Understanding Twitter Data with TweetXplorer

Fred Morstatter, Shamanth Kumar, Huan Liu, Ross Maciejewski
School of Computing, Informatics, and Decision Systems Engineering, Arizona State University
{fred.morstatter,shamanth.kumar,huan.liu,ross.maciejewski}@asu.edu

ABSTRACT

In the era of big data it is increasingly difficult for an analyst to extract meaningful knowledge from a sea of information. We present TweetXplorer, a system for analysts with little information about an event to gain knowledge through the use of effective visualization techniques. Using tweets collected during Hurricane Sandy as an example, we will lead the reader through a workflow that exhibits the functionality of the system.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining
; H.3.3 [**Information Search and Retrieval**]: Information Filtering

Keywords

Twitter Visualization, Retweet Network, Geospatial Analysis, Big Data

1. INTRODUCTION

The term "big data" describes data of a magnitude so large that it requires a change in methodology in order to process. Big data is often described by three "V"s: increased rate of data flow (*velocity*), heterogeneous types of media and perspective (*variety*), and enormous capacity (*volume*) [2]. These characteristics may lead one to believe that more data is always better. However, big data presents new challenges to existing data investigation systems and techniques. As data accumulates, it becomes harder to separate the wheat from the chaff. Our system addresses this "data paradox".

Social media data is undoubtedly big data. In June of 2012, Twitter alone reported 340 million tweets per day from 170 million active users[1]. In this sea of data we can find tweets containing eyewitness discussion of events that garnered worldwide attention in past couple of years. Social media data provides rich and expansive information about the pulse of populations worldwide. However, in order to obtain relevant information, one must know *exactly* what to search for. More often than not, we do not know exactly what we are searching for, but we know relevant search results when we see them. Using Twitter as an example[2], we present an effective approach to understanding big data and illustrate how one can start with a vague idea and iteratively dive into large volumes of social media data to find stories of interest. Our system connects human intelligence with rich data so that human clues can inform search and guide a user's query to form better conclusions about his data.

2. EXPLORING TWITTER DATA

To tackle the challenges of big data, organizations have evolved disciplined processes that guide the process of planning, collecting, analyzing, and curating their data. One model is the "Data Lifecycle" [9, p. 78], which includes the following phases: (1) "Plan & Prepare", (2) "Collect & Process", (3) "Analyze & Summarize", (4) "Represent & Communicate", and (5) "Implement & Manage". We follow these steps in the treatment of our social media data, outlining the decision making process at each step of the lifecycle. Support for the first two phases is provided by a companion system, TweetTracker [6]. TweetTracker collects and processes tweets matching the following parameters: hashtags, keywords, geographic boundary boxes, and Twitter usernames, chosen to track events on Twitter. Collected tweets are used by TweetXplorer. To help researchers deal with large volumes of social media data, we introduce TweetXplorer, as shown in Figure 1, a system that takes the user through the last three phases of the data lifecycle.

2.1 TweetXplorer

An analyst studying social media data will ask several questions to better understand their data. To begin, the analyst tries to find the most interesting days in the dataset (*when*). Once the analyst discovers the days of interest, they will want to enhance their understanding of the events that occurred on these days. To do this they will want to find important users and their tweets (*who*, and *what*), and important locations in the dataset (*where*). In the following sections we show how each of these can be extracted using TweetXplorer. For more information on our system, including videos of TweetXplorer in action, please visit the supplemental page[3].

[1] `http://blog.twitter.com/2012/03/twitter-turns-six.html`

[2] We choose Twitter for its openness in sharing data.

[3] `http://tweettracker.fulton.asu.edu/TweetXplorer`

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.
Copyright 2013 ACM 978-1-4503-2174-7/13/08 ...$15.00.

