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

Category: Research

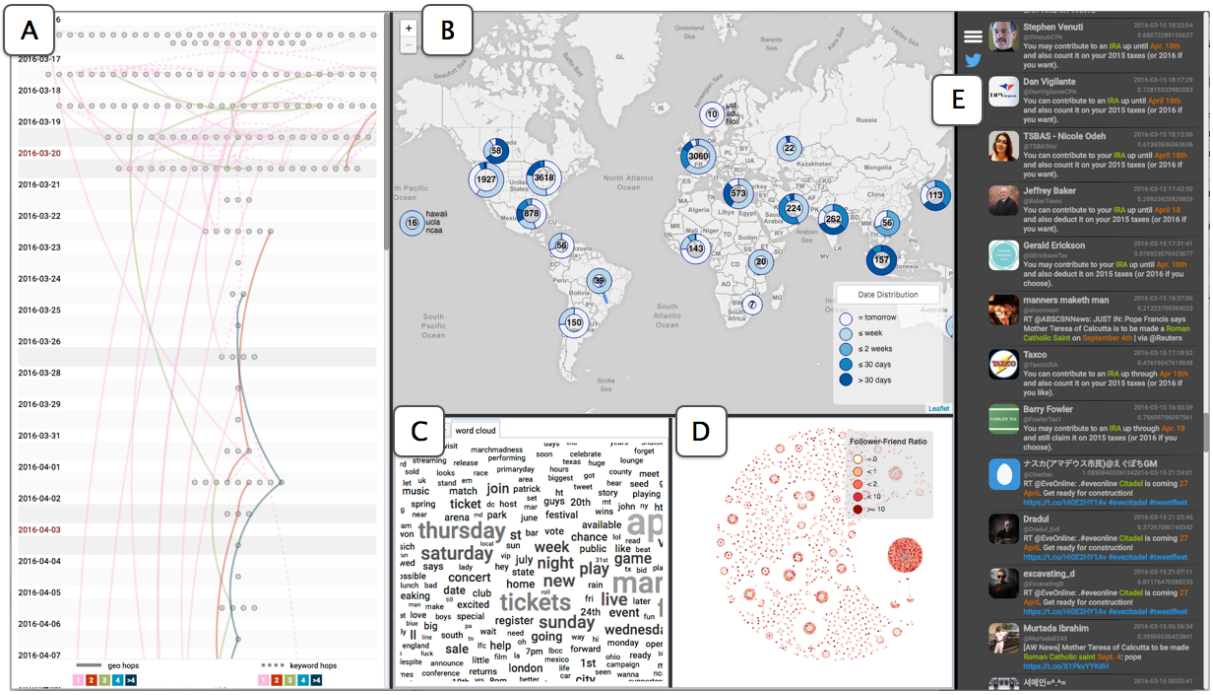

Fig. 1. CrytalBall interface. The interface is comprised of 4 main views: A) calendar view, B) map view, C) word cloud view, and D) social network view. The E) tweet panel is shown on demand.

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 on analyzing events derived from Twitter data.

Index Terms—Social media analysis, Event detection and analysis, visual analytics

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 [15]. And much work has been devoted to identifying past and ongoing events from social media [28].

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 [2], the Occupy Movement [5], as well as future function shutdown related events due to certain conditions such as inclement weather. Many of these events develop spontaneously and may not be known to public safety personnel, planners, facilities managers, or others who might effectively use the information. 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 [5]. Therefore, in addition to information regarding ongoing events, social media posts also capture information related to events that are likely to occur in the future. 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.

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. [15] 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 events by detecting a "change" in volume. Although this definition of events is general, it does not apply 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 highly unlikely that future events related posts would cause a change in the volume of social media streams. Therefore, in contrast to previous research on past and real-time event detection [26, 9, 8], 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.

The problem of identifying future events calls for new ways of detecting relevant information. Although posts regarding future events may not be significant in terms volume, they are likely to refer to two key attributes that characterize a future event: a future time and a location where the event will occur. Therefore, we define a future event as:

*"A potential occurrence that is associated with a **location** and a **date/time span** in the future."*

The location and time are two attributes that define a future event. In addition to the *where* and *when*, attributes including keywords and social network provide information regarding the *who* and *what* about future events. Moreover, measures such as the informativeness of the tweets associated with a future event and the credibility of the Twitter users provide information regarding the quality of the future event predictions. Our approach takes into consideration all of the aforementioned attributes when modeling future events from Twitter streams.

