Mining Twitter

Alexy Khrabrov and Gabriel Stocco
{*alexy,gfs*}@dartmouth.edu
Thayer School of Engineering
Dartmouth College
8000 Cummings Hall
Hanover, NH 03755

***Abstract*—Social Networks generally provide an implementation of some kind of groups or communities which users can voluntarily join. Twitter does not have this functionality, and there is no notion of a formal group or community. We propose a method for identification of communities and assignment of semantic meaning to the discussion topics of the resulting communities.**
Using this analysis method and a sample of roughly a month's worth of Tweets from Twitter's Gardenhose feed, we demonstrate the discovery of meaningful user communities on Twitter.

***Index Terms*—Twitter, Gardenhose, Community Identification**

I. Introduction

Twitter is a microblogging site and social network. Users of Twitter post or 'tweet' short (up to 140 character) messages. Users can follow others tweets and interact with others using the directed messages (denoted by @) and hashtags (denoted by #). Twitter provides a conventional social graph of news updates. However, due to its ubiquity, Twitter can also be treated as a sensor. Many individuals update their status often enough to provide background activity

Unlike many other social networks Twitter does not provide any kind of explicit community or group organization structure. As such, any notional community must be constructed solely from the directionality of the messages on the network and the content of those messages.

II. Lucene

Lucene is a powerful open-source search engine which uses flat-file inverted indices to provide fast full-text indexing and searching. The fundamental unit in Lucene is a document, which can have any number of field, each with a number of indexing parameters.

We created a custom analyzer for Lucene allowing us to preserve the meaningful special characters that appear in Tweets. The indexer tokenizes the text on characters which are neither letters, numbers or the @ or # symbols and creates a full-text search index on those tokens. We also store meta-data for each tweet, including date the tweet was posted to the twitter stream, if the tweet contains a hashtag, if the tweet is directed and who it was directed to..

We created two Lucene indices on a month's worth of tweets from the Twitter Gardenhose stream. The first index was created with each tweet as a document. A second index was then created on each user's corpora as a document. Using these two indices we created a number of API functions to perform textual analysis on the Twitter data.

A. Full Index API Functions

In order to ensure that we would receive meaningful results from the analysis of the twitter data we created a custom stop word list generator which would generate a stopword list based on the whole corpora of words used in tweets. Taking this stop word list and pruning out some entries by hand that were clearly meaningful we created a number of functions which performed textual analysis on the dataset.

1. topChatters - gets the users with the most conversation messages
2. twitUserIds - gets all the users and twits for a term
3. topWordsByTwit - gets the top words in the combined corpora of the list of tweets
4. topWordsByUser - gets the top words in the combined corpora of the list of users
5. userPairTopics - gets the top words for pairs of users
6. getStopList - creates a stoplist from the most commonly used words in the corpora of all tweets

B. User Corpora Index API Functions

In addition to the API functions which interact with the full index on a per tweet basis we created API functions which operate on the per user corpora index. These functions primarily use queries with very large numbers of terms to performcosine similarity analysis on the corpora of a user or a group of users against another user or group of users. Through this procedure we are able to compare a candidate to join a group to the corpora of the group as a whole in order to rank users who are considered to be on the fringe of a community for inclusion into the community.

III. Analysis Procedure

Given a search term, we can find either weak pairs: all the pairs of users who talk to each other and additionally each also use the search terms; or strong pairs: all the pairs of users who each individually use the search terms with at least one message between the two users containing the search term.

Having gathered this list of user pairs we then grow communities around each user pair. A user is a member of a community if they have been sent messages by two of the users of the community. Users are added recursively until there are no more users to add. Any users which have been sent messages by just one of the members of the community is considered to be the ’fringe’ of the community.

Finally we perform SIP and collocation analysis on these communities to find the resulting topics of conversation in the communities.

IV. Results

Using the aforementioned analysis procedure we found resulting communities and their related collocations and SIPs for a number of search terms. The graphs for these networks are large, and the full versions can be viewed at HTTP://WEBSITE.ADDRESS.

A. Glenn Beck

Glenn Beck is a controversial conservative cable television show host on Fox News. We performed a strict pair search for users who used both ”Glenn” and ”Beck” in a single tweet to find politically charged communities on Twitter. Given these communities we then performed a Collocation analysis.