

**Analyzing domain-based influence of user in Twitter and predicting the
most influential user in specific domain**



Project Proposal

by

Srushti Gangireddy

CWID: 805601788

CPSC 597

Project Advisor: Dr. Anand Panangadan

Project Reviewer: Dr. Doina Bein

Department of Computer Science

California State University, Fullerton

Contents

1. Abstract
2. Introduction
3. Requirements description
4. Design description
5. Implementation
6. Test and Integration
7. Installation instructions
8. Operating instructions
9. Recommendations for environment
10. Bibliography

1. Abstract

Social networking sites today has become the broadcast tools for information. They play the fundamental role as a medium to spread influence among its members. Word-of-mouth propagates information at a quick rate. There are many sites these days that people use to build the relations to share their personal interests, options and emotions. A decade ago, many companies were hiring different advertisement agencies to promote their brand/ product. The introduction of ad-blockers and ad-skip features in many applications help people to skip the advertisements. Social media websites, blogging sites has become a way of product broadcast in the digital era. Twitter, is one of the most-popular social networking service where in many people connect to each other and share their opinions or information on various topics. The information from various source propagates through different paths and reaches each user through one or multiple connections. Some people has the power of highly popularizing a brand or other product among their connections. Influence can be studied as the level of impact of something on the user. Twitter follows the “publish-subscribe” model. One user can “follow” other users and can later get notified by all their updates. The goal of this project is to predict the most influential user in specific domain by collecting the tweets of specific domain.

2. Introduction

a. Problem Definition:

Social influence captures the way people in social media influence each other by stating their opinions/ emotions/ behavior. In this project, we studied how user’s behavior is influenced by friends in their network and we tried to predict the influential user for a topic. The topic we chose for our study is “California Shooting”. We collected all the tweets related to California shooting over the period of a week when this event happened.

“Twitter influence” has been the interesting study to research over the past few years although there is ambiguity about quantifying influence. Some services such as klout, peerindex, twitter grader have gained attention by giving a numerical score that quantifies user’s social influence. Though there has been many studies no single study or algorithm has emerged as the best choice. Our focus is to answer the question: “Who is influential among the group of people speaking about certain topic?”. Early research studies in this area has found out that influencers play a major role in information propagation. Most influential twitter users are many orders of magnitude more influential than the average user and these users have the ability to spread information to larger audience. Cha et al. conclude that utilizing top influential users has potential benefits in marketing strategy [1].

Factors of influence for a tweet in twitter:

Recognition (Retweets)

Preference (Tweet likes)

Novelty (Number of out-links in a tweet)

Eloquence (Length of a tweet)

Factors of influence for a user in twitter:

Follower – following network

Due to the complexity and scale of twitter social network it is difficult to obtain fine grained information to measure the influence. Many studies have focused on measuring influence in the terms of number of followers, but some researchers have found out that only 22% of Twitter relationships are mutual [2].

Motivations for our study:

1. Explore how influence is exerted among people who are not personally acquainted.
2. Create accurate way to predict which twitter user can best propagate a message to others.

Defining influence:

“The ability to through one’s behavior on Twitter, promote activity and pass information to others”. There are many different ways in which users respond to the content of other user’s that demonstrate that they are influenced. User’s own tweet/ own behavior does not demonstrate the influence. However if a user tweets and many other users re-tweets the tweet, if the user introduces a new topic that is then discussed by others or if the sentiment of the content in user’s tweet is reflected in the re-tweets, then these are the examples of influence.

For our analysis, we considered the population to be the group of people speaking about “California Shooting”.

Hypotheses which direct our research:

1. A user’s total number of followers is not a strong predictor of user’s influence.
2. There exists meaningful influence effects apart from retweets, mentions

b. Project Objectives

1. To collect the twitter dataset and analyze the network structure of users as they discuss a relatively unpopular topic.
2. To identify the historic tweets related to this topic.

