International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 1 | Issue – 5

Seismic: A Self-Exciting Point Process Model for Predicting Tweet Popularity using Hashtags

Karthick.D
SRM University, Katankulathur,
Chennai, Tamil Nadu, India

Dr. G. Vadivu
SRM University, Katankulathur,
Chennai, Tamil Nadu, India

ABSTRACT

In existing paper they had used a full month of Twitter data to evaluate SEISMIC .In which the original data set contains over 3.2 billion tweets and retweets on Twitter from Octobor 7 to November 7, 2011.Also they only kept tweets such that it has at least 50 retweets, the text of the tweet does not contain a pound sign # (hashtag), and the language of the original poster is English. There are 166,076 tweets satisfying these criteria in the end. So here we are going to propose the mining of tweets with a particular #hashtags and going to formulate the number of retweets in an efficient manner ,so that it will be more efficient in terms of organizing particular categories while mining the popularity of retweets.

Keywords: *Information diffusion; cascade prediction; self-exciting point process; contagion; social media*

INTRODUCTION

Online social networking services, such as Facebook, Youtube and Twitter, allow their users to post and share content in the form of posts, images, and videos As a user is exposed to posts of others she follows, the user may in turn reshare a post with her own followers, who may further reshare it with their respective sets of followers. This way large information cascades of post resharing spread through the network.

A fundamental question in modeling information cascades is to predict their future evolution. Arguably the most direct way to formulate this question is to consider predicting the final size of a information cascade. That is, to predict how many reshares a given post will ultimately receive.

Our model gives only 15% relative error in predicting final size of an average information cascade after observing it for just one hour.

LITERATURE REVIEW:

The study of information cascades is a rich and active field. Recent models for predicting size of information cascades are generally characterized by two types of approaches, feature based methods and point process based methods. The process of adopting new innovations has been studied for over 30 years, and one of the most popular adoption models is described by Diffusion of Innovations. Much research from a broad variety of disciplines has used the model as a framework mentioned several of these disciplines as political science, public health, communications, history, economics, technology, and education, and defined Rogers' theory as a widely used theoretical framework in the area of technology diffusion and adoption. [1]

It is widely accepted that some time after the occurrence of a major earthquake the aftershock activity dies off and background activity surpases the aftershock activity. Prior to the next major earthquake, preseismic quiescence and then foreshocks are expected to appear in the focal region. Thus the seismic quiescence and related seismic gap have been studied by many seismologists for the purpose of earthquake predictions, from this it is also useful for the prediction of the the information cascade of future prediction of final popularity of retweets.[2]

Social network services have become a viable source of information for users. In Twitter, information deemed important by the community propagates through retweets. Studying the characteristics of such popular messages is important for a number of tasks, such as breaking news detection, personalized message recommendation, viral marketing and others. We cast the problem of predicting the popularity of messages into two classification problems: 1) a binary classification problem that predicts whether or not a message will be retweeted, and, 2) a multi-class classification problem that predicts the volume of retweets a particular message will receive in the near future. [3]

Modeling and predicting retweeting dynamics in social me-dia has important implications to an array of applications. Existing models either fail to model the triggering e ect of retweeting dynamics, e.g., the model based on reinforced Poisson process, or are hard to be trained using only the retweeting dynamics of individual tweet, e.g., the model based on self-exciting Hawkes process.Our model is motivated by the observation that the retweeting process of tweets could be generally characterized by a di_usion tree with only a handful of *key* nodes, each triggering a high number of retweets.[4]

Retweeting is the key mechanism for information diffusion in Twitter. It emerged as a simple yet powerful way of disseminating information in the Twitter social network. One interesting emergent behavior in Twitter is the practice of retweeting, which is the relaying of a tweet that has been written by another Twitter user. This can be done in one of two ways. First, one can retweet by preceding it with RT and addressing the original author with @. For example, “RT @userA: my experience with the new #iPad is great!” Second, Twitter also enables users to retweet easily with one-click. [5]

EXISTING SYSTEM:

In present they had used a full month of Twitter data to evaluate SEISMIC .In which the original data set contains over 3.2 billion tweets and retweets on Twitter from Octobor 7 to November 7, 2011.Also they only kept tweets such that it has at least **50 retweets**, the text of the tweet does not contain a pound sign # (hashtag), and the language of the original poster is English. There are 166,076 tweets satisfying these criteria in the end

PROPOSED SYSTEM:

So here we are going to propose the mining of tweets with a particular **#hashtags** and going to formulate the number of retweets in **an efficient manner**, so that it will be more efficient in terms of organizing **particular categories** while mining the popularity of retweets.

