyword list on the right. The flower glyph is designed to present values of five out of the seven measures of a future event introduced in Section 5, with the number of tweets shown in the center of the flower. The rest of the two measures on user credibility and network centrality are portrayed in the social network view (Section 6.4). Each petal in the flower glyph is filled based on the value of the corresponding measure.

Two timeline bar charts are shown for each event to present information regarding when the tweets related to one future event were posted and the number of tweets daily mentioning the same location and time pair in last 30 days (Figure 6B1 and B2). The first bar

chart provides information on when a future event was mentioned within a day. The second bar chart shows how a particular event could be mentioned within a 30-day window. The list of keywords that summarize an event is shown next to the bar charts (Figure 6C). The star (Figure 6D) on the left of the event allows users to bookmark an event of interest. Bookmarked events are saved in the CrystalBall system and users can later revisit the events. Overall, the Event List view presents detailed information regarding individual future event, thus facilitating the exploration and analysis of events of interests.

6.2 Map View: Where are the Upcoming Events?

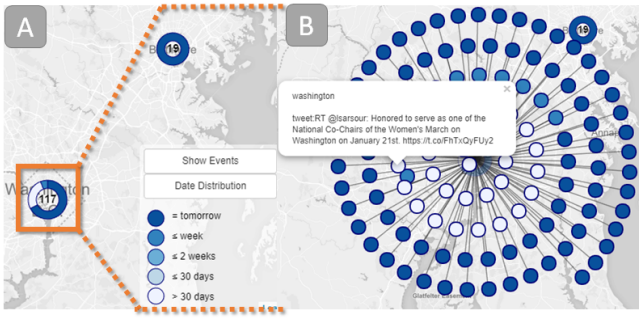

Figure 7: A) Overview of future events in the Washington D.C. area. The ring segments indicate that there are at least 3 different events in the area. B) Clicking the cluster for detailed events.

In addition to when will the events occur, another important aspect regarding future events is where they will occur. The Map view presents information pertaining to the locations of the future events (Figure 1B). As seen in Figure 7A, the ring at each location is designed to present information on how far in the future the events in one location will occur. The color assignment of the ring segments ranges from dark blue (tomorrow) to off-white (more than a month away). The number of event locations shown on the map is determined by the zoom level. As shown in Figure 7, the locations are shown at an aggregated level for an overview when zoomed out, while individual event locations will be shown upon zooming into a particular region. When clicking on an individual location, the tweets regarding the event occurring in this location will be shown (Figure 7B). The Map View is coordinated with other views in the interface, providing information based on selections made on date, keywords, or social networks. Details on the coordination between views will be described in Section 6.6. Overall, the Map view provides information on where the future events will occur, thus permitting users to explore future events based on location(s) of interest.

6.3 Word Cloud: What are the Events About?

To allow users to get an overview of what the future events are about without having to read the tweets, the CrystalBall visual interface includes a Word Cloud View showing a set of keywords summarizing the events (Figure 1C). To extract keywords to describe future events, we employ the Sparse Additive Generative Models (SAGE) that outperform the Dirichlet-multinomial generative models in both predictive accuracy and process speed because the SAGE learns more robust and simpler topics by focusing on high-frequency words with accurate counts [15]. We use SAGE to extract a number of keywords for each event. We rank the event keywords based on their weight in the probabilistic distribution. In the word cloud view, the size of each keyword indicates its occurrence in all events. The user can zoom in to improve the readability of the small size keywords.

The view is highly interactive and coordinated with other views. When a user selects a date or multiple events in the Future Events Overview, the Word Cloud View shows the keywords of the selected date or event accordingly. Keywords are highlighted in red when the user selects an event on the Event List View or on the Social Network View. In addition, the view filters keywords based on the current map extent. When the user zooms in to any particular location on the map view, the view shows keywords of the events that are in the current map extent, to help the user find future events based on her region of interest.