3. To detect the start of the discussion/ tweet about that topic and model the path of its propagation along the network. (This helps in understanding the path of information propagation of topic).
4. Study the information exchange graph to understand the influence of users on each other.
5. Find the most influential user in specific domain by studying all the possible factors that we think will play role to measure influence.
(Number of followers, retweets, mentions, information exchange graph, propagation time of message from some user)

c. Development environment

Programming Language:	Java Python 3.5
Tools and environment:	Eclipse as IDE for Java Code Pycharm as IDE for Python Code Tweepy API (Twitter API for python) Twython API (Twitter API for python) Networkx (Graph API in python) StanfordCoreNLP (To study the sentiment) Flask (Python web framework) Pandas (Python data analysis toolkit)
Operating System for development:	Windows
Hardware:	Intel CORE I7 Dell Inspiron 13 7000 Series

d. Objective of the study

The main objective of this study is to:

Evaluate existing studies and their suitability to identify influential users on twitter.

Provide the innovative approach to measure influence and rank users in twitter.

This is important since many previous approaches assessed influence based on number of followers and retweet count. Both approaches base measurement just on few parameters. Each of the studies has their own perspective to measure influence. Our goal is to study the group of tweets speaking about certain topic and find the influential user in that group of people by constructing the graph with users as node and degree if any of the retweets in the group of tweets are by the people in the user's network. We measured centrality on the graph formed and defined the central user as the most influential one.

e. Review of significant research

Influence is of great importance. If the influential users are correctly discovered, it can lead to lot of benefits in many different areas as politics, fashion, technology, education, marketing etc. Till now, studies have shown that Twitter has provided very good prediction capabilities. Since various companies are twitter users, free marketing can be done. Starbucks case study provides the interesting example. According to the case study, Starbucks has employed many social media tools as marketing instruments for customer knowledge management. It has redefined the role of its customers are contributors for innovation. Studies as these enforce the need to improve the work being done on influence and recommendation on social media.

Another research studies with the goal to find the backbone in the social network and use it for viral marketing. Sparsification of influence networks is the framework built by Yahoo Inc. and is a greedy algorithm which eliminates the links that do not play important role on how the information is propagated. It reveals the noise and accurately predicts the influencer. To this end, researches are based on the link existing in the graph of social network of a specific user or the retweet count of specific user.

f. Assumptions and limitations

Twitter API is limited by the rate limits. Also, the actual retweet propagation is not preserved and the source of the tweet is always the first tweeter. It is difficult to find the actual influencer in the network as all the re-tweets point back to the first tweet. We trace the path of propagation and find consider the relationship between the user and the tweet-retweet relationship to construct the graph and find the central user.

3. Research approach / methodology

We collected the tweets speaking about "California Shooting" as mentioned in the introduction. We tried to construct a graph with users as nodes and the edge between users if there exists a retweet between user A and user B, but the data revealed by the twitter API hides some useful parameters and also as stated in the limitations the source tweeter of the re-tweets is always the first tweeter irrespective of the path of its propagation.

We use python twython API to collect Tweets discussing the “Cal shooting” event and stored the tweets in MongoDB. MongoDB is the document based database so it was easy to store JSON tweets as documents in MongoDB. We cleansed the tweets collected in Java using rich data structures like hashmap, linked list and array list. We preserved only the tweet content, tweet information and the source tweeters information for our study.

We collected the list of the followers of all the users speaking in the tweets for our analysis. Since the number of users is 5672, it took a lot of time to collect the complete list of followers. We used multi-threading and allowed each thread to access Twitter API with different keys to avoid the rate limit errors.

After collecting the complete list of tweets, tweeter information and the follower list of all the tweeters, we formed a network/ graph with users as nodes. We drew an edge between user A and user B if user B has retweeted the Cal shooting tweet of user A and if user B is in the network of user A. We thought this would help in measure the actual influence than measuring the number of retweets of each user.

After forming the graph, we calculated the degree centrality and betweenness centrality of each user and labelled the users with more centrality to be more influential.

We used Folium which is the python library for tweet propagation visualization on twitter.

We developed a simple UI in flask, the python web development framework, html5, bootstrap and simple CSS.

4. Research results and analysis of results

Data processing:

Data collection and hardware setup

We need an application registered with Twitter to extract the data from Twitter. When the application is registered the keys are generated which lets us access the twitter API programmatically. The code to collect the tweets speaking about “California shooting” is written in python. The streaming API of twitter is used to pull the data continuously. The application is run on Ubuntu server. The code creates multiple threads to fetch the data of various tweeters and this helped in fetching more data. In case of hardware failure, we used amazon load balancer to use another server to gather the data. By default, twitter does not make all of the data to its developers unless you have “firehose” access.