METHODOLOGY:

In this project we are going to formulate the total number of retweets i.e, the popularity of the tweets with the help of SEISMIC (self exciting point process) algorithm .Initially with the help of the twitter application we made an authentication, with the help of Rstudio. In which it requires the following packages such as Seismic, twitter, devtools. After Installing necessary packages, then we have to retrieve the data from the twitter, generally we can search tweets in twitter using searchtwitteR(“example”). Then by defining the retweet pattern we were going to fetch the tweets that has been retweeted, over a period of time with the certain attributed such as, tweeted, time, date, is retweeted. The major methodology is that we were retrieving the data with hashtags, by mentioning the number of tweets that has to be retrieved. So After retrieving the tweets we are going to store it in a data frame to write it to a csv file. After the completion now we could be able to analyze the content graphically in exploratory with the total number of estimated parameters. The sequential process of the methodology is given with the flow diagram below.

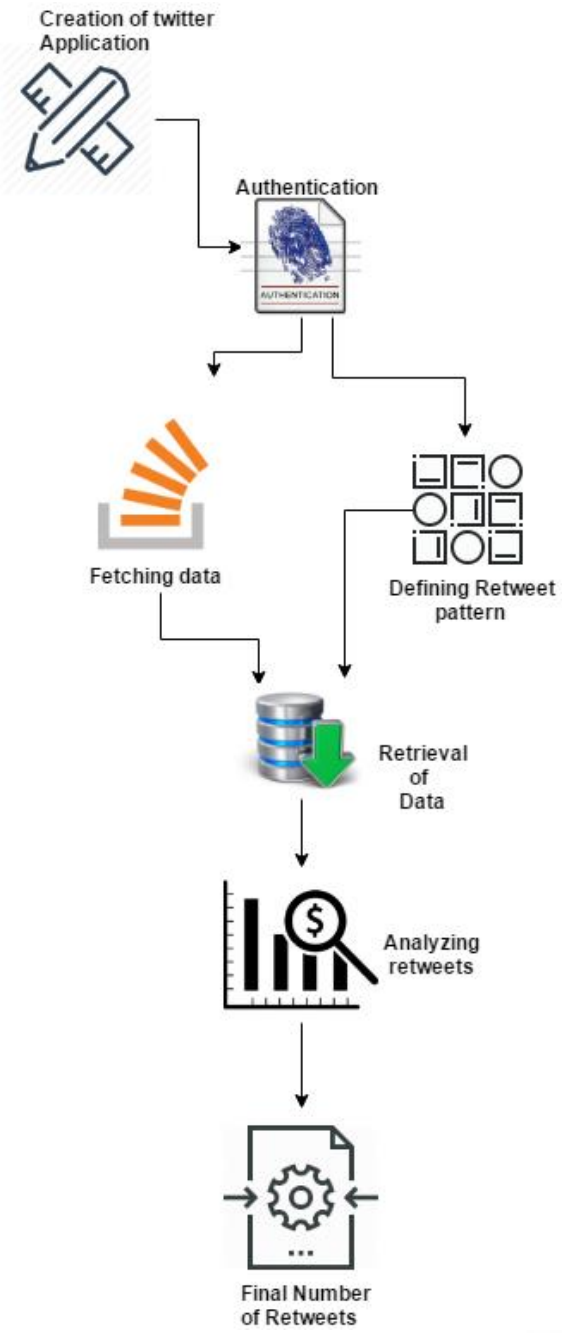

FIG 1: Methodology

ALGORITHM USED:

SEISMIC:

Purpose: For a given post at time t, predict its final reshare count

Input: Post resharing information: ti and ni for i = 0; : : : ;Rt.

SEISMIC models the information cascade as a self-exciting point process. In a self-exciting point process, each reshare not only increases the cumulative count by one, it also exposes new followers who may further reshare the post. This property is ideal to model the "rich get richer" phenomenon in information spreading.

SEISMIC implements a fast kernel weighted method to estimate the temporally evolving infectiousness, which fully characterizes an information cascade. Roughly speaking, it measures how likely the post will be reshared at that time. Then, if the infectiousness is smaller than a threshold, SEISMIC can accurately predicts the final popularity of the post.

