A Visual Analytics Approach to Summarizing Tweets

Ramik Sadana
Georgia Institute of Technology
ramik@gatech.edu

Yekyung Kim, Bongwon Suh
Seoul National University
{peiru, bongwon}@snu.ac.kr

Eunyee Koh
Adobe Research
eunyee@adobe.com

ABSTRACT
Corporations are increasingly monitoring Twitter to gain insights on entities of interest, such as products, brands or celebrities. However, the ever-increasing conversation on Twitter has made it difficult to identify the relevant themes from large corpuses of tweets. Some existing systems offer techniques to summarize tweets and highlight the overall Twitter activity. However, these systems seldom offer a comprehensive solution for an exploratory sensemaking task. In this work, we build on key principles of visual analytics and describe an end-to-end, visual exploration system for tweets that both presents overall summaries and supports analysis of any variations that exists in the activity.

Categories and Subject Descriptors
H.5.0 [**Information Interfaces and Presentation**]: General

General Terms
Design, Algorithms, Human Factors.

Keywords
Twitter, social media, visualization, content summarization

1. INTRODUCTION
Recently, there has been a tremendous increase in the use of Twitter, with the average number of tweets exceeding 20 million per hour [10]. These tweets are a rich source of data for identifying developing stories, examining public sentiment and even predicting the stock market [1]. This data is also of particular interest to corporations and industries that are increasingly monitoring tweets to gain insights on objects of interest, such as a product, a brand, or a person. Organizations use the Twitter API to request tweets relevant to one particular entity by specifying up to fifty keywords that relate to that entity. The API in turn returns a corpus of the most recent tweets that contain any of the specified keywords. By processing this corpus, organizations monitor the overall activity and mood related to an object.

These API requests, however, often result in massive amounts of tweets. It has been a continuing issue to process these large datasets to extract relevant information. Past research has presented many methods, several of which use summarization techniques [3][9]. Various commercial systems now exist that, given an input of tweets, present summarized versions that highlight the overall Twitter activity [11][12]. However, these systems seldom offer a comprehensive solution for an exploratory sensemaking task. Sensemaking has been defined as 'A motivated, continuous effort to understand connections (which can be among people, places, and events) in order to anticipate their trajectories and act effectively' [6]. In a sensemaking task, the aim of the user is to digest and comprehend the vast data in such a way that specific insights and actionable items are discovered. For large collection of tweets, this could entail first identifying periods of remarkable activity in the feed, either more or less than expected amount, and then identifying the causes for such deviations. Although existing systems offer comprehensive metrics on total amount of tweets and highlight the variability that might be present between different periods, they often do not support addressing why such variability exists.

In this work, we describe an end-to-end exploration system for tweets, which both presents overall summaries as well as supports analysis of any variations in trends. We focus on the three key principles of visual analytics, namely automatic analysis, effective visualization and fluid interaction [5]. The following sections discuss multiple strategies that we considered for summarizing tweets and highlight the visualization principles we used for designing our system.

2. RELATED WORK
A variety of past research has focused on Twitter summarization. O'Connor et al. [8] present TweetMotif, a system that takes an input phrase and extracts all significant activity on twitter that relates to the phrase. Sharifi et al. [9] present an approach to generate a one-line summary for themes within tweets using phrase reinforcement ranking. Harabagiu and Hickl [3] extend this model to generate larger size summaries, i.e., 250-words per summary.

The main objective of these works is to summarize tweets in a few sentences or just a few words. Most of these systems analyze the entire stream of Twitter data. However, some research also exists that focuses on selective subsets of tweets. Chakrabarti and Punera [2] generate summaries for sports topics by learning an underlying hidden state representation of the event via hidden Markov models. Louis and Newman [7] present a method for summarizing a collection of business-related tweets by aggregating tweets into subtopic clusters, which are then ranked and summarized.

For our system, we utilize some of the techniques described in these works. However, in our work we mainly utilize visual techniques for summarizing Twitter activity instead of purely textual methods used by the above-mentioned systems.

3. ANALYZING TWEETS
In this work, we are primarily concerned with tweets that pertain to a particular entity such as an organization, a company or a product. We collect these tweets by querying the Twitter API with up to fifty keywords related to that entity. The API in turn returns a collection of the most recent tweets that contain any of the specified keywords. Our aim is to develop an exploratory system for these tweets that both presents overall summaries as well as supports analysis of variations in trends.