1.2 Introducing CrystalBall

In this paper, we describe a visual analytics system, CrystalBall, that identifies, ranks, and visually presents future events. Our system provides event-oriented visualization based on automated algorithms, while permitting interactive exploration of events that are likely to occur in the future. Our work differentiates from past research in several aspects. First, much of prior work has focused on detecting past and ongoing events from textual sources [14, 22, 15, 20, 12] 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 certain types of future events.

We highlight four contributions of CrystalBall:

1. CrystalBall includes a general model to discover future events from Twitter messages.
2. CrystalBall integrates several metrics to characterize and rank the identified future events.
3. CrystalBall includes a new interactive visual interface that is tightly integrated with the modeling for exploring and making sense of future events.
4. We provide several case studies and a validation study 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, adn 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

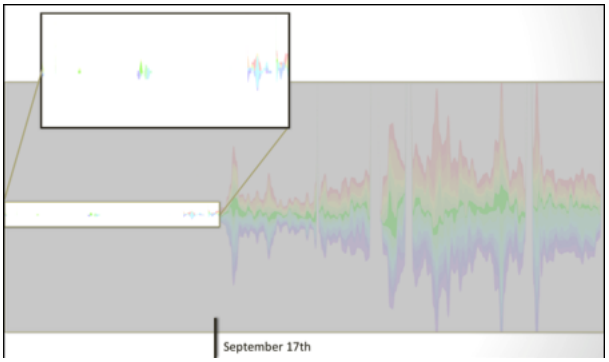

Fig. 2. The overall topic trends of the Occupy movement [13]. Precursor to the movement is enlarged and highlighted. The precusor is comprised of tweets related to organizing activities posted before Sept 17, 2011.

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 empowers individuals to initiate events that may spread locally or nationally. One example is the Bank Transfer Day event that was started in by a gallery owner and participated by an estimated 1 million citizens across the United States who share the same frustrations with big bank services [2].

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 [5]. Based on our data-driven analysis of the rise and fall of the movement [13], we discovered a precursor signaling organizing information well ahead of the official start date of the Occupy Movement. As shown in Figure 2, 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 to look for, it is highly unlikely that one would detect the organizing activities on Twitter prior to September 17. The small precursors to the

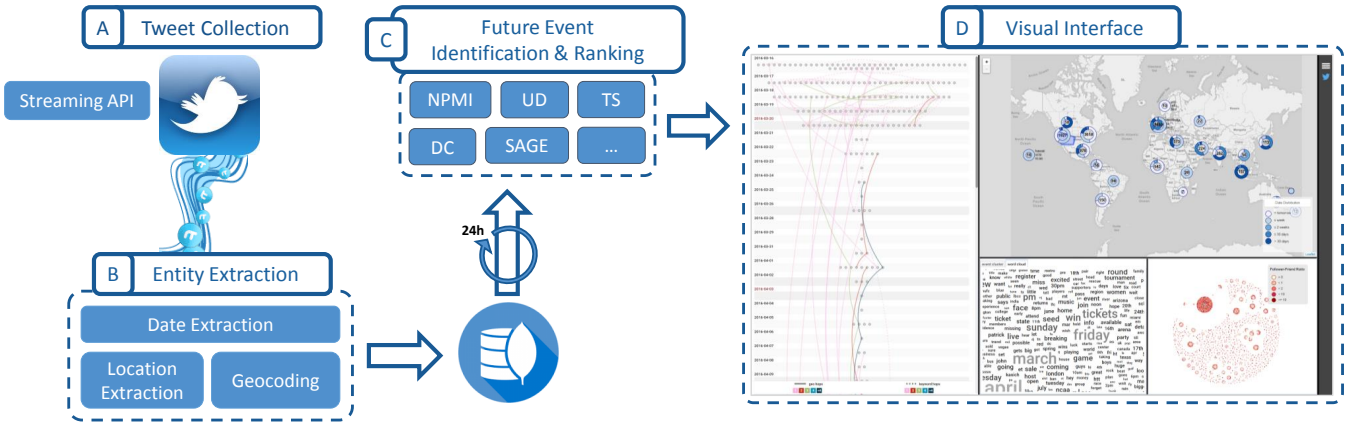

Fig. 3. System pipeline of CrystalBall: A) streaming Twitter data collection, B) named entity recognition and geocoding, C) future event identification and ranking, and D) interactive visual interface.