NETWORK DIAGRAM:

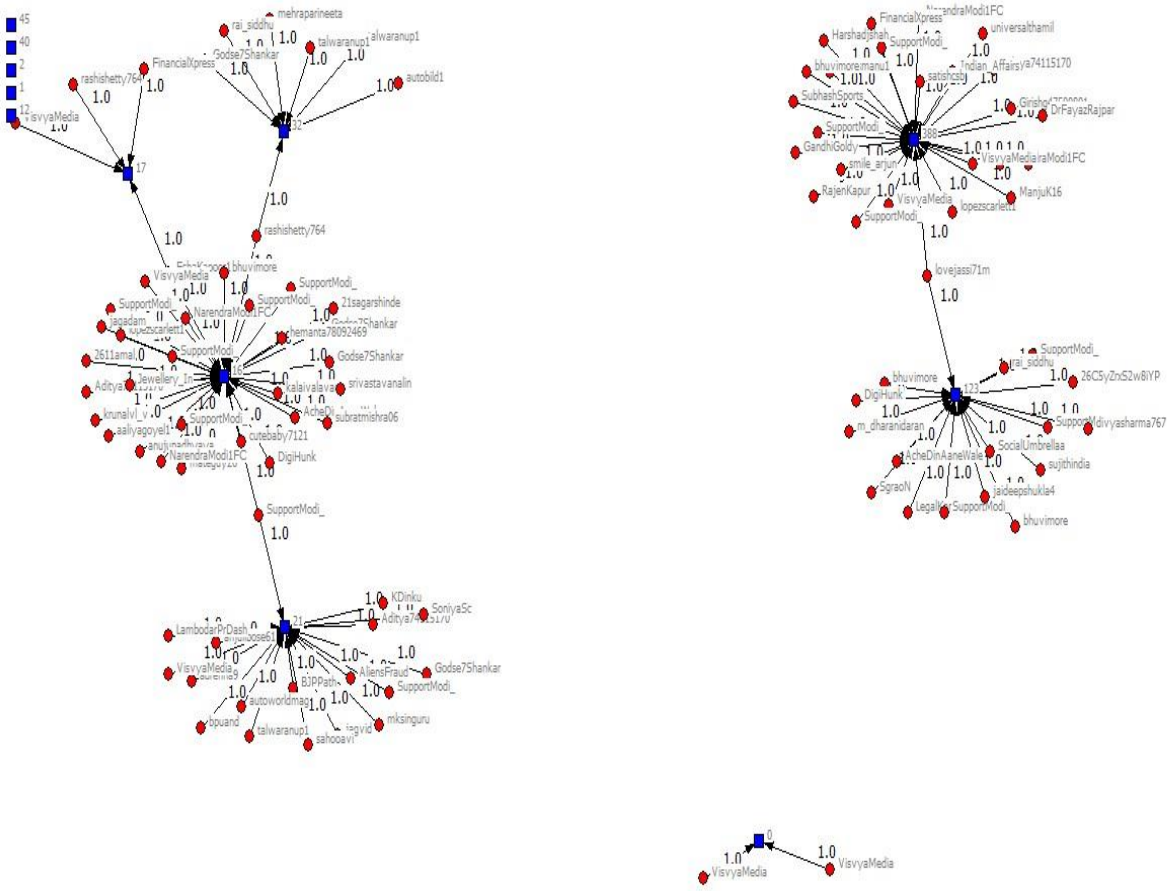

Fig 2: Network Diagram

BETWEENESS CENTRALITY:

In graph theory, **betweeness** centrality is a measure of centrality in a graph based on shortest paths.

CODE:

DISTANCE-WEIGHTED BETWEENNESS:

	DW	Betweenness
1	rashishetty764	5.379
2	Aditya74115170	12.303
3	Aditya74115170	41.807
4	Aditya74115170	5.695
5	sujithindia	3.020
6	mateguy26	12.400
7	NarendraModi1FC	12.512
8	SupportModi_	0
9	SupportModi_	0

10	universalthamil	0
11	Godse7Shankar	0
12	SupportModi_	0

EXPLANATION:

This is a simple stochastic algorithm to generate a graph. It is a discrete time step model and in each time step a single vertex is added. We start with a single vertex and no edges in the first time step. Then we add one vertex in each time step and the new vertex initiates some edges to old vertices. The probability that an old vertex is chosen is given by

$$P[i] \sim k \alpha i + a$$

sample_pa generates a directed graph by default, set directed to FALSE to generate an undirected graph. Note that even if an undirected graph is generated ki denotes the number of adjacent edges not initiated by the vertex itself and not the total (in- + out-) degree of the vertex, unless the out.pref argument is set to TRUE.