One highly relevant category of users for such a system is the social media manager (SMM). SMMs of an organization are often tasked with tracking conversations on twitter that relate to the organization, and highlighting any activity that may need to be addressed. To design a Twitter monitoring system better aligned

SIGIR '14, July 6–11, 2014, Queensland, Australia.

with their use case, we reached out to a few SMMs to understand their expectations from such a system.

Our discussions resulted in three main requirements:

1. The system must show the overall trend of conversation for the current and previous periods.
2. For the periods that show large variability in conversation, the system must support identifying the tweets that caused such variations.
3. For any tweet or topic that is highlighted, the system must reveal its past activity, particularly periods of large activity.

The above requirements gave us a concrete set of features to focus on. Since we wanted our system to largely be visual, we referred to the vast array of information visualization research. We found that Shneiderman's "*overview first, zoom and filter, details on demand*" mantra aligns well with the flow of data exploration presented in these requirements. As the smallest unit of analysis in our data is a tweet, displaying a single tweet, or a small group of tweets, fulfills the *details on demand*. For the *overview*, we considered various approaches. One approach presented in past research uses document summarization [3][9]. In the case of Twitter, this entails fetching tweets for a fixed period of time, such as a day, or for a particular event, such as Sochi Winter Olympics. Once fetched, these tweets are analyzed and a textual summary of the overall conversation is generated. This summary mainly contains the metrics on the tweets and, in some cases, a story generated by threading together themes that appear in representative topics or prominent tweets. From an end user perspective, presenting the user with this condensed, summarized form of tweets instead of raw tweets simplifies sensemaking by shortening the time it takes the user to process all the data.

One flaw with the above approach, however, is that there tends to be a large disconnect between the textual summaries and the actual tweets. Themes highlighted often appear out of context and the systems offer no easy linkage between the summary and the underlying tweets. Additionally, even though summaries are substantially more condensed than raw tweets, they still contain fairly long text and it takes SMMs considerable time to read them. Our goal is to shorten the time it takes SMMs to analyze the large corpus of tweets. One effective solution is to use *trending topics*. Trending topics are phrases that have seen a significant increase in conversation in the recent periods. These phrases are extracted from tweets where they predominantly appear as hashtags (words mentioned in tweets prefixed with a '#' symbol). Hashtags are often used for topic association. Thus, by looking at the list of the trending hashtags, a user could get a general sense of the landscape of the conversation happening on Twitter. In other words, trending topics act as a summary of the overall activity.

Social media sites such as *tumblr* and *YouTube* present trending topics their landing page. These topics offer a natural entry point for a user who isn't necessarily looking for anything specific. Although Twitter has not published the exact algorithm that it uses to detect these trending topics, the basic premise of the algorithm is to identify those topics whose frequency of occurrence in tweets from the current period is more than some expected value. This expected value is derived from the occurrences of the topic in past periods. There are other constituents of the algorithm such as the velocity of conversation, the significance of the profile that began the topic, and the geographic region represented by the conversation. The main constituent of the algorithm, though, is the expectation function.

4. TRENDING TOPICS

We use the approach of trending topics to summarize tweet activity. Since our focus is on tweets that are related to a particular entity, for this work we chose to concentrate on *Adobe Systems*. Adobe is a computer software company that primarily offers graphic design and publishing products such as Adobe Photoshop, Adobe Illustrator, and Adobe InDesign. The keywords we used to fetch tweets contained the names of all of Adobe's products as well as any promotions that were active at that time.

For a large organization like Adobe, the amount of twitter activity each day exceeds 50000 tweets. The conversation ranges from product announcements and sharing of work to bug reports, complaints against the organization, and other casual banter. Our aim is to present such a huge corpus of tweets to an SMM in a summarized, comprehensible manner.

Once we collect this data, the next step is to identify trending topics. In a preliminary analysis of the dataset, we extracted the hashtags from all the tweets. We observed that the majority of hashtags were the same as the keywords we used to fetch the tweets. Unlike on the Twitter website, where topics such as #Photoshop or #CreativeCloud might be effective as trends, when summarizing data for only Adobe-related tweets, these topics are far less useful. As a result, to expand the richness of trending topics, we increased the scope of the topics from just hashtags to any word within the tweet. One downside of doing this is that many single words inside tweets are non-descriptive and unintelligible. For instance, some words we observed in our data were *believe*, *rating*, and *judgment*. These words are ineffective as topics since they would not make much sense to a user unless presented in the context they appeared in. Hashtags based trending topics, such as on the Twitter website, sidestep this issue because quite often hashtags are a collection of words, e.g. #FreeJustinBieber and #HappyNewYear, and as a collection often make sense and are fathomable.

