Twitter Analytics: Architecture, Tools and Analysis

Rohan D.W Perera

CERDEC

Ft Monmouth, NJ 07703-5113

S. Anand, K. P. Subbalakshmi and R. Chandramouli

Department of ECE, Stevens Institute Of Technology

Hoboken, NJ 07030

Abstract—**We study the temporal behavior of messages arriving in a social network. We specifically study the tweets and re-tweets sent to president Barack Obama on Twitter. We characterize the inter-arrival times between the tweets, the number of re-tweets and the spatial coordinates (latitude, longitude) of the users who sent the tweets. The modeling of the arrival process of tweets in Twitter can be applied to predict co-ordinated user behavior in social networks. While there is sufficient literature on social networks that present large volumes of collected data, the modeling and characterization of the data have been rarely discussed. The available data are usually very expensive and not comprehensive. Here, we develop a software architecture that uses a Twitter application program interface (API) to collect the tweets sent to specific users. We then extract the user ids and the exact time-stamps of the tweets. We use the collected data to characterize the inter-arrival times between tweets and the number of re-tweets. Our studies indicate that the arrival process of new tweets to a user can be modeled as a Poisson Process while the number of re-tweets follow a geometric distribution. Our data collection architecture is operating system (OS) independent. The results obtained in this research can be applied to study correlations between patterns of user behavior and their locations.**

Index Terms:- Social networks, Twitter, API, modeling

I. INTRODUCTION

Online social networks (OSNs) have provided a means for users all over the globe to get in touch with each other to satisfy their professional interests (e.g., linkedIn [1]) as well as their personal interests (e.g., FaceBook [2], Twitter [3]). With a large volume of messages arriving from users in different parts of the world, it poses an interesting challenge to researchers to model the traffic on these OSNs. The first step to characterizing the traffic or population of messages in OSNs is to characterize the arrival process of the messages. Some key reasons describing the importance of developing stochastic models to characterize OSNs are as follows [4].

1) Social behaviour is complex: Stochastic models capture regularities in the behavior and allow for variability.
2) A well specified stochastic model allows an estimate of certain parameters associated with the observed outcomes.
3) The model can be used to obtain the distribution of several possible outcomes.
4) From the set of observed data, it is possible to obtain a model from which the observed data may have been generated.

Development of stochastic models to characterize the behavior of OSNs requires a large amount of data samples. As an example, in order to characterize the inter-arrival times between messages sent to a particular user, it is essential to obtain a large sample size of messages set to the user. Most of the current OSNs define a "friendship" relation between the users. A user of interest can send messages to or receive messages from only those users with whom he/she shares a relation of friendship. Alternatively, OSNs contain "groups" on certain topics and messages can only be exchanged between users that are registered in the group. These restrictions limit the amount of data that can be collected to study the messages that are sent to a user. Twitter is one OSN in which any user is permitted to send short messages called "tweets" to any other user or on any topic. The users exchanging messages neither need to maintain a relation of friendship nor need to possess membership to the same group. This enables collection of large amount of data in reasonably quick time, which, in turn, can be used to develop stochastic models to characterize the tweets. From the time Twitter originated, billions of tweets have been created with the current average being fifty-five million tweets a day [5]. This is a growth of 1,400% since 2009 [6]. While there are studies in the literature that model the arrival process of e-mails [7] and video traffic [8], there are not many studies that characterize the arrival process of messages in OSNs. One of the reasons is the lack availability of free data and the lack of an architecture that can enable data collection to study the statistics. Very recently, the arrival of tweets on earthquakes and their correlation to the earthquake map was studied [9]. However, the results cannot be directly generalized to other topics since the origin of tweets exclusively depends on the occurrence of a particular event- namely, earthquake.

In this paper, we demonstrate an architecture for twitter data collection and analysis for formulation and development of stochastic models of Twitter Data by considering a more general data set. Specifically, we focus on the inter arrival time between tweets and the frequency of users re-tweeting to the same user. This is because, tweets sent to a user depend on multiple events or many other factors like common interests, common profession, etc. Our data collection architecture is based on Twitter APIs developed in conjunction with PYTHON [10] and MySQL [11]. A Twitter network can be modeled as a directed graph as shown in Fig. 1 [12], in which vertices represent users. A directed edge from a vertex, v_1, to another vertex, v_2, indicates that the user v_1 is "following" user v_2 or user v_2 is "followed by" user v_1. Typically, the retweets of tweets sent by a user has a correlation with the number of users following the user of interest [13]. In our experiment we collect tweets from users that are following Barack Obama.
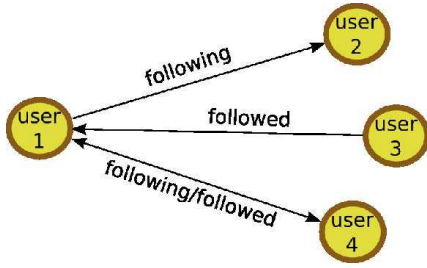
Fig. 1. Following/followed relationships in Twitter