6.4 Social Network: Who Posted Future Event Related Information?

To present relationship between the Twitter users who posted information related to future events, we construct a social network. The network is generated based on retweet (RT) and user mentions (@) information extracted from the tweets. The network presents how information related to future events flows among the Twitter users. The network has two modes: 1) as an overview, the network shows the numbers of events of location-time pairs (event mode, Figure 1D). In this mode, each node represents an event, with the links between nodes denoting two events sharing the same location and time. The number in a node represents the number of tweets of the event. 2) To provide detailed social network information for selected events/network clusters, the network shows relationships between the Twitter users of a selected location-time pair (Twitter user mode, Figure 9 left). Since an overview of one day's events could contain on average 150,000 tweets, the two-mode social network design provides flexibility of providing overviews on the event-level without too much detail on the retweet relationships.

We assign colors to each node based on the TFF ratios: 0 (Twitter accounts have no followers, possibly bots), less than 1 (bots or light users), less than 2 (normal users), less than 10 (popular users), and bigger than 10 (celebrities or news channels). The legend in the Social Network view explains the categories and color assignment. The links connecting the nodes denote either retweet or mention relationship. Based on our observation, each cluster in the social network is usually comprised of retweets about one future event.

At a glance, the Social Network view provides clusters of different sizes indicating the "popularity" of various future events. Hovering mouse over a node displays a tooltip showing the user name and the tweet. At the same time, other users that tweeted about the same event will be highlighted in the network, and the aspects regarding when, where, and what about the event will be highlighted in other views.

6.5 Details on Demand

To make sense of a future event, one usually needs to peruse the tweets in addition to the event characteristics presented in the visual interface. The tweets may provide external links that contain more information regarding the future event. To get to the tweets, one can click on the "Twitter" icon on the top right corner of the interface. A sliding panel will be brought up to show the set of tweets based on the current selection in the aforementioned four views (Figure 1E).

6.6 Interaction and View Coordination

The CrystalBall interface provides rich user interactions to support the exploration and analysis of future events from multiple angles. The views are coordinated to show relevant information based on users' actions within the interface. Users could begin exploration or analysis from any view in the interface.

Exploring future events starting with time: One can start the analysis of events from the Event Calendar. Hovering the mouse over an event (circle) in the calendar will highlight the corresponding event location in the Map View, the keywords related to the

event in the Word Cloud, as well as in the Social network. A user can also select events that may occur on a particular day by clicking on the future date, or explore linked events that share keywords or location by clicking on the links in the Event Calendar. The Map, Word Cloud, and Social Network views will change accordingly based on the selection of a subset of events in the Event Calendar.

Exploring future events starting with location: One can also enter the exploration of future events from the Map View. The Map View will show more events (as opposed to aggregated information) when zooming into a region of interest. When zooming in and out in the Map View, the keywords will be filtered based on the map extent. In addition, selecting an event on the map will update all other views to show detailed information regarding the particular future event.

Exploring future events starting with keywords: One can start the analysis from the Word Cloud View. Hovering mouse over a keyword in the Word Cloud View updates all other views in the CrystalBall to show all related events: the timeline view highlights the linked events; the Social Network View highlights all users who posted related tweets; and the map view shows all related geolocations. When the user selects a keyword, then the time line view shows related events in details in the event list view.

Searching for specific types of events: In addition to exploratory analysis of future events, one may want to look for events that are related to certain keywords. CrystalBall supports such focused investigation by providing a search function in order to identify events related to specific keywords and/or locations. For example, if one is interested in music events that will occur in the future, a keyword such as the name of a band can be used to identify such events. To initiate a search, one can click on the icon on the top right corner of the interface. A text search box will be shown to accept keyword input. All views in CrystalBall will be updated to show event results that are relevant to the query.

7 Case Studies

In this section, we report the findings as well as the strategies users have employed to arrive at the findings when using CrystalBall. We refer to the user's processes of exploration in CrystalBall as their strategies. The initial strategy involved entering the analysis from any one of the 4W's and then leveraging the coordinated views for information regarding the other 3Ws of the future events.