Table 1. Flow of bigram extraction from tweets

Step 1	Input tweet	Find 7 free actions for movie color correction #Photoshop via @youthedesigner #Psactions http://t.co/3ZQateeeVax
Step 2	Remove hashtags	Find 7 free actions for movie color correction | via @youthedesigner | http://t.co/3ZQateeeVax
Step 3	Remove URLs	Find 7 free actions for movie color correction | via @youthedesigner | |
Step 4	Remove refs.	Find 7 free actions for movie color correction | via | | |
Step 5	Remove stopwords	Find | free actions | movie color correction | via | | |
Step 6	Extract bigrams	Free actions, movie color, color correction

To make the trending topics we detect more informative, we use pairs of words (bigrams) instead of single words (unigrams). Pairs of words often create more meaning together than the words do individually. In the case of tweets, since the character limit is 140, there are only around 7 words on average. Given that tweets usually also contains URL links, user-mentions, and other similar entities, a bigram in a tweet aligns closely with the overall sense and offers a reasonable insight into what the tweet means.

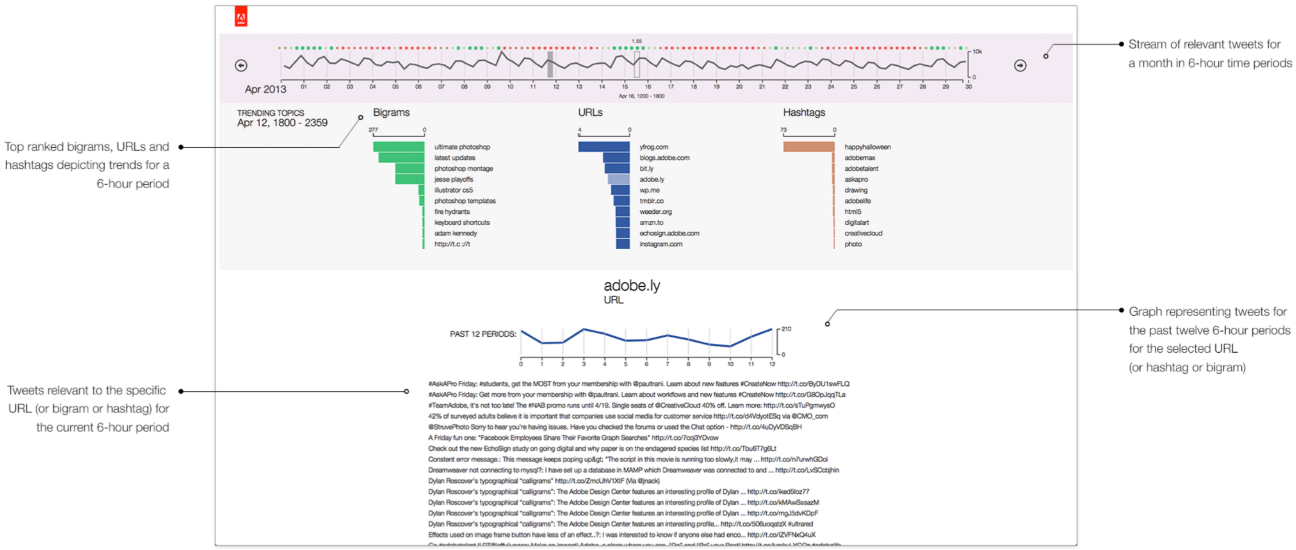

Figure 1. Snapshot of the system displaying meta-timeline, topic-lists and tweet-view.

Before we extract bigrams from tweets, we discard all the tweets that are non-English. Additionally, we also exclude tweets that have less than 4 words. On the remaining corpus of tweets, we use a three-stage technique to extract bigrams. In the first step, we strip a tweet of all entities that are hashtags, URLs, stop words or user-mentions (we exclude hashtags since we process trending hashtags separately, as discussed later). Next, we split the remaining tweet at the position of these entities to create smaller sentences. Finally, we extract bigrams from these smaller sentences and populate them into a database along with their timestamp. Table 1 describes this flow from an input string to the output bigrams.