We then extract the user ids and the exact time-stamps of the tweets. We use the collected data to characterize the inter-arrival times between tweets and the number of re-tweets to Barack Obama. Our studies indicate that the arrival process of new tweets to a user can be modeled as a Poisson Process while the number of re-tweets follow a geometric distribution.

The impact of these research findings is as follows. There are many models available in the queuing theory literature to study Markovian Queuing Systems with Poisson arrivals. Those results can be applied to study of behavior of user population in social networks like Twitter. Important parameters that can be determined are the mean growth and decay times of social networks i.e., mean time taken to grow a required size of population or the mean time for the population of a social network to decay to a specified value. This can be determined using the mean first passage time distributions. This, in turn, provides a means to determine the growth or loss of popularity of topics or user following in Twitter. This queuing analysis technique can also be applied to provide means of estimates of frequency of advertisements to generate the desired popularity for a topic or for a desired population of user following.

The rest of the paper is organized as follows. The architecture for the data collection process is described in Section II. In Section III, we outline the pre-processing and the analysis of the collected Twitter data. Conclusions are drawn in Section IV.

II. ARCHITECTURE AND TOOLS

Our experiment was run on a Dell Server Power Edge running Windows Server 2003, Python [10] and MySQL [11]. Fig. 2 describes our experimental set up.

Python enables a fast adaptation of the code. One can change the high-level layer of the application without changing the business rules that are coded within the modules. MySQL is an extensible, open storage database engine that offers multiple variations such as Berkeley DB, InnoDB, Heap and MyISAM. MySQL integrates seamlessly with a number of programming languages such as Python. Implementing the architecture using Python and MySQL provides the flexibility to adapt to other operating systems such as Linux, UNIX, etc. To write an application in python to interact with MySQL and Matlab we require an API (Application Programming Interface). The APIs used in our architecture are listed in Table I.
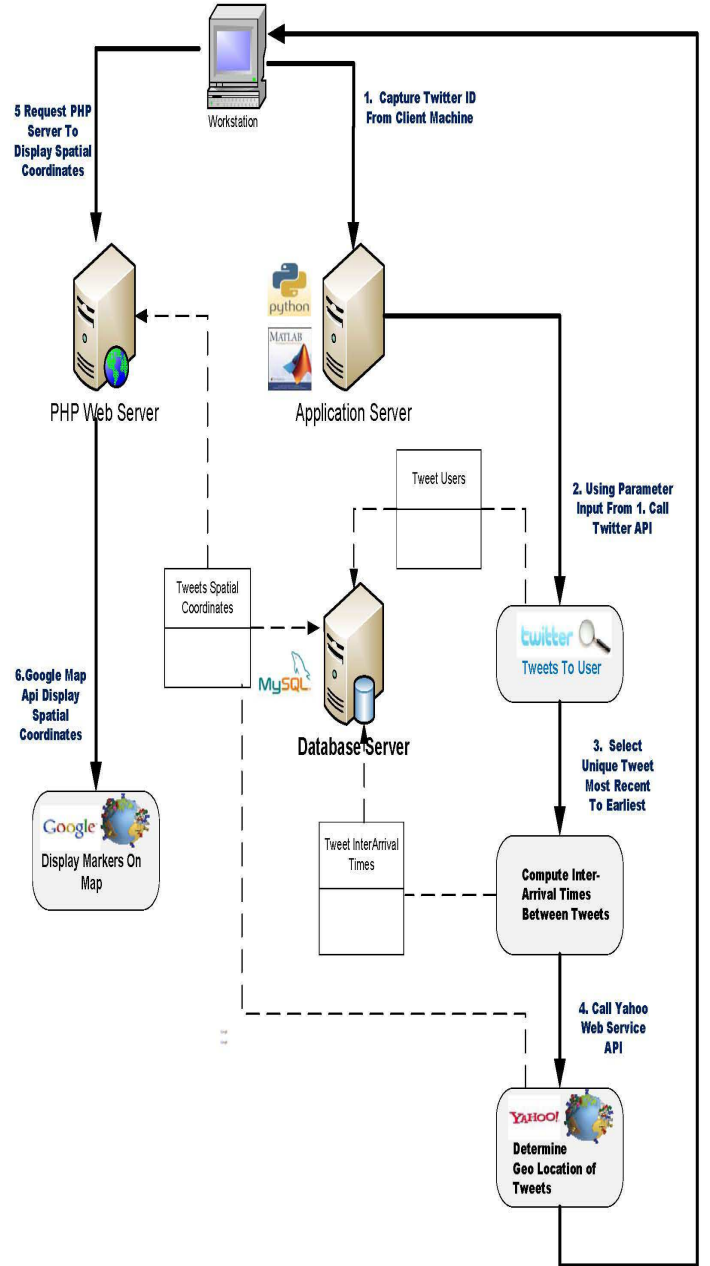

