Twitter Data Analytics – Analyzing Tweet Networks

Bychapura Parameswaraiah, Chandan (cbp140230)
Jiang, Ran (rxj131530)
Kuncham, Venkata Keerthana (vxk142230)

Department of Computer Science
University of Texas at Dallas
Richardson, TX 75080

***Abstract*— we analyze information propagation on Twitter social media. We specifically focus on identifying the most important people in flow of information through the twitter social media by building re-tweet network graph. Furthermore we analyze the visibility decay of tweets by analyzing the tweets time line. In the end we determine the polarity of tweets as positive, negative or neutral by mining the tweet text.**

Keywords—information propagation; twitter data analytics;

I. INTRODUCTION

It's well-known that Twitter's most powerful use is as tool for real-time journalism. Trying to understand its social connections and outstanding capacity to propagate information, we have developed a model to identify the evolution of a single tweet over a period of time. By doing so, we identify the most important people who are controlling the flow of information in this network. Insights about the most influential people on a retweet network will help analyzing the role of the users' activity in a community and if necessary change the direction in which the information is influencing the social community.

From all the content that people create and share, only a few topics manage to attract enough attention to rise to the top and become temporal trends which are displayed to users. By sampling the tweets under specific time interval we can visualize the trend line for a tweet or group of tweet over a period of time. Trend lines for tweets help us to identify the existing, persistence and visibility decay of trends among users of social media.

Tweets are frequently used to express tweeter's emotion on a particular subject. There are firms which poll twitter for analyzing sentiment on a particular topic. We finally analyze the sentiment of the tweets for a topic to determine the attitude of mass to a particular event or a topic under discussion.

II. DATA COLLECTION

A. *Tweets for a Query String*

As a first step, we collected streaming data from twitter using Twitter 4J for a particular event. For our project, we collected tweets for recent terrorist attack on Paris using "paris attack" as query string to twitter API.

B. *TOPIC modeling and Tweets collection for a TOPIC*

On a micro blogging site such as Twitter, the users have different perspective about an event and react differently to a particular situation. The collected tweets are combination and aggregation of all these different dimensions in which users view a situation and react to the same. Therefore, it becomes extremely important to identify the different dimensions in which the social media users are discussing a event.

In our example of "Paris attack", we found set of users more interested in discussions about terrorist organization "ISIS" which caused the attack, few others were tweeting about the victims of the attack and few more discussing about the country's political reactions towards this attack. To help us perform in depth focused analysis, it becomes necessary to identify the different topics on which the users are expressing their views for an event.

For our project, we have used Latent Dirichlet allocation (LDA) topic modeling algorithm to classify the collected tweets for an event under different topics representing various perspectives of user reactions to an event. The LDA model is highly modular and can therefore be easily extended. We have made use of MALLET, an open source implementation of LDA algorithm to determine the topics. The accuracy of using mallet for LDA algorithm is believed to be about 80%.

For building re-tweet graphs and comparing the results, we collected related tweets, tweets belonging to same topic. The query string was modified accordingly using key words generated for a topic by LDA algorithm. The Twitter API was searched repeatedly for each key word of the topic to form the final data set. This data set is used for all our further analysis.

C. *Friend and Followers data*

For generating re-tweet network graph and to establish relationship among different users in the network, we require information about the friends and followers of each user in the network.

We obtained this information by querying twitter API for each user individually as the API has a limitation on number of calls which can be made within a 15 minute time interval. We collected approximately 200 friends and followers for each user in the network.

D. *User's history re-tweets data*

For generating re-tweet network graph and to estimate the most probably originator of the re-tweet, we require information about past re-tweets of each user in the network.

We obtained this information by querying twitter API for each user individually as the API has a limitation on number of calls which can be made within a 15 minute time interval. We collected approximately 30K history re-tweets for each user.

III. ALGORITHMS

The algorithms used to implement different features of this project are described below:

A. *Information Propagation Graph (understanding most important people in a re-tweet network)*

There are several dimensions along which one may be considered important on a social media network. Measures of importance in social networks are called "centrality measures". Here, we use three measures applicable to a re-tweet network that are used most frequently in social media analysis to understand the most influential people. Each provides a different view of who is important in the network.
All the measures below assume the existence of re-tweet network graph, where each node represents a user and the directed link B → A represents the possibility of user B re-tweeting user A.

- Degree Centrality: Who gets the most re-tweets? One of the most commonly used measure. The calculation for this measure is simple: Count the number of links attached to the node, this is the degree centrality. Naturally, the node with highest count represents one of the most important a user in a network. According to Fig 1, Alice is considered one of the most important person in the network.
- Eigen Vector Centrality: Who is most influential? Degree centrality answers key question which is to say "how many people re-tweeted this node?" Eigen vector centrality builds upon this to ask how important are these re-tweeters?

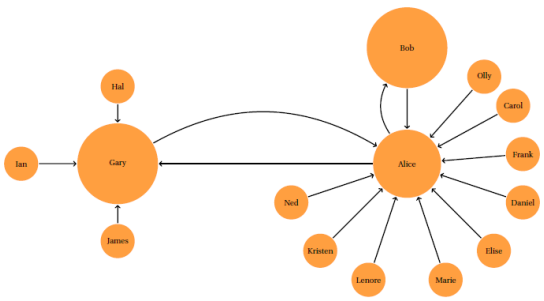

Figure 1: Example re-tweet network graph

In Figure 1, Bob who is largely ignored by degree centrality is also one of the important people controlling the flow of information because 'Alice' from whom the maximum re-tweets happened gets the information from Bob. Removing Bob from network would limit the flow of information.