We use this approach to process tweets each time we receive them from Twitter. The subsequent step is to identify the trending topics. Trending topics can be identified at different time scales, such as one hour, a few hours, one day, or longer. In order to balance noise with temporal relevance, we chose to use a six-hour period. A six-hour period is preferable over a one-hour period since one-hour periods are likely to contain spikes of activity for a particular topic. Using a six-hour window normalizes these spikes over a period long enough so that only steadily trending topics are highlighted. A one-day period would produce results even more stable than a six-hour period. However, we lose granularity and potentially miss out on important, short-term topics with a time scale that large.

We divide one day's tweets into four parts based on their timestamps. The groups thus generated are: 0000 - 0559, 0600 - 1159, 1200 – 1759 and 1800 – 2359. At the time during a day when each of these periods end, we process the trending topics for the past six hours. To do so, we compare the frequency of activity for the new bigrams in that period to the expectation from past periods. Any activity that is substantially larger than its expected value is identified as potentially trending. However, the number of tweets in each of the four periods of a day is different depending on people's work and sleep cycles. As a result, instead of comparing the frequencies directly, we compute a normalized value, which is the frequency of a bigram in a period divided by the total number of tweets in that period.

$$\text{Norm. MA}_n = 0.7 * \text{Norm. MA}_{n-1} + 0.3 * \left[\frac{\text{Frequency of bigram}}{\text{Total number of tweets}}\right]_n$$

For all bigrams that are detected in previous periods, we maintain a moving average of their normalized values for all subsequent periods. For a bigram detected in the current period, its normalized value is compared with its moving average from past periods. This ratio, called trend-strength, depicts the scale of variation from expected conversation for that bigram. Bigrams with the highest trend-strength in a period are the trending-bigrams. In the calculation of the moving average, we use exponential smoothing to diminish effect of the older periods on the current average. The purpose behind this is to ensure that we identify a trending bigram in a subsequent period even if it was trending in a preceding period.

$$\text{trend-strength}_n = \frac{\left[\frac{\text{Frequency of bigram}}{\text{Total number of tweets}}\right]_n}{\text{Normalized moving average}_{n-1}}$$

Along with bigrams, we also identify trending hashtags and URLs. This gives the user multiple entry points for data analysis. For URLs, we use the domain name as the unit of analysis. Since Twitter encodes URLs in a tweet in a custom form of 't.co', we first expand this URL to regenerate the original URL. From this original URL, we extract the domain name. Similar to bigrams, we maintain a moving average for each hashtag and URL domain we detect and use these moving average to identify the trending topics for each new period.

5. VISUALIZATION

We implemented our visualization system (Figure 1) for the web using HTML and JavaScript, using the D3 framework to draw the graphs. We structure the visualization based on the process of data analysis that emerged from our discussions with the SMMs. The interface begins with presenting the overview of all periods, the *meta-timeline*. The users can select a period to zoom to and view the trending topics for that period, the *topic-list*. Finally, for a topic, the user can reveal the underlying tweets corresponding to that topic, the *tweet-view*.

5.1 Meta-timeline

The top of the interface contains a timeline presenting the overall trend of tweets for a particular month. The data is presented for each day, further broken down into the 4 periods described above.

Data for other months can be accessed using the previous and next buttons. The overall conversation follows a daily trend, with the conversation being highest between 0600-1159, followed by 1200-1759. Due to this daily variability, it is difficult to spot any outliers. The outliers of interest are both a steep rise and a steep fall in the conversation for any period. To make it easier to identify these periods of abnormal activity, we show colored dots at the top of each period. The color of the dots depicts the change in conversation for that period - dark red depicting large drop and dark green depicting large increase.

We calculate the drop and increase by comparing the amount of conversation in a period with the expected conversation in that period. Past work has highlighted the seasonal nature of the tweets [4]. Based on that work, to correctly identify the trends, we use a measure of expected conversation for each period for each day of the week, thus generating a set of 28 (7 days x 4 periods) expected values. Each period's activity is compared with its respective expected value to generate a ratio. This ratio determines the color of the dot. Additionally, hovering over a period reveals this value.

$$\text{ratio} = \frac{\text{\# of tweets in period n}}{\text{Expected \# of tweets in period of week}}$$

5.2 Topic-list

Once the user clicks on the timeline to select a period for further analysis, the view presents three separate lists of trending topics, one each for bigrams, URLs and hashtags. Each list contains ten topics and the topics are ordered based on the trend-strength. Further, for each topic, a bar represents the exact value of trend-strength, highlighting the relative importance of each topic.