What data is gathered

6124 tweets speaking about the California shooting events are gathered.

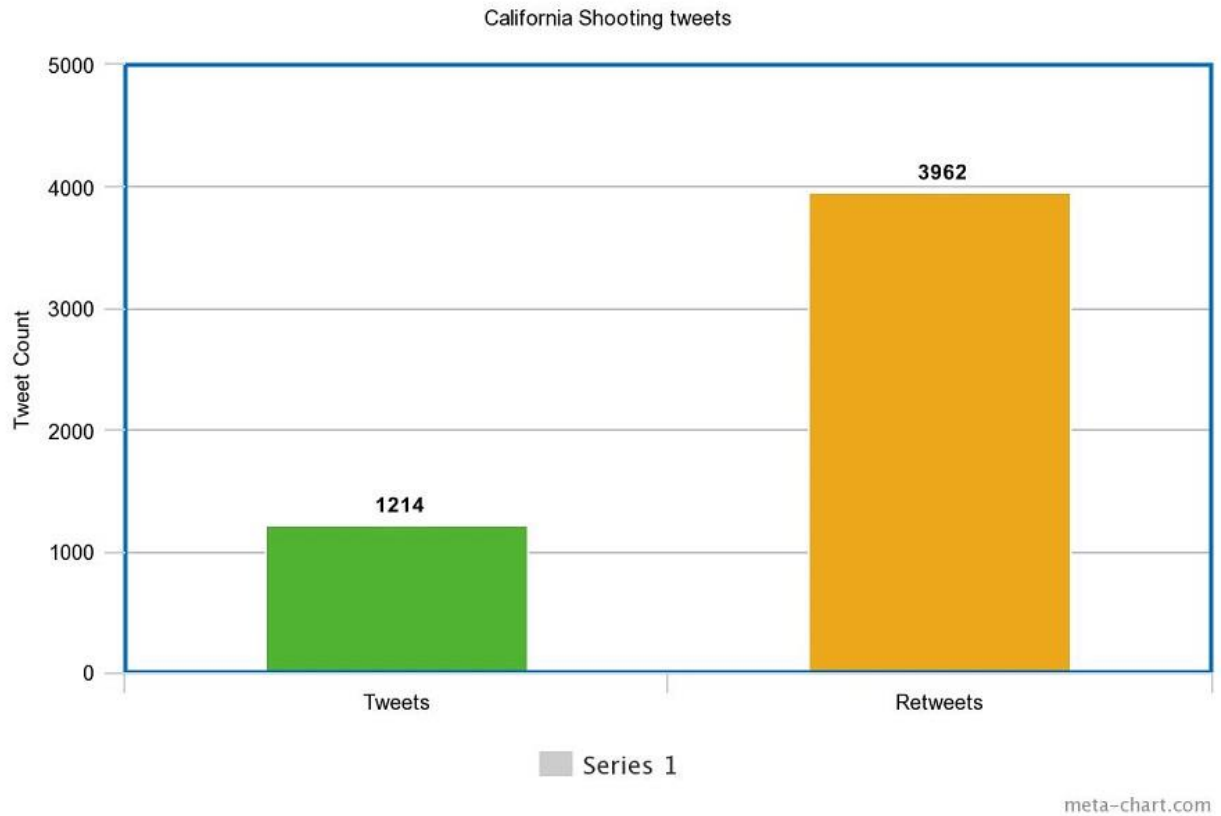


Image 1. Tweets and retweets

Tweets

TweetID	Tweeter	TweetContent	RTCount
930624906399834112	Rose4Austin2018	RT @kwilli1046: This is a tragic event for all of us. There are children involved. #CaliforniaShooting takes emotional toll on offiâ€¦	0
930624907138011136	FebAngelNYR	RT @kwilli1046: This is a tragic event for all of us. There are children involved. #CaliforniaShooting takes emotional toll on offiâ€¦	0
930624915543347200	armmoin	RT @JTClements_OF: Welcome to America, where Mass shootings are more common than respect and kindness. If,? donâ€™t understand anymore #Califoâ€¦	0
930624923852267521	Dawnhomes	RT @DeepStvte: UPDATE! Officials NOW say (backpedaling from 2 children killed) NO ONE was killed at the school but a â€œnumberâ€? of sâ€¦	0
930624938763014144	working_tg	RT @DeepStvte: SHOCKING VIDEO: Northern California resident (REHEARSES) describing his truck being stolen by shooting suspect, WHOâ€¦	0
930624941191462912	amandaowens1969	How terrible. https://t.co/vdXFHrUdBy	0
930624998326149120	ImJustAMel	RT @kwilli1046: This is a tragic event for all of us. There are children involved. #CaliforniaShooting takes emotional toll on offiâ€¦	0

Image 2. Tweets

Processing the data

The data collected is stored in JSON format as documents in MongoDB. Using the java application, the raw data is cleaned and only the information relevant to our study is saved.

We first set up a hadoop map reduce job to analyze the users with the more followers to analyze their role as the influencer.

```
{"XHNews": 11434236}  
{"firstpost": 1889608}  
{"SirJadeja": 917461}  
{"AmarUjalaNews": 645925}  
{"mid_day": 598716}  
{"moneycontrolcom": 571821}  
{"Tehelka": 516431}  
{"JagranNews": 470814}  
{"ecr9495": 404700}  
{"FinancialXpress": 394886}  
{"CopelandNetwork": 320736}  
{"haaretzcom": 306970}  
{"cnalive": 295348}  
{"RadioPakistan": 275231}  
{"WorldMedia4u": 228727}  
{"pablovillaca": 218613}  
{"Zee_Hindustan": 196461}  
{"ECR_Newswatch": 171102}  
{"KSLcom": 163641}
```

Image 3. Users ranked by their followers number

We then ranked the users by their retweets count.

```

4uslimIQ 1268
DeepStvte 1082
<willi1046 372
TrishaDishes 198
Dumptrump33 124
PeachyAmerican 43
JennJacques 37
SirJadeja 31
SpokesmanTweets 30
SGTreport 28
VickFury2018 21
GroovyFeline 21
BrownEyes_1023 20
tpowers6pack 19
TrueNevvs 18
VewtownAction 17
TheRealPockets 17
Rimiler 15
almostjingo 15
denudedmedia 13
CitizenKays 13
CopelandNetwork 11

