CSE578: DATA VISUALIZATION

Project Proposal:

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

Course Instructor: Dr. Chris Bryan

Submitted By:

Sangeetha Ramaswami(1224264691)

Sai Likhith Yadav Seera(1224791490)

Suraj Jyothi Unni(1219664641)

Riya Salian(1225625882)

Keval Shah(1225553979)

Preksha Gandhi(1219585913)

Domain Abstraction

The CrystalBall Visualization has many targeted users. Firstly, stakeholders and other potential users have their specific needs in identifying future related events. Retailers could also benefit by having the knowledge of the future gatherings at specific locations, such as malls, where timely and focused online advertising would be effective. Moreover, it could also help prevent unpleasant accidents, the Police can be well prepared for unforeseen events, lynch mobs or unruly rallies in their jurisdiction and take preventive and precautionary measures.

Data/Task Abstraction

A) Data Abstraction:

First visualization: The attribute Event Dates are Ordered with a cardinality of 30 (assuming, we are considering future event prediction for a month).
Emotion is another data abstraction attribute that is Categorical in nature with a cardinality of 6 representing emotions like anger, disgust, fear, joy, sadness and surprise.
Thirdly, Event Relation is a categorical attribute that has solid lines denoting events sharing the same location vs dotted lines representing events sharing the same keywords, hence resulting in a cardinality of 2.

Second visualization: Date Range is a categorical attribute (ordinal) with a cardinality of 5, helping us determine the date range in which the event occurred.
Geographic location is the second categorical attribute in the visualization with a cardinality equal to the number of events predicted.

Third visualization: The attributes in word cloud view are event keywords and the frequency of word in all events. The keywords are categorical data with cardinality as no. of keywords and the frequency of a word or its rank is shown by its size in the word cloud is quantitative data and cardinality is the maximum frequency.

Fourth visualization: There are five graphs in the fourth visualization. The flower glyph in the fourth graph shows 5 measures of future events, LR,NPMI,TS, UD, and Hashtag ratio. It also shows the no. of tweets.The value of each measure has a magnitude, so it is quantitative data with cardinality as the highest value of a measure. The 5 different measures in themselves are categorical data (cardinality:5) and no. of tweets is quantitative data with cardinality as the maximum no. of tweets.

In the group of bar charts in the event list view, the first bar chart shows the distribution of time when tweets are posted.In this the attributes are the hours and the no. of tweets. Hours are ordered, quantitative and cyclic with cardinality equal to 24, and the no. of tweets are quantitative and sequential.
The second bar chart displays the number of tweets over 30 days. The attributes for this chart are no. of tweets and the date. The no. of tweets is quantitative with cardinality as maximum tweets

in the 30 day gap and sequential data and the date is ordered, quantitative, and sequential data with cardinality 30.

The third bar chart shows predicted emotions at the event. The attributes here are the emotion values and emotion types. Emotion values are ordered, quantitative and sequential in nature with cardinality as maximum emotion value and emotion type is categorical in nature with cardinality 6 for the different emotion types.

The keyword summary shows all the keywords related to a future event. The data used for this is keywords that are categorical.

B) Task Abstraction:

First Visualization: The target can be to find out what events may occur on a specific future date(maybe a historically significant date) for which the action would be to lookup for the date and find events that are predicted to occur on that day.

Analyzing emotions associated with the events to determine the overall nature of the event can be considered another action-target pair.

Second Visualization: Exploring events based on geographic location to find out details of the events anticipated in a time range - for example, events in a span of 7-10 days or more.

Third Visualization: Through this graph we can discover the distribution of frequencies of different words and we can find which are the most prominent keywords by locating the outliers.

Fourth Visualization: The task abstractions for flower glyph are lookup features and discover distribution of different measures for a future event. It also helps to identify correlation between the different measures.

For the first bar graph we can discover the distribution of tweets posted at different time of the day that mention the future event.

With the help of the 2nd bar graph we can locate outliers and find when there is an increase in the no. of tweets mentioning the location time pair of the future event. We can also discover trends to understand the way the mention of event in various tweets increased over the days.

From the third bar graph we can identify features and understand the predicted emotion of the future event. The keyword summary can be helpful to browse features.

Visual encoding/Interacting Idiom

For this project, four visuals are under consideration. The first visualization attempts to differentiate between different events by using circles to symbolize marks and dates to indicate dates from tweets API data. A dotted or solid line is used to indicate the relationship between the tweets and the events to be known about in the future.

A confidence interval for the event that will have occurred with respect to the date and time is provided by the second visualization.Two bars are created for this, and the y-axis shows the probability that the event will occur, while the x-axis shows the data for one chart and the time for the other.Here, popular hashtags will be displayed as a word cloud based on the date chosen.

In the third, a spatial chart is utilized to encode the relationship between events and gives visualization based on where events are expected to take place. When you zoom in, a map of the events is displayed, at first providing a global overview of the activities.
The fourth visualization is word cloud. This helps to get an overview of future events without reading through the tweets. The size of each keyword in the cloud helps to analyze the number of its occurrences in all future events. Also when a date or an event is chosen in other visualization the corresponding keywords related to those events are shown in the word cloud view. When a user zooms in the map view the keywords related to the future events in that location are highlighted in this visualization by using different color schemes.