OUTPUT:

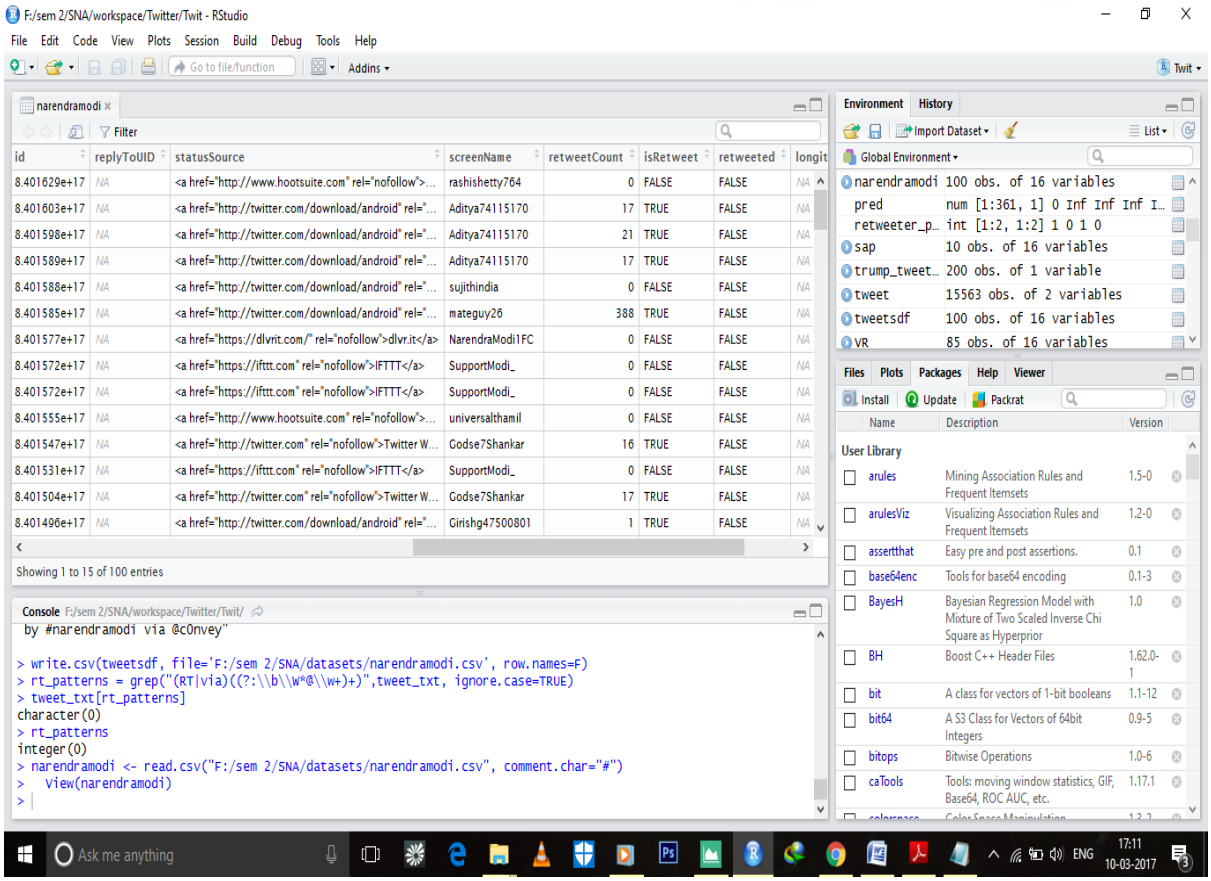

FIG 3: RSTUDIO

The above screen shot is a retrieval of retweets with the hashtag narendramodi

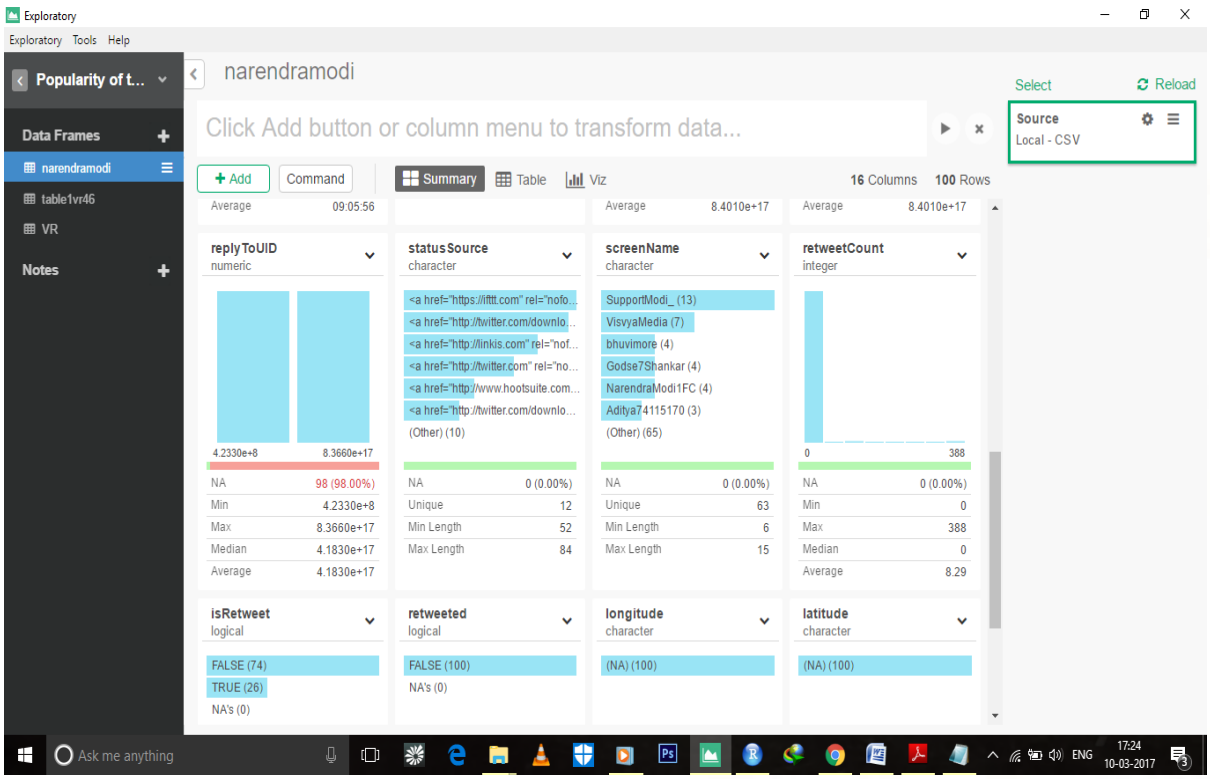

FIG 4: EXPLORATORY ANALYSIS

The above screen shot is a summary of retweets

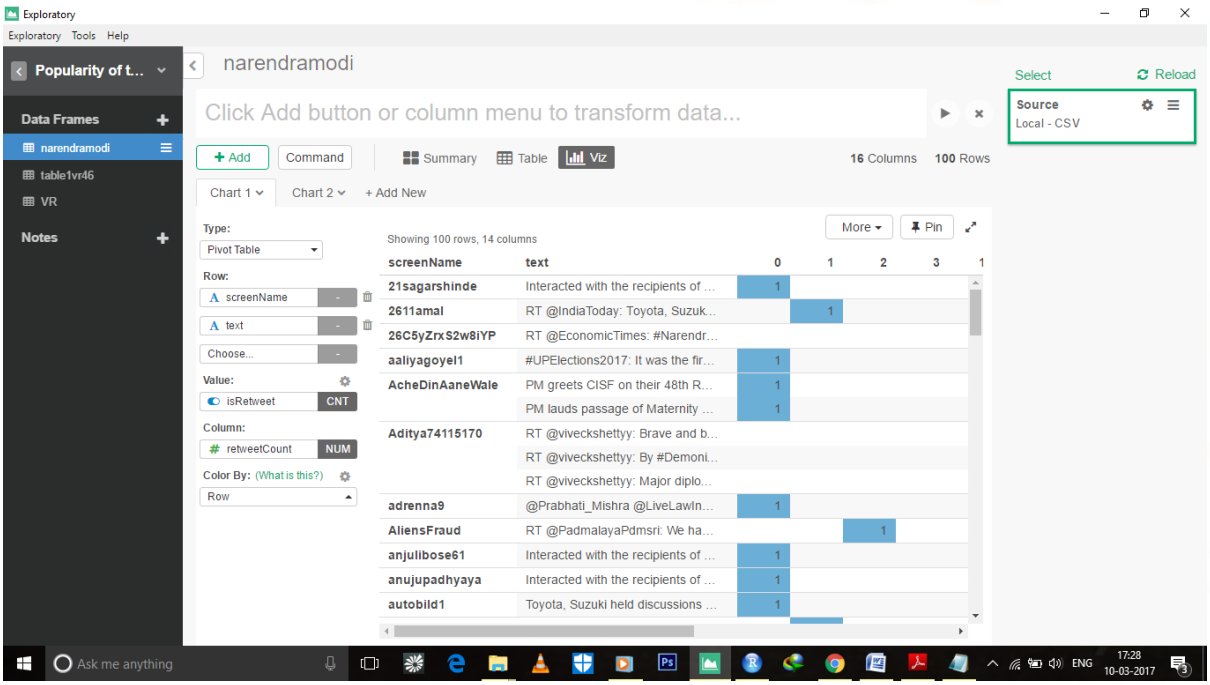

FIG 5: EXPLORATORY ANALYSIS

From here we infer the graphical data of the retweets

CONCLUSION

Our approach provides a theoretical framework for explaining temporal patterns ofinformation cascades.SEISMIC is both scalable and accurate. The model requires no feature engineering and scales linearly with the number of observed reshares of a given post. This provides a way to predict information

spread for millions of posts in an online real-time setting. SEISMIC brings extra flexibility to estimation and rediction tasks as it requires minimal knowledge about the information cascade as well as the underlying network structure. Thus the future enhancement could be come up with even more less relative error.

ACKNOWLEDGEMENTS:
This research has been supported by ucinet, exploratory, Rstudio.

RESULT:

The output is basically where the total number of retweets is analyzed, and then the Information cascade has been predicted. Below are some screenshots after implementing cascade.