5.3 Tweet-view

For each topic in the three lists, the user can select the topic to reveal the tweets associated with it. The tweets appear as a paginated list. Additionally, for the current topic, a linechart also depicts the overall flow of the tweets from the past 12 six-hour periods. This helps to identify periods of large or small activity.

The three views of the system map well to the process of data exploration that emerged from our discussions with the SMMs. Moreover, the system provides the ability to see multiple levels of data in the same view. This allows the user to link activities from different periods, which can lead to serendipitous discoveries.

6. CONCLUSION

In this article, we describe a web-based interface for visualizing summaries of twitter conversations. Focusing on requirements highlighted by social media managers (SMMs), we build on the visual analytics principles of automatic analysis, effective visualization and fluid interaction. Our system presents the user with a high level summary of tweets for multiple months, highlighting periods of key activity. For each period, the system displays trending topics in the form of bigrams, hashtags and URLs. Finally, the system supports fine-grained analysis by displaying individual tweets for each topic. With the help of the Adobe use-case, we illustrate the relevance of the system for monitoring Twitter patterns and trends for large organizations.

A natural follow up to this work is a user evaluation. User testing on the various features will provide helpful feedback on their usefulness. Since better feedback emerges when testing is conducted in a realistic setting for an extended period of time, we aim to deploy a working version of the system for use by SMMs over a period of a few weeks. However, we expect such an evaluation to generate various feature requests. This is primarily because the systems currently used by SMMs offer a lot of features for the overall analysis of tweets. And although the strengths of our system are of a broader nature, we believe additional features for overall analysis can be added.

We also would like to further optimize the bigram extraction process. For instance, some bigrams currently extracted are '*other suite'* or '*before photoshop'*. Optimizing the algorithm for better lexical and semantic processing could generate more useful results. Additionally, we can also use trigrams, or higher order n-grams, instead of bigrams. Similar to the transition from unigrams to bigrams, we expect trigrams and n-grams to be more useful than bigrams as representations of underlying topics.

Finally, we also would like to incorporate end-user customization and feedback into the system. Currently, the system presents a static list of extracted topics. A user might identify a topic to be inadequate, either because the topic is lexically weak or because it is not directly related to the entity whose data is being analyzed. In those situations, we would like to provide the user the ability to exclude those topics and specify the cause. This information can then be used to update the set of keywords and refine the results. This would also help to optimize the topic detection algorithm.

7. REFERENCES

[1] Bollen, J., Mao, H. and Zeng, X. Twitter mood predicts the stock market. *Jour. of Computational Science,* 2011, 1–8.

[2] Chakrabarti, D. and Punera, K. Event Summarization Using Tweets. *AAAI Conf. on Weblogs and Social Media* 2011.

[3] Harabagiu, S. and Hickl, A. Relevance Modeling for Microblog Summarization. *AAAI Conf. on Weblogs and Social Media* 2011.

[4] Java, A., Song, X., Finin, T. and Tseng, B. Why We Twitter: Understanding Microblogging Usage and Communities. *Proc. of 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*.

[5] Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J. and Melançon, G. 2008 Visual Analytics: Definition, Process, and Challenges. *Information Visualization*.

[6] Klein, G., Moon, B. and Hoffman, R.R. Making Sense of Sensemaking 1: Alternative Perspectives. *IEEE Intelligent Systems* 2006.

[7] Louis A. and Newman T. 2012. Summarization of business-related tweets: A concept-based approach. *Proc. of 24th Intl. Conf. on Computational Linguistics (COLING)*.

[8] O'Connor, B., Krieger, M. and Ahn, . TweetMotif: exploratory search and topic summarization for Twitter. *In Proc. of AAAI Conference on Weblogs and Social* 2010.

[9] Sharifi, B., Hutton, M. A. and Kalita, J.K. Experiments in Microblog Summarization. *IEEE Second International Conf. on Social Computing (SocialCom)* 2010.

[10] Smith, C. 2014. By The Numbers: 104 Amazing Twitter Stats. *http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/*

[11] Topsy. http://topsy.com/analytics

[12] Twitonomy. http://www.twitonomy.com