```

Image 4. Users ranked by their retweet count

Limitations

It was not possible to predict the influence of the user based on the number of likes of their tweets or the number of actual re-tweets as the actual retweet path is not preserved in the data.

Some of the tweets were missing in between. So we have decided to perform centrality measure on the network of users and define more central user as the influencer.

Centrality measure

Graph: A graph G is an object formed by the set of vertices and the set of edges that connect the vertices. The vertex set and the edge set of the graph are denoted by $V(G)$ and $E(G)$. If two vertices are joined by the edge they are said to be neighbors.

Social network can be conveniently modeled by a graph. Users can be represented as the vertices and the relationships between them can be represented by edges.

In most networks, some edges or vertices are more central than the others. To quantify this, centrality indices were introduced. There are many ways to quantify relative importance of a node in the network. The most basic centrality measure is the degree of the node and the most common factor of other centrality measures is the distance between nodes in the network. Standard centrality indices are betweenness, closeness and eccentricity.

Centrality: Centrality is the characteristic of individuals in social network. The centrality score conveys how the user fits in the overall social network. Specific users with high centrality score are often key canals of information. Finding out centrality score helps in spreading the information in the social network faster.

Degree centrality: Degree centrality is the simple measure that counts how many neighbors a node has. We have two measures of degree in directed node: in-degree and out-degree. In-degree measures the number of incoming links and out-degree measures number of outgoing links. A node is important if it has more neighbors or many in-links.

Betweenness centrality: Betweenness centrality measures the extent to which a node lies in the paths between other nodes. Vertices with high betweenness is proved to have high influence within the network because of their control on information passing between other nodes. The removal of these nodes may cause the disrupt in the communication between other nodes in the network.

$$C_b(i) = \sum_{j < k} g_{jk}(i) / g_{jk}$$

G_{jk} = number of geodesics connecting j k

G_{jk} = number of geodesics that the node “i” is present in

Geodesic path is the shortest path through a network between two vertices.