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 event that is going to occur on September 17, individual 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. [30] 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. [8] proposed methods to detect "local" events in real-time from tweets by measuring the entropy of the tweets spatial signature. Becker et al. [9] also extracted events from social media by learning similarity metrics that enable online clustering of events. Along the same line, Ritter et al. [26] presented TWICAL, an open-domain event extraction and categorization system for Twitter. Compared to previous work, Ritter et al. coined the term "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. [25] 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 [27]. 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, 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 [14]. 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. [28] 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 [20], 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 [15], 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 [22], 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 3. CrystalBall integrates multiple components, including data collection from the Twitter Streaming API [6](Figure 3A), entity extraction(Figure 3B), future event identification and ranking (Figure 3B), and an interactive visual interface(Figure 3D). 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 3A). Batches of tweets are then sent to the entity extraction component (Figure 3B), which involves date extraction and mapping, location extraction, and geocoding. We employ Stanford SUTIME [11] for the date extraction and TweetNLP [24] for the location extraction. For the geocoding,

we use OpenStreetMap Nominatim [23]. 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. 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 3C). 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 Identifying 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)[10] 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 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 co-occurrence of a location and a time differs from what we would expect it to be on the basis of the probabilities of the individual occurrences and the assumption of independence. A completely uncorrelated location-time pair would receive a PMI of 0 [10]. 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 tighter correlation between the when and where aspects of a future event. Therefore, we save location-time pairs with positive NPMI scores for further analysis.

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 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 6 additional measures that characterize the future events. From these a linear regression model is built to integrate all measures into a unified metric for ranking the events. Previous work has ranked tweets based on informativeness and trustworthiness of the content [19]. Our 6 additional measures were developed by 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 4 measures that are related to the informativeness of the tweets regarding individual events. We build on findings from prior work [19] and propose measures including link ratio, hashtag ratio, user credibility, and user diversity.

Link and hashtag ratio As found in previous work [19], 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" [19]. 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 a 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 [29]. 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

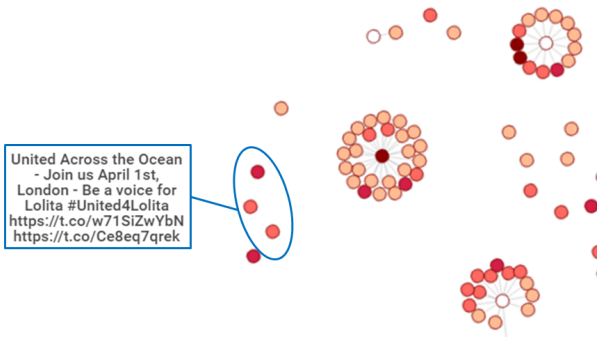

Fig. 4. Tweets from 3 twitter users in the blue oval are identical but with no retweet and mention information. Therefore the nodes are not connected to each other.

of unique users divided by the total number of tweets related to one future event $\frac{\#uniqueusers}{\#tweets}$.

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 [18]. 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 The degree centrality of the social network linked to a possible future event reflects how connected the tweets are. However, when we analyzed the tweeting behavior centered around an event, 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 (an example can be seen in Figure 4). 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 [21].

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 rule out any types of events before users have a chance to see them since different users may be interested in a 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.

5.2.3 Ranking future events

The 7 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.

To rank the identified future events, we build a multiple linear regression model to evaluate the relationship between the explanatory variables and a response variable denoting the overall quality of a future event.

$$y = \beta_0 + \beta_1 NPMI + \beta_2 LR + \beta_3 HR + \beta_4 UD + \beta_5 DC + \beta_6 TS + \beta_7 TFF$$

To train the linear regression model, we develop a labeled dataset comprised of extracted future events during 3 days (approx. 600). For the training dataset, two coders went through the extracted future events and labeled the events as either true or false. The false labels are given to time-location pairs mainly associated with tweets from bots or in cases where there does not seem to be a future event related to the time-location pair. To keep the events of diverse types, the coders were instructed to avoid bias towards certain types of events during the labeling process. Instead the goal is to rule out identified future events that contain larger number of unreliable tweets (such as from a bot).

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 an interactive visual 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 [15], our visual interface includes 4 main views, with each centered around one of the 4Ws. The interface is mainly developed using D3.js [3], Leaflet [4], PHP and Javascript.