7.1 Case Study of Social Unrest in Multi-Day Protests

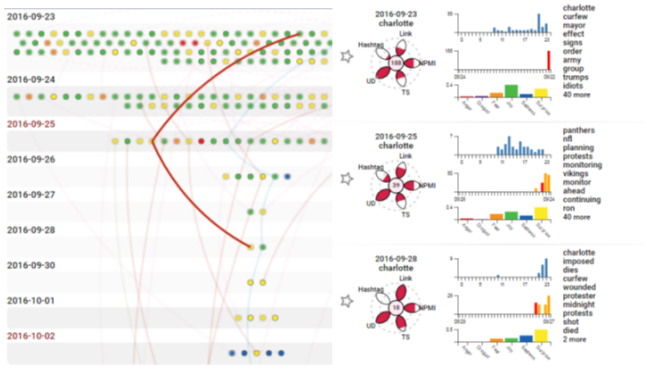

Figure 8: Several future events detected in CrystalBall following the most violent and disruptive night of protests in Charlotte, N.C. on September 22, 2016.

In this case study, we use CrystalBall to explore week-long events relating to the September 2016 protests in Charlotte, N.C. The protests were sparked by the police shooting of Keith Lamont Scott around 4:00 p.m. on Tuesday, September 20, 2016. While the initial protest was spontaneous without much organizing activities on Twitter, CrystalBall can be used to track and identify (future) follow-up events that are related to the shooting. Within hours of the shooting, a crowd grew near the Northeast Charlotte apartment complex and transitioned from peaceful to combative by nightfall. That night as the protesters turned violent, riot police were forced to deploy tear gas multiple times to disperse the crowds in multiple locations. The next evening, Wednesday September 21, larger crowds of protesters converged to the city center. Around 8:00 p.m. the protests turned chaotic after the point-blank shooting of a civilian protester led to widespread vandalism followed by the governor declaring a state of emergency at 11:00 p.m. and deploying National Guard troops.

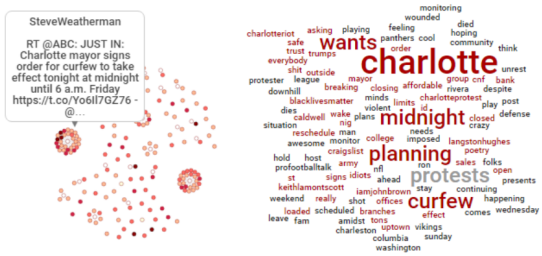

Figure 9: The Word Cloud and Social Network Views indicating the upcoming curfew in Charlotte, N.C. imposed at midnight on September 23, 2016.

Following the first two days of protests, we want to explore how CrystalBall can investigate multiple future events through a variety of views and interactions.

Exploring Days after the Initial Protests Setting the date to September 22, nearly 36 hours after the initial shooting, a user started the exploration to investigate follow-up events using CrystalBall. The user first zoomed in the Map View to focus on the Charlotte region. By hovering on location Charlotte in the Map View, the user discovered several events highlighted in the Event Calendar View that may occur on different dates. The user then switched to the Event Calendar View to investigate the events further. As seen in the left of Figure 8, a solid line connecting three events that are going to occur in Charlotte can be found in CrystalBall. Next, the user clicked on the solid line to examine details of the three events in the Event List View (Figure 8 right), the user now sees that the future events are related to the upcoming city curfew, postponed music shows, and the potential cancellation of the Carolina Panthers' upcoming NFL home game. Out of these events, the user was able to identify and drill down on the largest (by tweet volume) surrounding the city curfew that started midnight of the next morning (September 23). The user then moved to the social network to detect large clusters of tweets that relayed information and responses to the upcoming curfew (Figure 9).

Curious if this trend continued to the next day, the user changed the date to September 23 and found multiple new future events. Using the Emotionbar to scan for polarizing emotions, the user spotted an event with high disgust and fear regarding "charlotte", "crooked", "hillary", and "blacklivesmatter" (Figure 10). Interested in learning more, the user transitioned to the Social Network to discover a cluster of retweets by presidential candidate Donald Trump. Specifically, in his tweet, Trump mocked Hillary Clinton's "bad judgment" to cancel her upcoming trip to Charlotte on September 24[2]. After the analysis, the user commented that he appreciates CrystalBall's capability to enable him to identify such relevant but

[2]https://twitter.com/realdonaldtrump/status/779503099492831233

unexpected event.