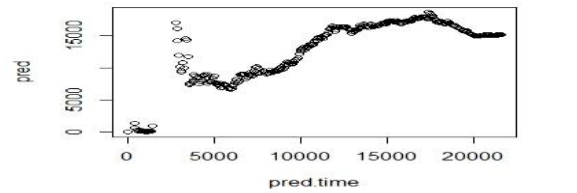

FIG 6: CASCADE CURVE

REFERENCES:

1) S. Gao, J. Ma, and Z. Chen. Modeling and predicting retweeting dynamics on microblogging platforms. In WSDM' 15, 2015.
2) A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. Biometrika, 58(1), 1971.
3) L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In WWW '11, 2011.
4) E. M. Rogers. Diffusion of innovations. Simon and Schuster, 2010.
5) B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be retweeted? large scale analytics on factors pacting retweet in twitter network. In SOCIALCOM '10, 2010.
6) https://twitter.com/mottbollomy/status/ 127001313513967616.
7) D. Agarwal, B.-C. Chen, and P. Elango. Spatio-temporal models for estimating click-through rate. In WWW '09, 2009.
8) E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In WSDM '11, 2011.
9) R. Bandari, S. Asur, and B. A. Huberman. The pulse of news in social media: Forecasting popularity. In ICWSM '12, pages 26–33, 2012.
10) A.-L. Barabasi. The origin of bursts and heavy tails in human dynamics. Nature, 435:207, 2005.
11) J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In WWW '14, 2014.
12) R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. PNAS, 105(41), 2008.