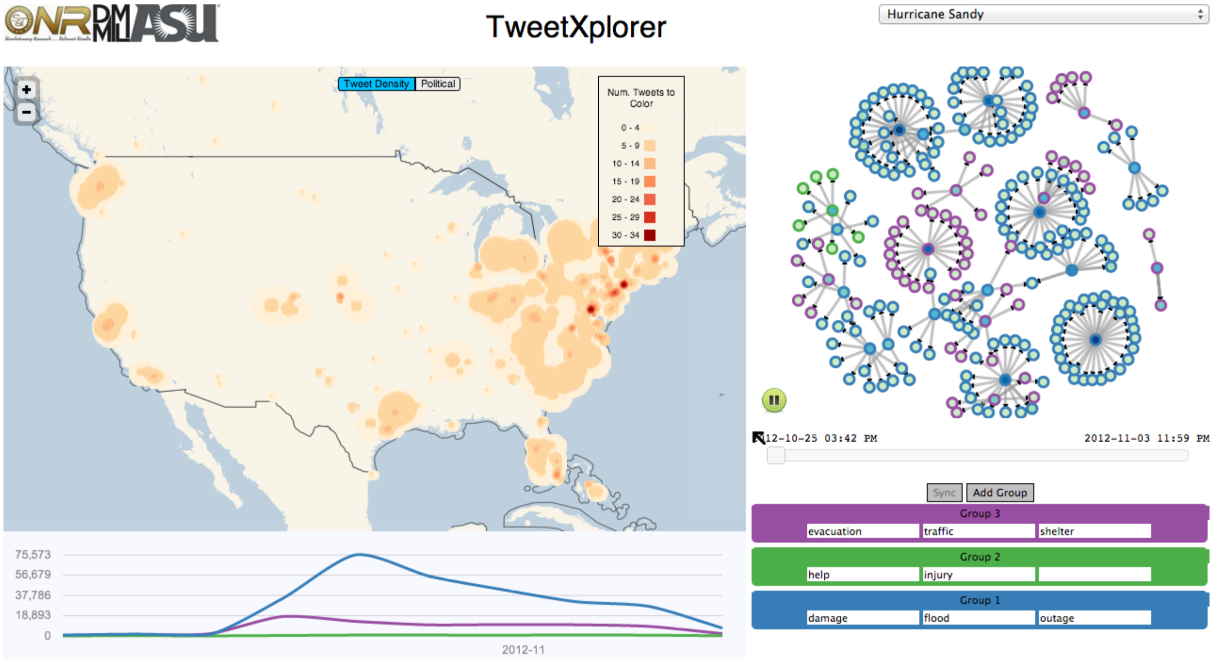
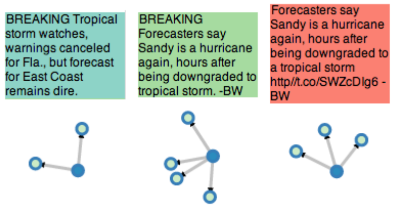


Figure 1: TweetXplorer - A system for visualizing big social media data.


BREAKING Tropical storm watches, warnings canceled for Fla., but forecast for East Coast remains dire.

BREAKING Forecasters say Sandy is a hurricane again, hours after being downgraded to tropical storm. -BW

Forecasters say Sandy is a hurricane again, hours after being downgraded to a tropical storm http//t.co/SWZcDIg6 - BW

Figure 2: Selective network filtering of a user

Creating Meaningful Queries: Traditional keyword search is iterative and it typically allows users to obtain a deep understanding of only one aspect of the data at a time. In social media analysis there are often multiple avenues an end user could investigate. We cater to this by allowing users to select multiple groups of keywords for analysis. By creating different groups, the user can instantly compare different pathways to see which path would best lead them to relevant information. It also helps in simultaneous investigation of the data along different dimensions. An example of groups of keywords a user could create is shown in the bottom right of Figure 1.

Discovering Interesting Time Periods: TweetXplorer offers a view into the data that shows the number of tweets matching a user's query occurring on each day in the dataset. This can be used by an analyst to identify interesting time intervals, where the importance of an interval is determined by the number of tweets posted on that day and "zoom in" to the specific intervals to investigate the data further.