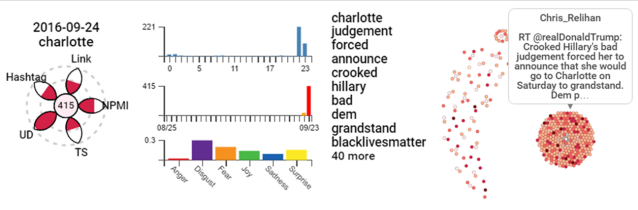

Figure 10: A related but unexpected event for Hillary Clinton's cancelled campaign appearance in Charlotte, N.C. following the aftermath of the Charlotte protests.

7.2 From exploration to focused analysis (the PEGIDA UK movement)

This case study illustrates how CrystalBall enables users to explore and discover possible leads, and then perform more focused investigation based on the initial discovery.

When performing exploratory examination of the events discovered from June 1, 2016 within CrystalBall, the user noticed a future event in Rotherham, UK that may occur on June 4. The user came across this event by hovering mouse over different events in the Calendar View and glancing at the location and keywords related to each event. When clicking on the event in the Calendar View, multiple retweets on "PEGIDA march Rotherham this Saturday" are shown in the tweet panel. Being unfamiliar with PEGIDA, the user searched for more information on the web on the entity PEGIDA, and discovered that it is a political movement stands for "Patriotic Europeans Against the Islamisation of the West", which has offshoots in multiple countries including Germany and the UK.

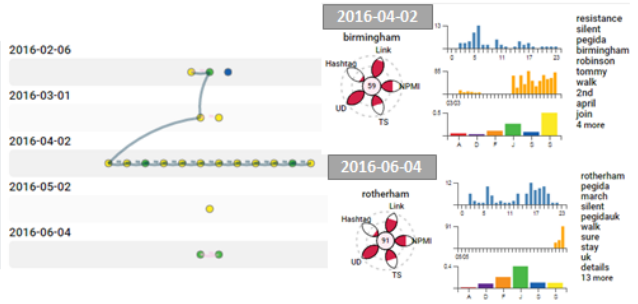

Figure 11: The search result of keyword "PEGIDA". The 2nd PEGIDA UK movement was in Birmingham, England on April 2, 2016. The 3rd PEGIDA UK march was in Rotherham on June 4, 2016.

The information discovered so far raises the user's interest on the entity PEGIDA. Following this lead, the user then leveraged the **search function** within CrystalBall by inputting "PEGIDA" as a keyword. As result of the search, multiple events related to PEGIDA were shown in the CrystalBall interface (Figure 11). The user found that from February 6 to June 2016, there is at least one event related to PEGIDA at the beginning of each month. The events include 4 marches that are organized by PEGIDA for different causes.

CrystalBall enables the identification of such future events ranging from one week to one day ahead. To verify if the marches did occur as seen in CrystalBall, the user performed validation analysis of the marches after the dates they were forecasted to occur by performing web searches of PEGIDA and the day of the events. The user verified that all the planned marches have indeed taken place, with each march participated by hundreds of protesters, as reported by news media including Reuters, the Guardian, and Breitbart. Note that the news articles were published to report the protests/marches, while CrystalBall enables users to discover the events while they were being planned but haven't yet occurred.

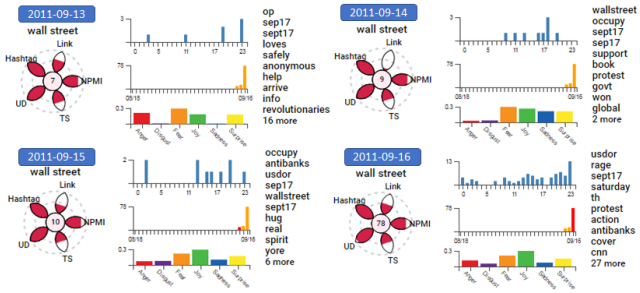

Figure 12: The Occupy Wall Street event protest on September 17 is identified in the CrystalBall interface from September 11-14's data.