Algorithm

The Twitter Streaming API is used to collect and analyze data online. This API pulls thousands of streaming tweets from around the world daily.These tweets are then stored in MongoDB and are utilized when required. A stream of tweets is then sent for Entity extraction, which identifies the location, date, and geocode of the tweet from the provided tweet.For the extraction and mapping of the streaming tweet's date, location, and geocode, algorithms such as Stanford SUTIME, TweetNLP, and OpenStreetMap Nominatim are employed.

The location and time reference of the stream of tweets are identified in this section of entity extraction. These discovered times and dates are then mapped to the calendar date.If a tweet contains references to future times, dates, or locations, it is taken into consideration for further processing and is saved in the database. If not, the tweets are not taken into account for additional analysis.Further, these extracted tweets are utilized for future event identification. These tweets about future events constitute a small percentage from the processed tweets. Therefore, we extract small future event signals from a large base of noisy data.

Normalized Pointwise Mutual Information (NPMI) and **Pointwise Mutual Information (PMI)** are the attributes used for the extraction of these signals which provides information about the future events.Here, NPMI and PMI are calculated for finding out the correlation between the attributes like location and time attributes of future event from the given tweet data.

Higher the NPMI score, higher would be the correlation between the location-time pair. This leads to a higher chance of detecting weak signals of future events from the noisy data. User credibility and diversity indicate the informativeness of event tweets. **Degree of Centrality** and **Tweet Similarity** are the other two measures that indicate event tweet cohesion.

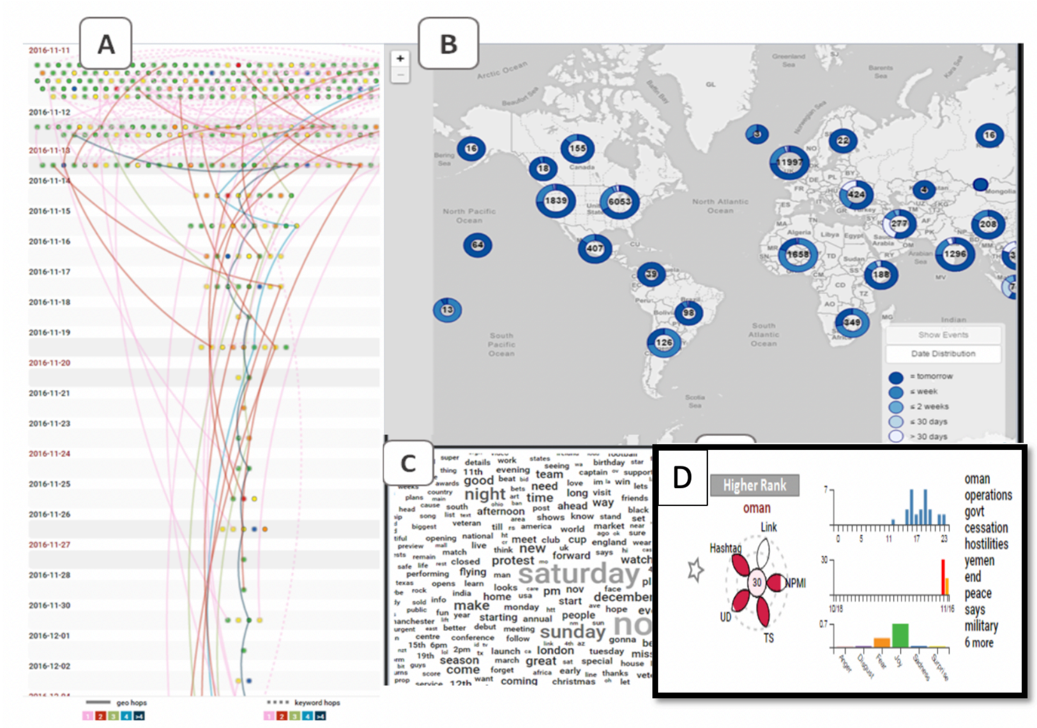
The **Link Ratio** is calculated as the number of tweets containing links over all tweets that are related to a possible future event. The **Hashtag Ratio** is calculated as the number of tweets containing hashtags over all tweets that are related to a possible future event.Informative tweets are more likely to be posted by credible users. The **Twitter Follower-Friend (TFF)** ratio has been considered to measure the user's credibility.TFF is the ratio of followers to friends.

User diversity is related to the informativeness of sources of tweets. If one user tweets regarding one potential future event, then it would likely be a bot which is programmed to send periodic tweets.To measure degree centrality between these tweets, we consider the retweets and mentions regarding one possible future event and then calculate DC by using Freeman's General

Formula for Centralization.Twitter Similarity has been created in order to overcome this situation. The similarity of a set of tweets linking to one future event is calculated using Levenshtein Distance.

RankSVM is used as a feature ranking method, which describes the relationship between explanatory variables and the response variable.

Mockup Image



A) Event Calendar View B) Map View C) Word Cloud View D) RankedSVM model view

Extensions

1) The visualization currently categorizes events into five types. Instead, a more detailed approach to event categories will be used. For example, there is a social event category here that can be subdivided into concerts and arts events.

2) A complicated graph is used to illustrate the relationship between the current event and the future event in time event-based visualization. Instead, a graph resembling a location event graph will be adequate to demonstrate the relationship.