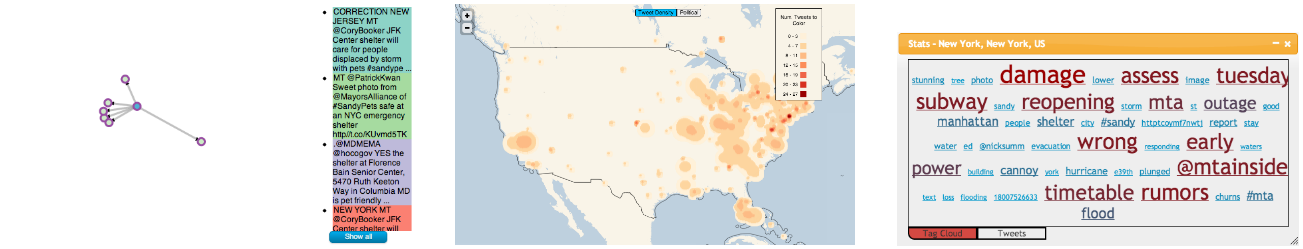
Representing Important Users and Tweets: We use Twitter's "retweet" functionality to help identify prominent users and tweets. By viewing the retweet network, the user can see which tweets were retweeted the most. The network visualization in TweetXplorer is implemented using the D3 visualization toolkit[4]. Each node in the network represents a user and the edges between nodes represents a retweet relationship. The network visualization uses the force directed layout [5] to place nodes and edges on the graph. The layout is designed to highlight nodes with high connectivity. The computation is prohibiting, $O(n^3)$, where n is the number of nodes in the network, which is relaxed in D3 with the Barnes-Hut approximation. Even with this approximation the algorithm can still involve significant computation, so we trim the network to 3 hops away from the original tweet. Through this visualization we are still able to find users at an individual level while making the dataset more manageable by showing only users and tweets which receive approval from the community. In Figure 3, we present an example retweet network.

Node information is encoded using two colors. The outer color of the node represents the keyword group it is associated with, which is determined by taking a simple majority of the group of his individual tweets. The darkness of the inner-color reflects the number of times a user was retweeted, with darker colors corresponding to higher retweet counts.

Additional node information can be summoned by clicking the node, which brings up an info pane, an example can be seen in the right part of Figure 3. The top of the panel contains the name of the user represented by the node, followed by the individual tweets authored by the user. The background color of each tweet corresponds to a slice in the pie chart, which summarizes the number of times each tweet was retweeted. The global retweet network can also be filtered to show a tweet-specific network, shown in the second row of Figure 2. This can be obtained by clicking a tweet in the first row of the figure. The network also automatically filters when a region is selected on the map.

Communicating Salient Locations: When a tweet is

[4] `http://d3js.org/`



- CORRECTION NEW JERSEY MT @CoryBooker JFK Center shelter will care for people displaced by storm with pets #sandype ...
- MT @PatrickKwan Sweet photo from @MayorsAlliance of #SandyPets safe at an NYC emergency shelter http//t.co/KUvmd5TK
- .@MDMEMA @hocogov YES the shelter at Florence Bain Senior Center, 5470 Ruth Keeton Way in Columbia MD is pet friendly ...
- NEW YORK MT @CoryBooker JFK Center shelter will

Stats - New York, New York, US

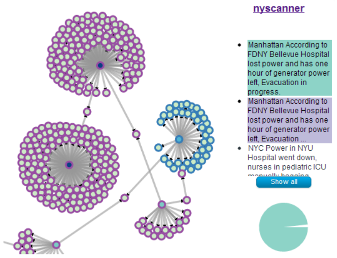
Show all
Tweet Density
Political
Num. Tweets to Color
0 - 3
4 - 7
8 - 11
12 - 15
16 - 19
20 - 23
24 - 27
stunning tree photo damage lower assess image tuesday
subway sandy reopening storm mta st outage good
manhattan people shelter city #sandy httptcoymf7nwtj report stay
water ed @nicksumm evacuation wrong responding early waters
power building cannoy york hurricane e39th plunged @mtainside
text loss flooding 18007526633 timetable rumors churns #mta
flood
Tag Cloud
Tweets

(a) Retweet network of @humanesociety discussing pet-friendly evacuation shelters.