13) H. Daneshmand, M. Gomez-Rodriguez, L. Song, and B. Schölkopf. Estimating diffusion network structures: Recovery conditions, sample complexity & soft-thresholding algorithm. In ICML '14, 2014.
14) P. A. Dow, L. A. Adamic, and A. Friggeri. The anatomy of large facebook cascades. In ICWSM '13, 2013.
15) N. Du, L. Song, M. Yuan, and A. J. Smola. Learning networks of heterogeneous influence. In NIPS '12, 2012.
16) R. Durrett. Probability: theory and examples. Cambridge university press, 2010.
17) M. Gomez-Rodriguez, J. Leskovec, D. Balduzzi, and B. Schölkopf. Uncovering the structure and temporal dynamics of information propagation. Network Science, 2:26–65, 4 2014.
18) R Vidhya, G Vadivu, Research Document Search using Elastic Search, Indian Journal of Science and Technology 9 (37), 2016.
19) T. Y. J. Naga Malleswari , G. Vadivu, "Map reduce: A Technical Review", Indian Journal of Science and Technology, Vol 9(1), January 2016.
20) K.Sornalakshmi, G.Vadivu, "A Survey on Realtime Analytics Framework for Smart Grid Energy Management", International Journal of Innovative Research in Science, Engineering and Technology, March 2015.

FUTURE ENHANCEMENT:

- Enhancing the retweet count even more with less relative error.
- Improvizing the efficiency of mining.