Fig. 2. Twitter data collection architecture

API Name	Download Location
MySQLdb	http://mysql-python.sourceforge.net/
mlabwrap	http://mlabwrap.sourceforge.net/

In order to collect the required data for our analysis, it is essential to use an API that is not only of low complexity of implementation, but also compatible with the python software. We implement two Twitter APIs namely REST and Twython to enable our data collection process. Representational State Transfer (REST)[14] is an API that allows a user to query and extract tweets based on customized parameters. As an example, it is possible to extract tweets by topic, a twitter user id, a specified date range, etc. To obtain spatial data about users who tweeted (e.g., full name, location, description, etc), we use a Python API called Twython. The Twython API calls are listed in Table II.

TABLE II
PYTHON/TWYTHON API CALLS

Method	Description
searchTwitter	searches on topic and retrieve tweets
showUser	Returns the User Profile

We use the Twython API, ShowUser, to determine location of tweets. However this API provides only the city and the state but does not provide the latitude and longitude coordinates. To obtain the spatial coordinates, we use a Restful API web service provided by Yahoo [15], to which, the address, city and state are provided as input parameters and the output is an XML file which contains the latitude and longitude. Fig.3 illustrates an example of using the Yahoo web service to determine the geographical location. In this example, The

Fig. 3.   Yahoo geo web service

latitude and longitude for 701 First Ave Sunnyvale California are 37.416397 (north of the equator) and 122.025055 (i.e., west of the Greenwich meridian), respectively. Additionally, a web server running PHP and Apache is used to plot the spatial

points. The MySQL database is shared between the PHP web resources and the python twitter collection program.

III.   DATA COLLECTION AND ANALYSIS

The parameters of interest are the Twitter ids, and the time when the tweet was sent. However, there are many attributes that can be obtained, e.g., Twitter user's name, language preference, date they joined twitter, description, etc. Fig .4 illustrates examples of tweets that were sent to Barack Obama that were collected using the architecture described in Section II. The data collection procedure is detailed in Section III-A

List Of Tweets

tweet text	tweet date	tweet user
@BarackObama Hey bro can I get an invite to tomorrow's bill signing? Pretty pleeeeeeease.	2010-03-29T18:01:20Z	eamartinez
@BarackObama - 3,503,874 followers, go to www.congress.org and tell your Congress members what you want for America.	2010-03-29T17:59:28Z	sarabialb
@BarackObama Schools can get away with financial aid fraud! Check out my blog on the topic http:hardtopic.blogspot.com/spref=tw	2010-03-29T17:58:55Z	Jagboi24
@BarackObama RT @Jagboi24: @KyraCNN Schools can get away with financial aid fraud! Check out my blog http:hardtopic.blogspot.com/spref=tw	2010-03-29T17:57:19Z	Jagboi24
@BarackObama will you follow me at twitter?	2010-03-29T17:51:42Z	martijnusername

Fig. 4.   Tweet messages to President Barack Obama

and the data analysis is described in Section III-B.

A. Data Collection

Fig. 5 describes our data collection process. The twitter data collection program captures two input parameters, i.e.,

1) Twitter id
2) time interval in minutes

The Twitter Id is used to extract tweets sent to the specified id. In our analysis, we collect the tweets sent to President Barack Obama. The time interval is the periodicity with which the twitter API is called. In our experiments, we consider a time interval of five minutes. Thus, every 5 minutes a request is made to Twitter using the Twitter APIs, to obtain tweets sent to Obama. The experiment was conducted for a period of 6 days from 14th to 20th of April, 2010, to obtain about one and a half million data points. The structure of the Twitter data that is collected. is illustrated in Fig 6. The key attributes which we extract are

1) TweetId
2) TimeStamp (Time of Tweet)
3) Tweet User (Twitter ID of user who sent Tweet)