(b) Heat map of tweets 1 hour before landfall.

(c) Tag cloud of most commonly-used words in NYC 1 hour before landfall.

Figure 4: Multi-faceted view of TweetXplorer displaying Pre-Landfall information for Hurricane Sandy.



nyscanner

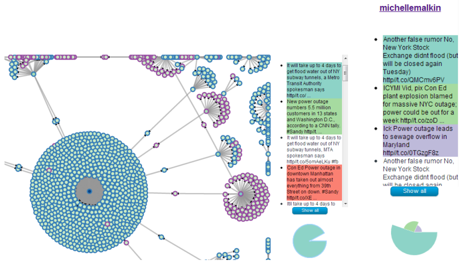
- Manhattan According to FDNY Bellevue Hospital lost power and has one hour of generator power left, Evacuation in progress.
- Manhattan According to FDNY Bellevue Hospital lost power and has one hour of generator power left, Evacuation ...
- NYC Power in NYU Hospital went down, nurses in pediatric ICU

Figure 3: A retweet network showing users from different keyword groups. The right shows an information panel.

created, the author has the option to "geotag" their tweet. In Figure 4(b) we see a view of the map showing a heatmap of the tweets matching the query "evacuation". The values in the heatmap are generated using a 2D kernel density estimate. This visualization enables the user to immediately see regions of interest on the map. Here, we see that the darkest region in view lies on the Eastern Seaboard of the U.S. We zoom in to this region, causing the heat map to disappear and show each individual tweet on the map. To investigate tweets from a specific region, we brush the region. Brushing the region around New York City brings up a popup showing the tweets published from this region and a tag cloud of the most prominent words in the dataset, shown in Figure 4(c). This advanced view can help the user to understand the topics of discussion in specific areas during a crisis and also assist them to improve their queries.

The map component connects network information by displaying geotagged retweets on the map. This helps to find *where* individual tweets are receiving attention. The map overlays the network information to help direct the user to locations that retweet the initial message. This can help the user to identify other regions of interest, but which may not necessarily be high-traffic areas. For more information, see the video referenced at the beginning of this paper, or the supplemental images on our web page.

Discovering User Patterns: In addition to more macro-level views of the data, TweetXplorer also offers a glimpse into the patterns of individual tweeters. TweetXplorer allows the user to zoom into the tweeting behavior of users to find characteristics of the users discussing the event. To get a better understanding of the user's behavior, we break their behavior into the following categories: *when*, *what*, *where*, and *how*, shown collectively in Figure 5. Figure 5(a) shows the time of the day and days of the week *when* the user prefers to tweet. We can also see the topics that interest



- It will take up to 4 days to get flood water out of NY subway tunnels, a Metro Transit Authority spokesman says http//t.co/ ...
- New power outage numbers 5.5 million customers in 13 states and Washington D.C., according to a CNN tally. #Sandy http//t. ...
- It will take up to 4 days to get flood water out of NY subway tunnels, MTA spokesman says http//t.co/5omAeLKu #fb
- Con Ed Power outage in downtown Manhattan has taken out almost everything from 39th Street on down. #Sandy http//t.co/XE ...
- Itll take up to 4 days to

michellemalkin

- Another false rumor No, New York Stock Exchange didnt flood (but will be closed again Tuesday) http//t.co/QMCmv6PV
- ICYMI Vid, pix Con Ed plant explosion blamed for massive NYC outage; power could be out for a week http//t.co/zoD ...
- Ick Power outage leads to sewage overflow in Maryland http//t.co/0TGzgF8z
- Another false rumor No, New York Stock Exchange didnt flood (but