6.1 Event Calendar: When will the Events Occur?

The Event Calendar presents a list of future events that have been identified and ranked. The Event Calendar shows an overview of events (future event overview) by default. More details about the selected subset of events (event list) will be shown upon user interaction.

6.1.1 Future events overview

The future events overview (Figure 1A) is designed to present the identified events and the connections among the 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 weekend days showing in dark red. 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. Multiple future events may be related. We categorize the relatedness by events sharing keywords or location. To visually represent the relationship among events, links are drawn to connect events that may be related. Links that connect two or more events indicate two possibilities:

1. A link connecting two events that share the *same location* is drawn as a solid line. Hovering the mouse over the link will invoke a pop-up window showing the future events occurring in the same location (Figure 5A).

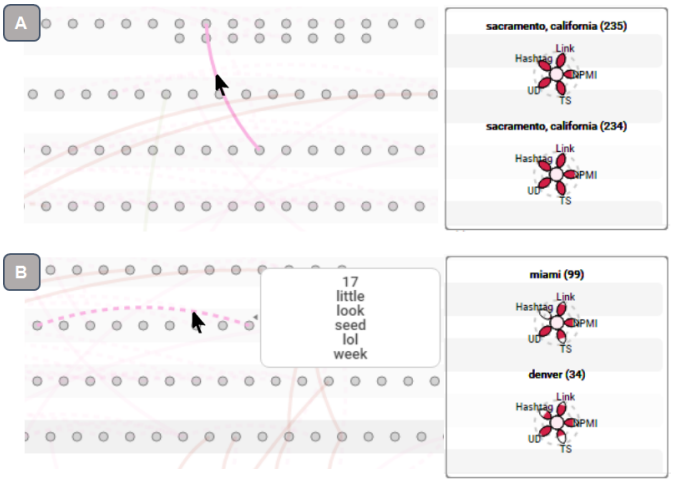

Fig. 5. Future events overview. A) Solid lines indicate events that share the same locations, while B) dotted lines indicate events sharing the same keywords.

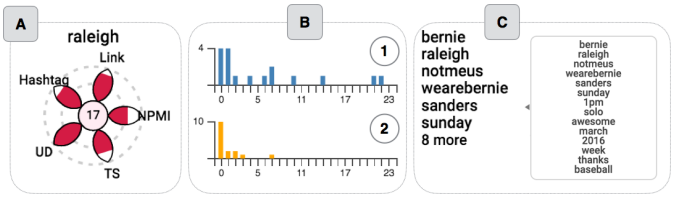

Fig. 6. Event List view. A) the flower glyph visualizes 5 measures of the future event (LR, HR, NPMI, UD and TS) with the number of tweets in the center. B) The top bar chart shows distribution of the tweet posting time and the bottom chart shows inter arrival time of the tweets. C) Keyword summary of a future event.

2. A link connecting two events that share *2 or more keywords* is drawn as a dotted line. Hovering the mouse over the link will trigger a tooltip displaying the shared keywords (Figure 5B).

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 1A, a distinct color is assigned based on the number of hops, with one hop connecting two events. Overall, the Future Event Overview (Figure 5) 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

The Event List view is designed to present the characteristics of the selected events. Each row shows the characteristics of one future event identified by CrystalBall. As seen in Figure 6, each row consists of 4 elements: a flower glyph on the left, two timeline bar charts in the middle, and an event keyword list on the right. The flower glyph is designed to present values of 5 measures of a future event, with the number of tweets shown in the center of the flower. The 5 out of 7 measures introduced in section 5 include link and hashtag ratios, the NPMI, user diversity, and tweet similarity. The other two measures on user credibility and network centrality are portrayed in the social network view, which will be introduced in section 6.4. Each petal in the flower glyph is filled based on the value of the corresponding measure. During the design process, we started representing the 5 measures as a star glyph, as shown in Figure 6A. Through an informal evaluation, we found that users want to get a more accurate sense of the actual value of each dimension relative to the maximum value. Therefore we designed the flower glyph, with the maximum value denoted by the overall petal shape and the actual value presented by the filled level.

For each event, two timeline bar charts are shown to present information regarding when the tweets related to one future event were posted and the inter-arrival time of the tweets (Figure 6B). The posting time of the tweets provide information on when a future event was mentioned, while the inter-arrival time of the tweets could potentially reveal patterns of different types of events [17]. The list of keywords that summarize an event is shown next to the bar charts (Figure 6C). 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?