- Betweenness Centrality: When information travels through network, it takes the most convenient path possible. The most convenient path in the a network is the shortest path. Betweenness centrality measures the shortest paths in which the user is in the sequence of nodes in the path.

All of these measures make use of a re tweet network graph. On Twitter, information spreads primarily through re tweeting. The resulting Tweet is called a re tweet. When we visualize re tweets we are essentially visualizing the flow of information in the network. Re tweets are marked by the characteristic prefix "RT" followed by the name of the user who originally published the Tweet. An important yet subtle property of this twitter data is that one can only identify the original source of the information and not the intermediate users along the information propagation path. If a graph is drawn using raw data provided by twitter, the graph would look like (b) in Figure 2, while in actual scenario, the information propagation occurred like (a).

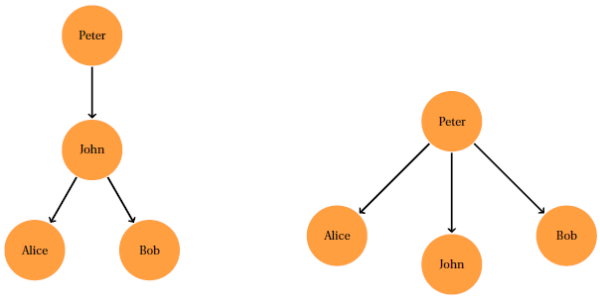

Figure 2: Re-tweet network graph example

Hence, for our project we have devised an algorithm, Figure 3, to build a directed graph for a re-network which provides a better approximation of how the information factually traversed through the network. We then analyze this graph to determine the important people on the re-tweet network by calculating the centrality measures discussed above.

B. *Modeling timeline for Tweets (persistence and decay of tweets)*

Twitter API provides a way to model time line of a user determining the tweets made by user over a period of time along with other related statistics. But, there is no processed data which concentrates on how a particular tweet or group of tweets evolved among twitter users over a period of time.

Information about the Trend lines for tweets help us to identify the existing, persistence and visibility decay of trends among users of social media. Visualizing the trends of the tweets might yield important information on the next course of

ALGORITHM

Step 1: Determine the original user (the owner) of the Tweet.
Step 2: Arrange re-tweets in the ascending order of their occurrence along with the user who made the re-tweet
Step 3: Assign the first user as child for the original user.
Step 4: For each of the re-tweet user after step 3, do
    Step 4.1: List all the users (*parentList*) who have re-tweeted prior to current user
    Step 4.2: compare *parentList* with the User's friends/follower list (*fList*)
        If (compare (*parentList* & *fList*) is Null)
            Assign original owner as parent node
        else if (compare (*parentList* & *fList*) == 1)
            Assign matched user from parentList as parent node
        else if (compare (*parentList* & *fList*) > 1)
            Step 4.2.1: compare *parentList* with the history re tweet count (*historyCount*)
                If (historyCount == 0 for all parentList)
                    Assign one of the users from parentList randomly as parent Node
                else
                    Assign user with highest historyCount as parent Node
Step 5: End

Figure 3: Algorithm for building re-tweet network graph

action that has to be taken for a social media reaction. We have modeled a sampling algorithm as explained in Figure 4.

ALGORITHM

Step 1: Determine the minimum timestamp and maximum timestamp of the tweet occurrence.
Step 2: Sample the total time into buckets each for a five minute interval.
Step 3: Assign each tweet to exactly one of the buckets
Step 4: Count the number of occurrences of each tweet in each interval.

C. Sentiment Analysis of tweets

Opinion mining is the task of judging whether a document expresses a positive or a negative opinion or no opinion regarding a particular object or topic. For Sentiment analysis of tweets, we use strategy of bag-of-words model. We create lists of 'positive', 'negative' or 'neutral' words and judge a document based on whether it has a preponderance of positive or negative words (and judge it neutral or 'objective' if it has few words of either category.

Since the text in tweets is fairly limited the bag of words strategy showed consistent results.

IV. TOOLS AND TECHNOLOGIES

a) MALLET: open source tool for LDA algorithm. Used this tool for topic modeling

b) Twitter APIs (OAuth): to crawl twitter social media and collect the real time tweets data and user information.

c) Rest Web Service: to communicate with twitter server

d) Apache Spark: for data processing and analytical result derivation.

e) Cassandra: for row by row result set data processing.

V. CONCLUSION

We have outlined an algorithm to model a re-tweet network and identify the most important people who control the flow of information in a mass social media such as Twitter. The algorithm provides fairly accurate results when tested on very small network. Experiments have to be conducted to order to determine the accuracy and consistency of the algorithm used. More comprehensive machine learning techniques such as maximum likelihood estimators can be employed to further enhance the accuracy of the algorithm.

The timeline analysis of the tweets fairly determines the evolution of tweets or group of tweets over a period of time. This can be further analyzed and different dimensions such as geographic location of the tweets; user category etc. can be introduced to this analysis to gain deeper insights.

The data collection is largely limited by the restriction on number of API calls which can be made in order to collect large data set.

REFERENCES

[1] Twitter data Analytics, Shamanth Kumar, Fred Morstatter, Huan Liu. (material *references*)

[2] Information propagation on Twitter, Eldar Sadikov, Maria Montserrat Medina Martinez.

[3] http://mallet.cs.umass.edu/. Open source LDA implementation toolkit.