The Twitter Connection [Working Title]

Gabe Benjamin, Cole Harbeck, and Adrian Wan
Swarthmore College
{gbenjam1, charbec1, awan1}@cs.swarthmore.edu

1. PROJECT PROPOSAL

1.1 Problem

A description of the problem you plan to solve and the motivations for doing so (i.e., why this problem is interesting/important).

Twitter is a social media platform with 284 million users active per month, tweeting approximately 500 million tweets per day. [9] At a maximum of 140 characters per tweet and about 1 byte per character[1] this represents an informational flow of over 70GB of text per day. An average book is about 2MB in size [2], so Twitter users are collectively writing about 35,000 books a day.

With this wealth of textual information that is often supplemented with geo-location data and content-connecting hashtags, it is no surprise that a multitude of tools have arisen to harvest the information encoded in tweets. Twitter's API [10] has been used to build visualizations of tweet locations and several third party apps have been created: Tweet Ping is a website that tracks live tweets and displays them like lights on a global map [4] [3]; A World of Tweets is a similar website that builds heatmap of tweets [5]; Tweet Beam creates a wall of tweets based on a given hashtag or search query [8]. These are just a few of the many apps that utilize the live stream of tweets from Twitter.

Existing products that leverage the informational wealth of Twitter are mostly interested in geolocation, and are either purely historical or purely live. Few, if any, current products focus on the connection between tweets or go beyond simply using geolocation or maybe a keyword. We seek to move away from this problem space by creating a product that:

- Uses both historical and live-stream data
- Focuses on relationships between tweets, users, and content, rather than on just their location
- Displays these relationships in a visually pleasing way.

[1]Issues arise with different encodings, but we ignore these for the purposes of our back-of-the-envelope calculation.

1.2 Goals

The goals you have for the project. What constitutes success and how will you evaluate it?

Our goal for this project is to have a fully functional visualization depicting the popularity of historical and recent tweets (based on retweets) and the connectivity of tweets (based on @replies and hashtags). We envision TwitCon(WT) working as follows: users access a webpage which starts blank and immediately starts populating with live tweets. The tweets grow in size as they are retweeted, tweets that are replies are connected, and tweets are clustered by conversation rather than by location. Coloring could be used based on mood — a plug-in affect detection module could be an interesting demonstration of the modularity of our work[2]. The window changes in size to accomodate the largest (most retweeted) tweets. Users can move the window around by dragging and clicking, and can zoom in and out by scrolling. Tweets contain a user handle and/or (user-provided) real name, a snippet of content, and can be clicked on to access the full tweet. Movement towards the edge of the window displays options, such as a toggle for historical context (continuous scale at day variation up to perhaps the month scale — the maximum amount of caching depends on API and storage limitations), mood, and so on.

The primary goal of this project is to create a (relatively) novel visualization of relationships between Twitter users. As such, success can be measured by the quality of the insights that TwitCon(WT) provides. By exploring the visualizations produced — see §1.3 for an articulation of the trends that we hope to see — we would consider the quality of the insights gleaned from TwitCon(WT) a marker for its success.

Equally central to this project is how we can present interesting information in aesthetically pleasing, intuitive ways. New information is no benefit if it does not elucidate the data that it is gleaned from. To have value, visualizations must be pleasing to the eye and easy on the mind. Hatnote's Listen to Wikipedia project [6] serves as our primary inspiration for a simple visualization (and auralization) that beautifully presents data that may have otherwise fallen to the wayside (additions to and subtractions from articles, additions of new users and pages).

Thus, our goals can be summarized as follows: we wish to create a product that

[2]G. Benjamin has authored early work on a mood-detection module that we would be interested in applying.

Course paper for Swarthmore College course CS 91: Cloud Systems and Datacenter Networks

- Provides interesting insights into the culture of Twitter by mapping interactions of users
- Displays both historical and live-stream data
- Presents an aesthetically pleasing, intuitive way to understand Twitter from a different, global perspective.

1.3 Hypothesis

Your hypothesis: given the work you intend to do, what are the results you expect to see? How does this work help to solve the problem?

The main difference between TwitCon(WT) and currently existing visualization tools is that we choose to focus on the relationships between tweets rather than on user location. Current Twitter visualization tools allow users to see where and when people are tweeting. This visualization method is no longer novel, and may not actually provide any new information about the Twittersphere, as geographical distribution for tweets correlates closely with population distributions [7][3].

Our work will allow us to draw different conclusions from Twitter data. Although we cannot predict exactly what kind of insights can be provided by this new way of visualizing Twitter data, by visualizing popularity and connectivity of tweets, examples of what we expect to be able to see include:

- Types of tweets, including:
  - Those from celebrity figures, characterized by a large number of retweets and replies
  - Announcement and news tweets that are predominantly retweeted
  - Personal tweets interconnected by @replies that form threads of conversation.
- Types of tweets that are spawning large conversations
- How hashtags start trending (likely a popular post originating the hashtag is retweeted and/or replied to).

Creating a system that can substantiate these example insights would be a good measure of the success of our work. It would justify this shift of focus from geographical location to personal context and interaction, and potentially spawn a new genre of Twitter visualizations.

1.4 Environment

Characterize environment you intend to operate in. Does your project operate on Amazon Web Services? on the general Internet? in a data center?

We intend to operate using Amazon Web Services. TwitCon(WT) consists of an always-on EC2 instance that is responsible for "listening" to Twitter via Fabric and distilling the presentable information from it, a Hadoop cluster that caches results such that users can view the current live stream over a variable-range historical context (up to several weeks), and a Web UI that allows users to view and interact with the data. The webpage will likely be hosted on the EC2 machine; implementation details can be fleshed out on the fly. Visualizations could potentially employ libraries/tools such as d3.js [1].

2. REFERENCES

[1] M. Bostock. d3.js - data-driven documents, 2014.

[3] Yes, this is an xkcd citation. No, this citation doesn't actually directly apply to our argument. Yes, we should probably get a better source.

[2] L. Daly. How big is the average epub book?
[3] D. D'Orazio. Map tweets in real time across the world with tweet ping.
[4] F. Ernewein. Tweet ping, 2013.
[5] Frog Design. A world of tweets, 2010.
[6] S. LaPorte and M. Hashemi. Listen to wikipedia, 2014.
[7] R. Monroe. Heatmap.
[8] Tweet Beam. Tweet beam, 2011.
[9] Twitter Inc. About Twitter, Inc., 2014.
[10] Twitter, Inc. Twitter developers, 2014.