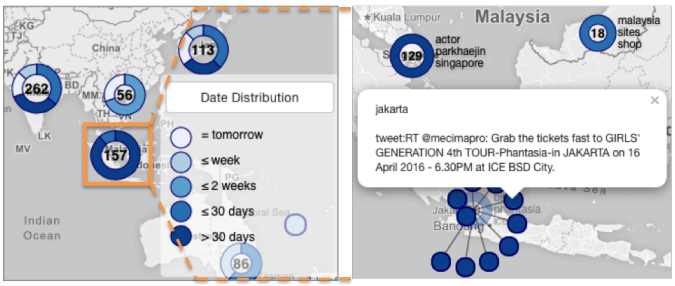

Fig. 7. Left: Overview of future events in the Southeast Asia region. Right: Zooming into Malaysia for detailed events.

In addition to when will the events occur, another important aspect regarding future events is where will they occur. The Map view presents information pertaining to the location of the future events (Figure 1B). As seen in Figure 7 left, 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 off-white (tomorrow) to dark blue (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 7 right). The Map View is coordinated with other views in the interface, providing filtered 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 organized information based 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 the user to get an overview of what the future events are about without reading the actual 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) [16] that outperform the Dirichlet-multinomial generative models in both predictive accuracy and process speed. We use SAGE to extract a number of keywords for each event. We consider the probabilistic distribution of the keywords for each event and keep the keywords that make significant contribution for describing the event. In the word cloud view, the size of each keyword indicates its occurrence in all events. The user can zoom and pan the view 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 an event on the event timeline view, 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 events that are in the current map extent, to help her to find future events on her point 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. In the social network view (Figure 1D), each node represents a twitter user. We assign color to each node based the TFF ratios: 0 (twitter accounts have no follower, possibly bots), less than 1 (bots or light users), less than 2 (normal user), 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 connected to one future event. However there are cases that multiple clusters that are not connected to each other are related to one future event (Figure 9).

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 an individual future event, one usually needs to peruse the tweets in addition to the event characteristics presented in the visualizations. 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 4 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 the tweet 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 the 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 a query. 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

We refer to the user's processes of exploration in CrystalBall as their strategies. In this section, we report the findings as well as the strategies users have employed to arrive at the findings when using CrystalBall. 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 3 Ws of the future events.

7.1 What's going on in locations with multiple future events?

One user started the exploration of future events with the Event Calendar, which presents *when* the events will occur as well as the relationship among future events. The participant conducted the exploration on March 15, 2016, therefore the events shown in CrystalBall are from that day forward. She first perused the events on each of the future dates by hovering the mouse over individual events to see the corresponding information on location, keywords, and social network. She then started to explore the links that connect events sharing either locations or keywords. The participant was in particular interested in links (solid lines) that connected multiple events over different dates, which denoted multiple events that will occur in the same location. She identified two "multi-hop" links that connected "London" and "Chicago" (Figure 8) respectively and examined these links further. The Event List view in the center of Figure 8 shows representative future events that are going to occur in London and Chicago, with one representative tweet placed next to each event. The two events in London are related to demonstrations against animal cruelty. The event that is going to occur on April 1 is a protest supporting retiring an orca and having her reunited with her mother. The other event is a march planned for April 30 to stop lion trophy hunting. Following the links provided in the tweets one can get more information regarding the upcoming events, including how many people are planning to participate and how to get involved.

7.2 Looking for "popular" future events

Another common strategy we observed users employ was to look for salient features in different views. The salient features include large clusters in the Social Network view that denote "popular" future events, or darker rings in the Map view that represent events in the far future.

Upcoming events with lots of discussions on Twitter Figure 9 shows the social network view for future events extracted from tweets posted on March 14, 2016. Hovering the mouse over the large clusters could provide a quick overview of the popular future events that already sparked lots of discussions on Twitter. The largest cluster (Cluster 1 in Figure 9) is comprised of tweets from Justin Bieber fans on the Purpose tour in Sacramento the next day. Another large cluster (Cluster 2) is formed by tweets related to the ongoing presidential campaigns. People expressed that they are going to vote for Ted Cruz in the upcoming primary in Peoria, IL. Cluster 3 shows tweets about "Putin orders start of withdrawal of troops from Syria next Tuesday". Note that although two separate clusters related to the same event are not connected, our analysis consolidates the two clusters since they refer to the same location-time pair. Cluster 4 is comprised of tweets condemning the Murder of Hindu Activist Raju in Mysuru, Karnataka. This group of tweets also contains information regarding the protest that was planned for the next morning. This case study illustrates how