Normalized by

$$C_b^*(i) = c_b(i) / (n-1)(n-2)/2$$

Closeness centrality: Closeness centrality measures the mean distance from a vertex to other vertices.

Twitter graph: In our project we constructed graph with users as nodes and the edge between user A and user B if user B is in the network of user A and if user B has re-tweeted at-least one cal-shooting tweets of user A. After constructing the graph, we perform the degree centrality measure and betweenness centrality of the users in the network and rank the users in the order of their centrality.

Results:

Below are the results of the users ranked by degree and betweenness centrality measures.



Image 5. Frontend of project

Locations from which users tweeted



Image 6. Location of the users who participated in the discussion of California shooting

Graph of users.

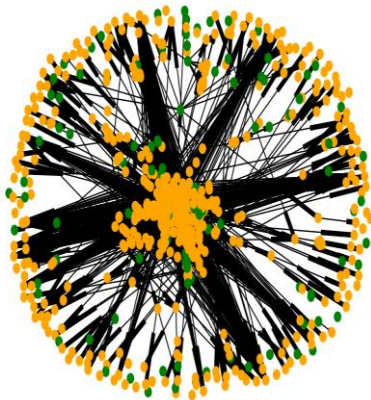
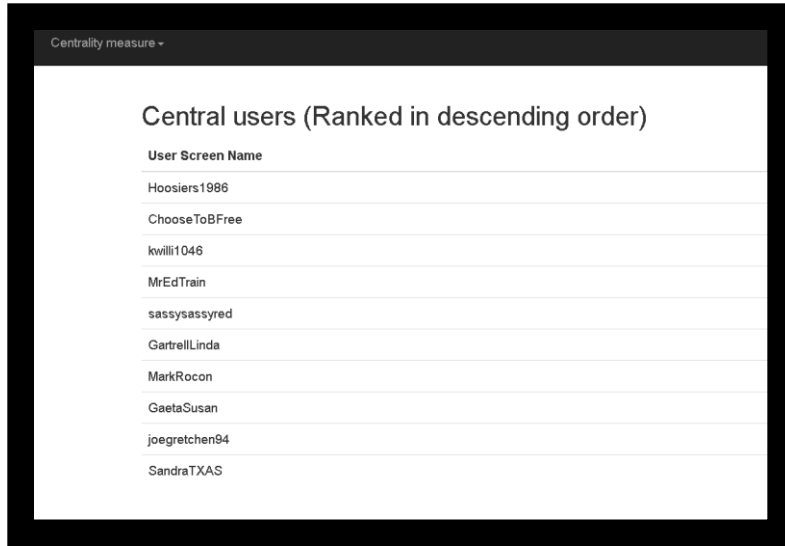


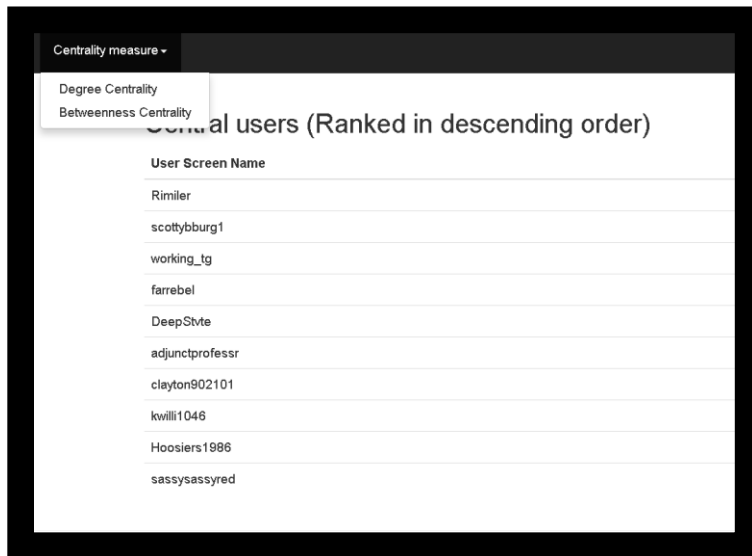
Image 7. Graph of users



The screenshot shows a web interface with a dark header bar containing a dropdown menu labeled "Centrality measure". Below the header, the title "Central users (Ranked in descending order)" is displayed. A table lists the top 12 users by degree centrality.