8 Validation Studies

8.1 The Occupy Movement

One way to evaluate CrystalBall would be to test whether it can discover a well-known past event prior to the event day from historical data. Therefore, in this case study, we use historical 1% Twitter data sample collected in 2011 to analyze the Occupy Movement that started in New York City on September 17, 2011. We analyzed the tweets posted between September 9 to 16 to see whether we can identify the precursors to the first protest. Around 123,000 tweets were analyzed by CrystalBall, out of which only 120 tweets contain hashtag "occupy" based on our preliminary data analysis. The dataset serves as a good benchmark to evaluate whether CrystalBall is able to detect very small signals from overwhelming noise.

We performed exploratory analysis of this dataset with CrystalBall just like one would with streaming Twitter data. Starting from September 11 to 16, we can consistently discover in CrystalBall from each day's data that there is a future protest event going to occur on September 17, 2011. Figure 12 shows the results of the upcoming occupy event identified between September 13 to 16 (the dates on when the future event were discovered are presented in blue rectangles). The earliest date for which CrystalBall was able to identify the upcoming event was on September 11. Although the precursor signal is very small as there are only single digit tweets related to the upcoming event, CrystalBall is still able to identify them as possible future events. The signal got consistently stronger over the next several days. The keywords posted each day regarding the upcoming event changed over time. Keywords summarizing tweets published on September 12 contain the keyword "anonymous", which is a group that participated in organizing the Occupy Wall Street movement. The hashtag "#usdor" is among the keywords summarizing tweets posted on September 15. The hashtag refers to an organization "US Day of Rage" [4] that also helped to organize the first #Sep17 protest.

In summary, in this validation study, we assessed whether CrystalBall can discover a known event prior to when the event actually occurred. Analyzing Twitter data posted a week before the Occupy Wall Street movement, CrystalBall was able to identify the future event and permitted the analysis of organizing activities. Our retrospective study showed that there were periodic mentions of preparation for the Occupy Wall Street event for several weeks before the event [12]. Thus if we had been monitoring streaming Tweets during this time, it is possible that CrystalBall could have enabled the discovery and characterization of the event earlier.

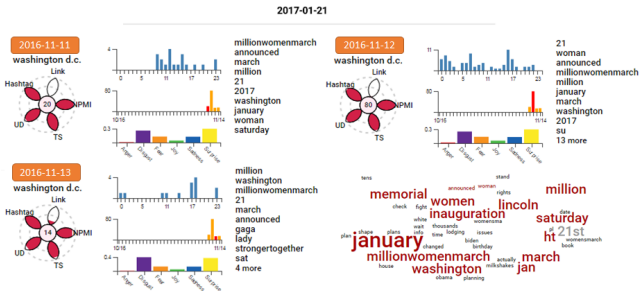

Figure 13: The Women's March on January 21, 2017 first appeared in the CrystalBall interface on November 11, 2016.

8.2 Tracing its Roots: Women's March on Washington

In this case study, we use CrystalBall to monitor live Twitter stream day by day to identify large scale social protests in a noisy environment. Specifically, we use the recent January 2017 Women's March on Washington as an example and present information on the earliest detection of the protest following the November 8, 2016 United States presidential election. More importantly, by identifying the first detection we can then use CrystalBall's multiple date features to track the development of the movement up until the event itself.

As a user was exploring the events identified from November 11's data, the day after the election. The user started with the Event Calendar View and found a flurry of post-election events. After identifying multiple large, global locations (e.g., India, U.K., Canada, London, Israel), she found "Washington" spanning across multiple dates. Selecting that location, she then found multiple future events up to Inauguration Day (January 20). The keywords describing the future events include "millionwomenmarch", "washington", "january", etc. Examining tweets related to these events revealed that the organizational activities of Women's March occurred soon after the election. CrystalBall enables the identification of early signal as the event is being organized.

9 Limitations

In this section, we discuss three limitations of CrystalBall along with possible solutions.

First, CrystalBall cannot extract a comprehensive list of future events from daily streaming Twitter data. Instead, it focuses on future events that will occur at a physical location. The modeling of future events based on location-time pairs requires the future events to be mentioned on Twitter in a specific way, namely with a location and a date reference (such as tomorrow, this weekend, next Thursday, etc.). Although CrystalBall can already extract on average close to 300 future events per day, it will miss events that are not being mentioned explicitly by time and location.