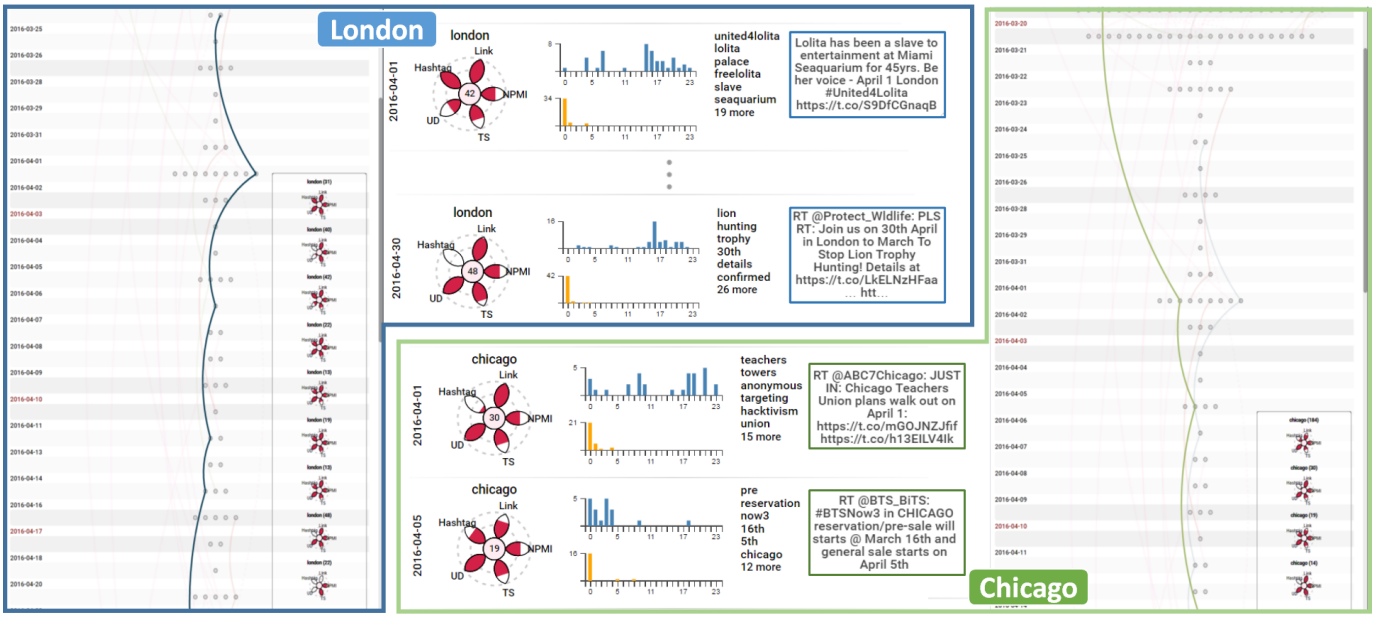

Fig. 8. The figure shows 2 "multi-hop" links connecting London and Chicago respectively. The events are extracted from March 15, 2016 data.

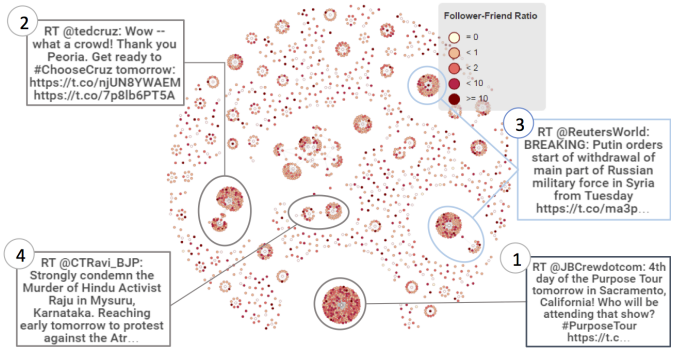

Fig. 9. Investigating future events by examining large clusters.

one may identify future events that receive more attention on Twitter by perusing larger clusters in the Social Network view.