(a)

(b)

Figure 6: (a) Reports from @cnnbrk on flooding and power outage on east coast. (b) A twitterer debunking the rumor about NYSE.

the user by viewing a tag cloud of his most common hashtags, shown in Figure 5(b). This helps us understand *what* topics motivate him to tweet. To understand *where* the user tweets from, we provide a map displaying the user's geotagged tweets, shown in Figure 5(c). To show *how* the user tweets, we provide a pie chart depicting his client distribution, shown in Figure 5(d). This can help to understand the user's tweeting conditions.

3. CASE STUDY

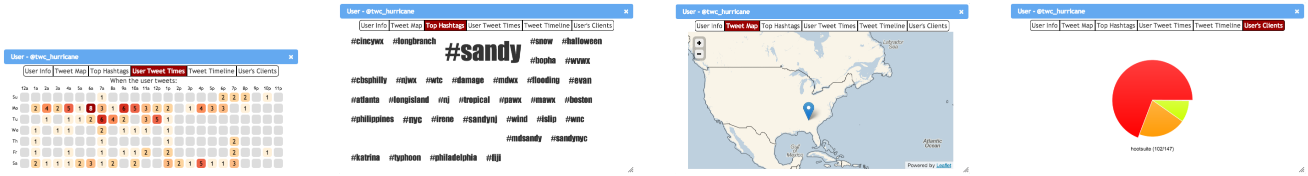
The final step of the data lifecycle is to put the interpretations of the data to use. Here we will present an example of TweetXplorer to interpret Twitter data using a dataset containing tweets from Hurricane Sandy, a massive storm that devastated the East Coast of the United States in late October, 2012. We collected some of the discussion of this event on TweetTracker for further investigation in TweetXplorer.

3.1 Hurricane Sandy Data

The parameters used to collect the data are shown on our supplemental web page. We collected 5,639,643 tweets during the period: October 25th, 2012 to November 3rd, 2012. For the purpose of case study, we partition the dataset into three distinct epochs: pre-landfall (2012-10-29 00:00 - 2012-10-29 17:59), landfall (2012-10-29 18:00 - 2012-10-30 23:59), and recovery (2012-10-31 00:00 - 2012-11-01 12:00). We select keywords to help understand the important topics in each time period, as shown in the bottom right of Figure 1. We will enumerate each epoch and discuss some findings.

3.2 Pre-Landfall

In the hours leading up to Hurricane Sandy's landfall, we see discussion along different paths. In Figure 4(a), we see



When the user tweets:

	12a	1a	2a	3a	4a	5a	6a	7a	8a	9a	10a	11a	12p	1p	2p	3p	4p	5p	6p	7p	8p	9p	10p	11p
Su								1											2	2	2		1	
Mo		2	4	2	5	1	8	3	1	6	5	3	2	2		1	4	3	3		1			
Tu			1		1	1	2	6	4	2		3	5	1										
We		1		1	1			2		1	1	1		1										
Th		1						1						1						2				
Fr		1			1			1		1	1	2		2						2			1	
Sa		2	1	1	1	2	3	1	2		2			3	2	1	5	1	1	3				

(a) Times the user tweets by hour of day and day of week.

(b) Tag cloud showing users preferred hashtags.

(c) User's geolocated tweets.

(d) Distribution of how the user publishes tweets.

Figure 5: Views of the user component in TweetXplorer.

that one of the most highly-retweeted tweets is a tweet mentioning the availability of pet shelters in evacuation areas. Looking at the geotagged tweets produced during this epoch we observe that tweets contain general hashtags and topics of conversation. As the storm nears we see people move towards terms describing specific issues, such as "rumors", "damage", and "subway", shown in Figure 4(c).