A second limitation of our approach is related to falsely identifying past events as future events. Some tweets use a present tense especially when they refer to news headlines. Traditionally news headlines are written in a present tense to attract the readers' attention. Tweets containing the news headlines regarding past events will likely to be falsely recognized as future events. Furthermore, we observed retweets of an original tweet regarding a future event are posted after a fair amount of time (e.g. several hours and possibly days). Since our date reference relies on the time stamp of the individual tweets, the "late" retweets would create wrongly resolved dates. An example would be late retweets of "tomorrow", the resolved date would be one day after the retweet was posted.

The third limitation stems from the time and location extraction.As previous researchers noted, the named entity extraction (NER) algorithms are sometimes trained and evaluated by annotated datasets created from news or Wikipedia [27]. However, the noisy nature of Twitter data would lead to significant accuracy reduction of the NER algorithms [27]. In addition, since tweets are posted from everywhere in the world, some locations are inherently ambiguous while others are duplicates (e.g., Columbia, South Carolina and Columbia, South America). However, for the aforementioned cases, the highly interactive and exploratory interface of CrystalBall will be useful for users to run through several possible events quickly (including their links). In future versions, we could develop feedback mechanisms to allow users to mark non-future or mis-identified events. Such a method to incorporate domain expertise and knowledge into systems represents one of three broader themes of future work.

10 Future Work

There are three promising areas of future work for event detection visual systems.

First, the rise of open-source distributed computing engines (e.g., Apache Spark [35]) can facilitate a wider suite of machine learning algorithms for future systems. These tools can enable scale and increased the speed of processing through real-time processing engines (e.g., Apache Kafka [20]). Moreover, advances in natural language processing like word embedding models could provide innovative visualizations for varying spatial-temporal patterns within Twitter content (e.g., ESTEEM model [6]).

Second, the integration of multiple data streams along with additional state data (e.g., one week moving average of events per location) can provide a more holistic perspective to event detection. One weakness of CrystalBall is that it only includes data from a single social media platform; alternatively, future platforms that employ more sophisticated processing engines (e.g., Spark) could integrate multiple data streams from other social media platforms (e.g., Facebook, Instagram) along with non-social media data such as news articles, Google trends, and Wikipedia.

Last but not least, there is a need for evaluating existing and developing new visualization techniques for the future analysis. The best way to convey events that are inherently spatial and temporal are yet to be explored. Design studies will be conducted with target users to develop and evaluate ways to represent the attributes that characterize future events and more importantly the relationship between events.

11 Conclusion

In this paper, we presented CrystalBall, a visual analytics system that identifies, ranks, characterizes, and visually presents future events from streaming Twitter data. In addition, we introduced seven analysis measures for the future event identification and characterization in order to provide rich and detailed information of the future events. Several case studies were presented to demonstrate the efficacy of the CrystalBall interface for future event discovery and analysis.

Acknowledgments

This work is supported in part by the Army Research Office under contract number W911NF-13-1-0083 and the U.S. Department of Homeland Securitys VACCINE Center under award no. 2009-ST-061-CI0002.

REFERENCES

[1] Bank Transfer Day. https://www.facebook.com/Nov.Fifth/. Accessed: 2016-03-29.

[2] Occupy Wall Street. http://occupywallst.org/. Accessed: 2016-03-29.

[3] The Streaming APIs. https://dev.twitter.com/streaming/overview. Accessed: 2016-03-29.

[4] US Day of Rage. https://www.facebook.com/US-Day-of-Rage-199185230105826/info/?tab=page_info. Accessed: 2016-03-29.

[5] H. Abdelhaq, C. Sengstock, and M. Gertz. Eventweet: Online localized event detection from twitter. *Proceedings of the VLDB Endowment*, 6(12):1326–1329, 2013.

[6] D. Arendt and S. Volkova. Esteem: A novel framework for qualitatively evaluating and visualizing spatiotemporal embeddings in social media. In *Proceedings of the Association for Computational Linguistics*. ACL, 2017.