7.3 Identifying and analyzing events of specific types

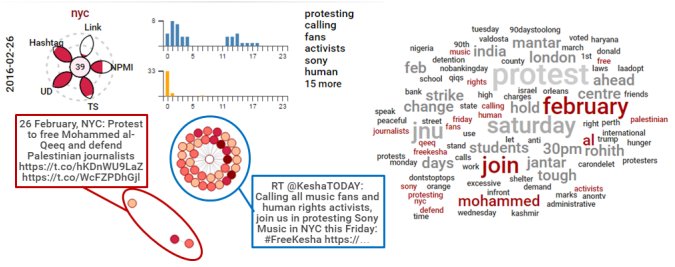

Fig. 10. One potential event in NYC is discovered on Feb 22, 2016 as a result from keyword search "protest". Upon examination, two protests will occur in New York City on Feb 26, 2016: #FreeKesha and the detention of Mohammend al-Qeeq.

CrystalBall provides exploratory event analysis capability to permit the discovery of a diverse range of events. After some initial analysis, users may develop an interest in particular types of events. CrystalBall also permits such focused event analysis. A user can perform directed search using keywords, which may lead to the identification of related events of even smaller scales. To initiate the search, users can click on the icon on the top right corner of the CrystalBall interface to get to a text search box. Users can then type in keyword(s). CrystalBall will first query for tweets containing the keyword(s), and perform "event identification and ranking" in real time on the set of tweets. One search we observed a user perform is looking for "protest" related events. After the user typed in the keyword, the interface presented a list of days with protest-related events. The user then clicked on a recent date to investigate these future events. Figure 10 shows upcoming events in NYC observed on the day of February 22, 2016. The interface shows an event that is going to occur in three days on February 26, 2016. The keywords related to the selected NYC events are highlighted in the word cloud. Upon further examination in the Social Network view, there are two protest-related future events that are being organized to take place in NYC on February 26. The larger cluster refers to #FreeKesha protest at Sony headquarters in support of her attempt to terminate her recording contract. The smaller cluster refers to a protest at the G4S headquarters (a security company employed by Israeli prisons and detention centers) to protest the detention of Mohammed al-Qeeq, a Palestinian journalist held by Israel. In summary, CrystalBall allows focused analysis of certain types of events based on users' interest.

8 The Occupy Movement, A Validation Study

One way to evaluate the CrystalBall interface would be to test whether CrystalBall can discover a well-known past event prior to the event day from historical data. Therefore, in this case study, we use historical twitter data collected in 2011 to analyze the Occupy Movement that started in New York City on September 17, 2011. We first attempted to look for tweet archives that contain tweets published before September 17, 2011. Unfortunately, we found no such data set. Therefore we decided to generate a synthetic dataset that resembles the streaming tweets we currently collect. Luckily, we were able to leverage the data collected during a previous analysis on the Occupy Movement [13]. The dataset comprises tweets containing the hashtag "#occupy". Within the dataset, we ran a search for tweets related to organizational activities on the Occupy Movement before September 17, 2011. The resulting dataset contains 550 tweets that were posted before September 17, 2011. After going through the date and location extraction

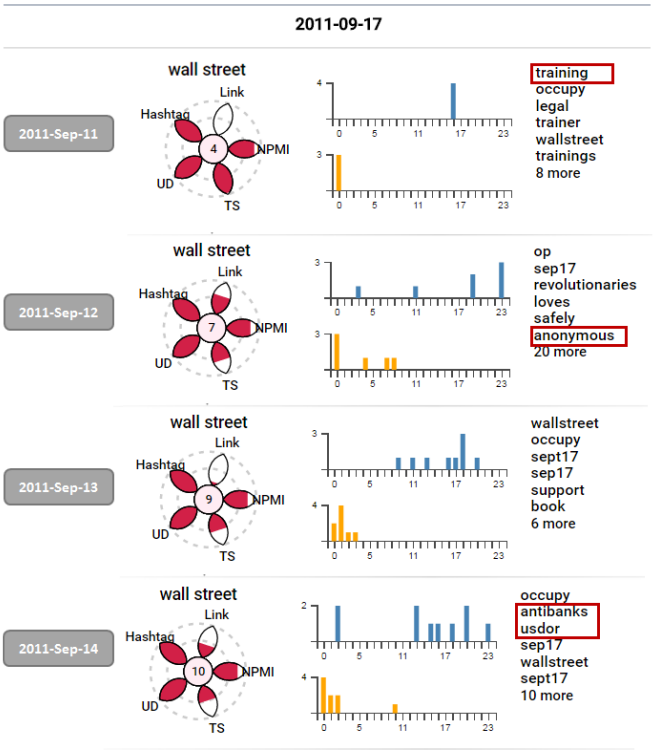

Fig. 11. The Occupy Wall Street event protest on Sept 17 is identified in the CrystalBall interface from Sept 11 to Sept 14's data.

process, we are left with 120 qualified tweets with each tweet containing a location-time pair. These tweets were posted between September 9 and 16.