B. Data Analysis

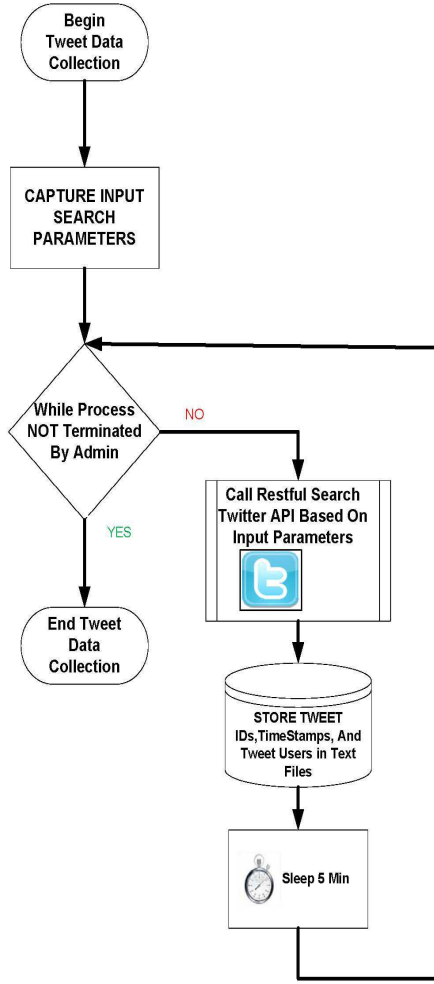Fig. 7 illustrates the overall Data Analysis procedure. The parameters extracted using the described procedure are

Fig. 5. Twitter data collection procedure

1) Computing Inter-arrival between tweets
2) Number of Tweets Sent to Obama by the same user
3) Determining Spatial Coordinates of Tweets

After the raw data was imported into the database, it was found that off the total number of one and half million tweets, only 5000 of these tweets were unique (determined using the tweet ids). The unique Tweets were ordered from the highest to the lowest tweet id to compute the inter-arrival time between tweets. From the unique tweets, frequency of tweets by each user and its spatial co-ordinates were determined and plotted on a map as shown in Fig 8 Despite having many tweets we were only able to determine 95 spatial coordinates. This is attributed to the fact due that the information provided by
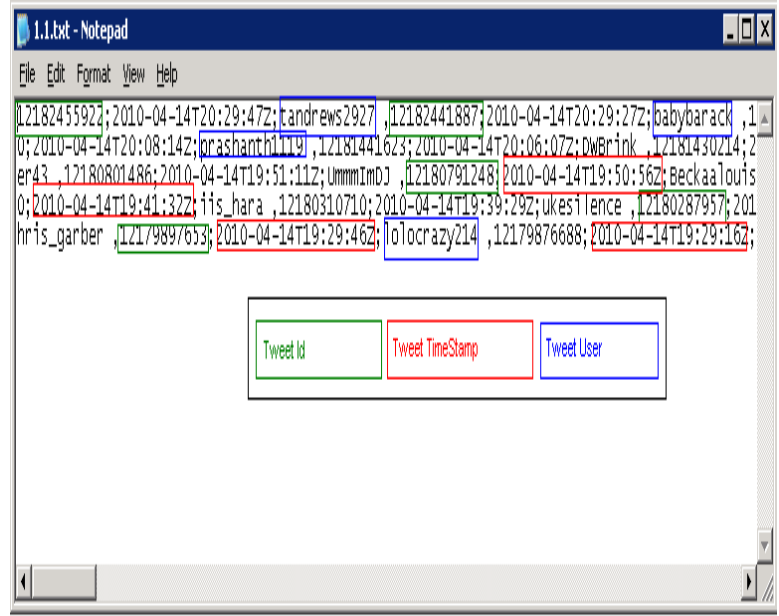

Fig. 6. Raw data structure of tweets

users to twitter may lack accuracy or may be incomplete, e.g., a user may provide only the state and not the city or may provide an incorrect city/state.