[7] 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*, pages 291–300. ACM, 2010.

[8] G. Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40, 2009.

[9] A. X. Chang and C. D. Manning. Sutime: A library for recognizing and normalizing time expressions. In *In LREC*, 2012.

[10] I. Cho, W. Dou, D. X. Wang, E. Sauda, and W. Ribarsky. VAiRoma: A visual analytics system for making sense of places, times, and events in roman history. *Visualization and Computer Graphics, IEEE Transactions on*, 22(1):210–219, Jan 2016.

[11] R. D'Andrade and M. Egan. The colors of emotion. *American ethnologist*, 1(1):49–63, 1974.

[12] 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. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[13] W. Dou, X. Wang, W. Ribarsky, and M. Zhou. Event detection in social media data. In *IEEE VisWeek Workshop on Interactive Visual Text Analytics-Task Driven Analytics of Social Media Content*, pages 971–980, 2012.

[14] 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 *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 93–102. IEEE, 2012.

[15] J. Eisenstein, A. Ahmed, and E. P. Xing. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 1041–1048, USA, 2011. Omnipress.

[16] L. C. Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978.

[17] H. Guo, J. Huang, and D. H. Laidlaw. Representing uncertainty in graph edges: an evaluation of paired visual variables. *IEEE transactions on visualization and computer graphics*, 21(10):1173–1186, 2015.

[18] H. Huang, A. Zubiaga, H. Ji, H. Deng, D. Wang, H. K. Le, T. F. Abdelzaher, J. Han, A. Leung, J. P. Hancock, et al. Tweet ranking based on heterogeneous networks. In *COLING*, pages 1239–1256, 2012.

[19] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.

[20] J. Kreps, N. Narkhede, J. Rao, et al. Kafka: A distributed messaging system for log processing. In *Proceedings of the NetDB*, pages 1–7, 2011.

[21] M. Krstajić, E. Bertini, and D. A. Keim. Cloudlines: Compact display of event episodes in multiple time-series. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2432–2439, 2011.

[22] V. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet Physics Doklady*, volume 10, pages 707–710, 1966.

[23] D. Luo, J. Yang, M. Krstajić, W. Ribarsky, and D. A. Keim. Eventriver: Visually exploring text collections with temporal references. *Visualization and Computer Graphics, IEEE Transactions on*, 18(1):93–105, 2012.

[24] OpenStreetMap. Nominatim. http://wiki.openstreetmap.org/wiki/Nominatim. Accessed: 2016-03-29.

[25] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. Association for Computational Linguistics, 2013.

[26] K. Radinsky and E. Horvitz. Mining the web to predict future events. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 255–264. ACM, 2013.

[27] A. Ritter, S. Clark, O. Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics, 2011.

[28] 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.

[29] A. Sadilek, H. A. Kautz, and V. Silenzio. Predicting disease transmission from geo-tagged micro-blog data. In *AAAI*, 2012.

[30] D. Thom, H. Bosch, S. Koch, M. Wörner, and T. Ertl. Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages. In *Pacific visualization symposium (PacificVis), 2012 IEEE*, pages 41–48. IEEE, 2012.

[31] S. Volkova, Y. Bachrach, M. Armstrong, and V. Sharma. Inferring latent user properties from texts published in social media. In *AAAI*, pages 4296–4297, 2015.

[32] F. Wanner, A. Stoffel, D. Jäckle, B. C. Kwon, A. Weiler, D. A. Keim, K. E. Isaacs, A. Giménez, I. Jusufi, T. Gamblin, et al. State-of-the-art report of visual analysis for event detection in text data streams. *Computer Graphics Forum*, 33(3), 2014.

[33] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu. Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In *Proceedings of the 21st international conference on World Wide Web*, pages 71–80. ACM, 2012.

[34] Y. Yang, T. Pierce, and J. Carbonell. A study of retrospective and online event detection. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 28–36. ACM, 1998.

[35] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin, et al. Apache spark: A unified engine for big data processing. *Communications of the ACM*, 59(11):56–65, 2016.