3.3 Landfall

Hurricane Sandy made landfall on Oct 29, 2012 at 20:00 EST [5]. First reports of flooding start to arrive around this time, accompanying links to images of flooding. As the storm progresses, we observe several reports of power outage from NYC and nearby areas. Due to the power outage, we discover reports of hospitals which were forced to evacuate their patients. In Figure 3, we can see two clusters of users connected by common retweeters discussing this. In Figure 6(a), one can observe that @CNN's tweets claim that more than 5.5 million people were without power in the region. More interestingly, most of @CNN's retweets came from their tweet reporting the flooded NYC subway tunnels, which was retweeted 678 times. Reports from the NYC area paint a similar picture, with top hashtags discussing power outages and flooded subway tunnels. At the same time, false rumors were also spreading on Twitter. As seen in Figure 6(b), there were several reports of flooding of the NYSE building. Agencies such as @weatherchannel tweeted this information and later retracted it, as it was discovered to be false.

3.4 Recovery

While analyzing the Twitter activity after the storm we notice that the most prominent tweet during this time discusses repair resources. Also, the discussion in NYC focuses on the words "damage", "power", and "flood", indicating an attention shift towards post-storm topics. Supporting images are available on our supplemental web page.

4. RELATED WORK

Visualization of events has helped bring increased situational awareness to first responders in crisis scenarios. [8] uses keyword-based search to bring situational awareness to rescue crews in flood and fire scenarios. [4] assists high-ranking commanders in battlefield scenarios through the use of map layers and textures.

Geographical visualization is central to our system. In [3] the authors' mapping system utilizes a map's legend to show statistical properties of the data. In [7] the authors present a tool designed to gather situational awareness from tweets.

Network visualizations can play an important role in the analysis of large datasets. Many systems have been proposed to visualize large networks, such as Gephi [1].

5. FUTURE WORK

In this work we have demonstrated TweetXplorer, a visual analytics system that can help a user gradually obtain deeper insight into Twitter data. Going forward we will create new modules for TweetXplorer, which focus on other forms of information including sentiment analysis and detecting dynamic communities. We will extend the keyword grouping interface to help the user discover keywords and hashtags central to an event, and the network component to show different types of networks in Twitter, such as hashtag co-occurrence, and Twitter friendship network.

6. ACKNOWLEDGEMENTS

This work is sponsored by Office of Naval Research grants N000141010091 and N000141110527.

7. REFERENCES

[1] M. Bastian, S. Heymann, and M. Jacomy. Gephi: An Open Source Software for Exploring and Manipulating Networks. In *ICWSM*, volume 2, 2009.

[2] E. Dumbill. What is Big Data? `http://radar.oreilly.com/2012/01/what-is-big-data.html`, January 2012.

[3] J. Dykes, J. Wood, and A. Slingsby. Rethinking Map Legends with Visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):890–899, 2010.

[4] E. Feibush, N. Gagvani, and D. Williams. Visualization for Situational Awareness. *CG&A, IEEE*, 20(5):38–45, 2000.

[5] T. Fruchterman and E. Reingold. Graph Drawing by Force-Directed Placement. *Software: Practice and experience*, 21(11):1129–1164, 1991.

[6] S. Kumar, G. Barbier, M. A. Abbasi, and H. Liu. TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief. In *ICWSM*, 2011.

[7] A. MacEachren, A. Jaiswal, A. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford. SensePlace2: GeoTwitter analytics support for situational awareness. In *IEEE VAST*, pages 181 –190, oct. 2011.

[8] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. Microblogging During Two Natural Hazards Events: What Twitter may Contribute to Situational Awareness. CHI '10, pages 1079–1088, 2010.

[9] H. Whitney. *Data Insights: New Ways to Visualize and Make Sense of Data*. Morgan Kaufmann, 2012.

[5] `http://www.nhc.noaa.gov/archive/2012/al18/al182012`