To add "noise" to make the current dataset resemble the streaming tweets, we merged the "Occupy" tweets with tweet samples we found online [1]. We were able to find large collections of tweets that were posted between September 9 to 16, 2012. We changed the tweets' time stamp to dates from September 9, 2011 to September 16, 2011 accordingly. As a result, 123,000 qualified tweets were added into the dataset. Thus we now have a benchmark collection to test whether CrystalBall could identify a known event prior to the event date.

We performed exploratory analysis of this simulated dataset with CrystalBall just like one would with streaming Twitter data. We did not find any future events related to the Occupy Movement in CrystalBall before September 11, 2011.However, starting from September 11 to 16, we can consistently observe in CrystalBall that there is a protest event that is going to occur on September 17, 2011. Figure 11 shows the results of the upcoming occupy event identified between September 11 to 14 (the dates on when the future event were discovered are presented in gray boxes). The earliest date for which CrystalBall was able to identify the upcoming event was on September 11. The precursor signal is very small as there are only 4 tweets related to the upcoming event. The signal got consistently stronger over the next several days. The keywords posted each day regarding the upcoming event changed over time. Tweets posted on September 11 are related to medic and legal support, as well as communication and facilitator training. 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 14, 2011. The hashtag refers to an organization "US Day of Rage" [7] that also helped to organize the original #Sep17 action to Occupy Wall Street.

In summary, in this validation study, we assessed whether CrystalBall can discover a known event prior to when the event actually occurred. Specific to the Occupy Wall Street movement, CrystalBall was able to identify the future event a week ahead and permitted the analysis of organizing activities. Our initial study showed that there were periodic mentions of preparation for the Occupy event for several weeks before the event [13]. Thus if we had been monitoring streaming Tweets during this time, it is possible that CrystalBall would have enabled the discovery and characterization of the event earlier, even though it did not pick up future event indicators on September 9 and 10.

9 Limitations and Future Work

In this section, we discuss limitations of our approach and potential future work.

We recognize that CrystalBall will by no means 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 100 future events per day, it will miss events that are not being discussed by referring to the time and location describing when and where the events will occur.

Another 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 location and time extraction and geocoding. As previous researches have pointed out, it is very difficult to evaluate the accuracy of named entity extraction results with no ground truth corpus. In addition, since tweets are posted from everywhere in the world, some locations are inherently ambiguous. However, for the aforementioned cases, the highly interactive and exploratory interface of CrystalBall will be useful for users to run through several possible events (including their links) quickly. What we are considering for future work is to allow users to mark non-future or mis-identified events. This is in some sense similar to crowdsourcing in that we will then develop methods to consolidate the marks from multiple users and take actions to remove or correct the false positive event results.

10 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 7 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.

References

[1] Archive team: The twitter stream grab. https://archive.org/details/twitterstream. Accessed: 2016-03-29.
[2] Bank transfer day. https://www.facebook.com/Nov.Fifth/. Accessed: 2016-03-29.
[3] D3.js. https://d3js.org/. Accessed: 2016-03-29.
[4] Leaflet. http://leafletjs.com/. Accessed: 2016-03-29.
[5] Occupy wall street. http://occupywallst.org/. Accessed: 2016-03-29.
[6] The streaming apis. https://dev.twitter.com/streaming/overview. Accessed: 2016-03-29.

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

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

[9] 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.

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] J. Eisenstein, A. Ahmed, and E. P. Xing. Sparse additive generative models of text. 2011.

[17] A. Ferraz Costa, Y. Yamaguchi, A. Juci Machado Traina, C. Traina Jr, and C. Faloutsos. Rsc: Mining and modeling temporal activity in social media. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 269–278. ACM, 2015.

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

[19] 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.

[20] 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.

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

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

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

[24] 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.

[25] 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.

[26] A. Ritter, O. Etzioni, S. Clark, et al. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112. ACM, 2012.

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

[28] 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.

[29] 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.

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