User Screen Name
Hoosiers1986
ChooseToBFree
kwilli1046
MrEdTrain
sassysassyred
GartrellLinda
MarkRocon
GaetaSusan
joegretchen94
SandraTXAS

Image 8. Users ranked by their degree centrality measure



The screenshot shows the same web interface as Image 7, but with the "Centrality measure" dropdown menu open, showing "Degree Centrality" and "Betweenness Centrality". The table below lists the top 12 users by betweenness centrality.

User Screen Name
Rimiler
scottybburg1
working_tg
farrebel
DeepStvte
adjunctprofessr
clayton902101
kwilli1046
Hoosiers1986
sassysassyred

Image 9. Users ranked by their betweenness centrality measure

5. Project flow

1. Collect the keywords to collect the tweets speaking about California tweets.
(#Calshoot,#Calshooting,#Californiashooting,#ShootinginCalifornia,#NorthernCalshooting).

2. Run the script `collect_tweets.py` with the search words as the arguments to collect the tweets related to certain topic and store them in the local mongodb database on the server.
3. Import the tweets in mongodb to the system as the json file.
4. Run `calshootinganalyzetweet.java` to parse the json tweets and analyze the tweets.
5. Run `calshootingtweetsretweets.java` to get the tweet and retweet count.
6. Run `calshootingparse.java` to get the list of users with top follower count and the list of users with top retweet count.
7. Run `GetTweeterFollowers.py` to get the list of followers of each tweeter that participated in the discussion. (This script runs many scripts with multiple threads inside and may take long time)
8. Run `TweetsOnMap.py` to visualize the tweeters location on the google maps.
9. Run `propagateTweetOnMap.py` to construct the graph and perform centrality measure.
10. Python/Web-application folder has the code for the web front end part. Run `__init__.py` to start the web application and see the list of influential users.

Note: There are many other scripts that perform analysis on the tweets and the list of tweeters. For ex, `fetchTimeline.py` in `activeusers` folder gets the number of tweets the users has posted from the beginning and gives the list of users ranked in the order of their tweets count.

6. Conclusion and future work

From the analysis of the results it can be concluded that the influence of the user is dependent on their network structure and their follower – following relationship as well as the tweet- retweet relationship. Many research studies indicate that different types of network centralities generally identify the same nodes to be important. The importance given to each node depends on the centrality measure. Leaf nodes have importance in distance centrality but really less in betweenness centrality. No centrality measure is best for all situations. It is important to understand the network structure and select the centrality measure that is more appropriate. Many changes can be made to the web application. Along with the network structure tweet content and the attributes of the tweet like hashtags, links, urls and references can be considered to determine the influence of the user.

7. References

1. Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In Fourth International AAAI Conference on Weblogs and Social Media, May 2010.
2. Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In WWW '10: Proceedings of the 19th international conference on World wide web, pages 591-600, New York, NY, USA, 2010. ACM.
3. Chua, A. Y. & Banerjee, S., 2013. Customer knowledge management via social media: the case of Starbucks. Journal of Knowledge Management, pp. 237-249.
4. Abellera, L. V., & Panangadan, A. (2014-2015). OCICATS (Online community input classification to advance transportation services) – a GIS-based decision-support tool

Deployment details:

The steps to run the project are as below:

1. Download the source code.
2. Install Pycharm as Python IDE and Eclipse as Java IDE.
3. Import java source code to Eclipse and python source code to pycharm.
4. Install tomcat as server container.
5. Add the server to IDE.
6. Run the project and files.

Project Schedule:

2017	August				September				October				November					Summary	
Tasks:	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	5	Hours	Percent
Requirements	10	10	5	5														30	16.25%
Design			10	10	15	10												45	17.24%
Implementation					5	5	15	15	15	10	10	10	12	8				95	46.79%
Test							3	3	3	5	5	5	3	3				30	10.83%
Readme														4				4	1.97%
Demonstrate																	4	4	1.97%
Documentation													2	2	3	3		10	4.92%
Hours	10	10	15	15	20	15	18	18	18	15	15	15	17	17	3	3	4	210	100.0%