From the inter-arrival times computed using the collected data, we determine their normalized frequencies. This is done by computing the statistical frequencies and normalizing the maximum frequency to one. From the extracted inter-arrival times, we compute the mean inter-arrival time, $\overline{T}$, and define the arrival rate as $\lambda \stackrel{\triangle}{=} \frac{1}{\overline{T}}$. We fit an exponential probability density function (pdf) for the inter arrival times of tweets that are sent to Barack Obama. If T is the random variable that denotes the inter-arrival times with sample space, $\mathcal{S}_T = \{x | x \in (0, \infty)\}$, then we fit the pdf, $f(x)$, given by

$$f(x) = \lambda e^{-\lambda x} \qquad (1)$$

and the cumulative distribution function (cdf), $F(x)$ given by

$$F(x) = 1 - e^{-\lambda x}. \qquad (2)$$

To model the number of re-tweets to Barack Obama, we proceed as follows. Let N be the random variable that denotes the number of re-tweets. The sample space of N is $\mathcal{S}_N = \{n | n \in \{0, 1, 2, 3, \cdots, \}\}$. We fit a geometric probability mass function (pmf), given by

$$\Pr\{N = k\} = \rho^k (1 - \rho), \qquad (3)$$

where $0 < \rho < 1$. The expected value of N, $\overline{N}$, is then given by

$$\overline{N} = \frac{\rho}{1 - \rho}, \qquad (4)$$

i.e.,

$$\rho = \frac{\overline{N}}{\overline{N} + 1}. \qquad (5)$$

Fig. 7. Twitter data evaluation procedure

Fig. 8. Spatial coordinates of Twitter users tweeting to President, Barack Obama

tweets is 0.07, yielding an accuracy of 93%.

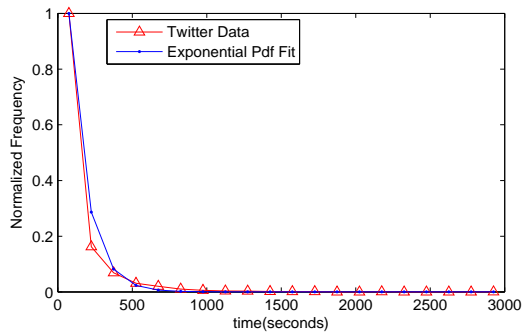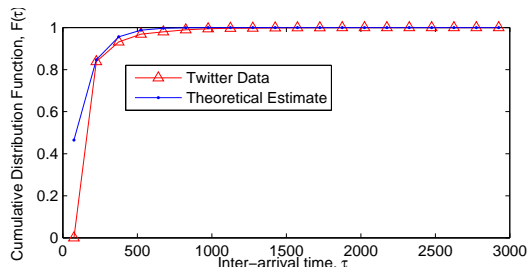| Graph	$Parameter$	RMSE
Fig.9	$\lambda = 0.0083$	0.0274
Fig.10	$\rho = 0.7542$	0.0674

IV. CONCLUSION

We developed an architecture that used PYTHON and MySQL in conjunction with Twitter APIs to obtain tweets sent to specific users. We used our architecture to obtain the inter-arrival times between the tweets, the number of re-tweets and the locations of the users sending the tweets. An exponential pdf was fitted to model the inter-arrival times of tweets with 97% accuracy and a geometric pmf was used to model the number of re-tweets with 93% accuracy. Our results can be used in conjunction with available results on queueing theory, to study the transient and steady-state behavior of social networks. The proposed architecture can be used to monitor OSNs for correlation between user behaviors and their locations. The application of the obtained results to study the growth of population in OSNs is under investigation.

REFERENCES

[1] [Online]. Available: http://www.linkedIn.com
[2] [Online]. Available: http://www.facebook.com
[3] [Online]. Available: http://www.twitter.com
[4] M. Tranmer. Statistical models for social networks. [Online]. Available: http://www.methods.manchester.ac.uk/events/2010-02-08/tranmer.pdf
[5] Tweet preservation. [Online]. Available: http://blog.twitter.com
[6] Twitter statistics: The full picture. [Online]. Available: http://thenextweb.com/socialmedia/2010/02/22/twitter-statistics-full-picture
[7] R. D. Malmgren, J. M. Hofman, L. A. N. Amaral, and W. D, J, "Characterizing individual communication patterns," *Proc., Intl. Conf. on Knowledge, Discovery and Data Mining (SIGKDD'09)*, Oct. 2009.
[8] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "YouTube traffic characterization: A view from the edge," *ACM SIGCOMM Conf. on Internet Measurement*, Oct. 2007.

Therefore, the parameter, ρ, can be obtained by computing the mean number of re-tweets, $\overline{N}$, from the collected data and applying (5).

Figs.9 and 10 present the fitted models for the inter-arrival times of tweets and the number of re-tweets, respectively. It can be observed that the expressions in (1), (2) and (3) model the inter-arrival times and the number of re-tweets, respectively. The accuracy of the models is measured in terms of the root-mean square error (RMSE). Table III lists the RMSEs of the models fitted in Figs. 9 and 10. It is observed that the RMSE is about 0.03 for the exponential fit for the inter-arrival times, thus yielding a 97% accuracy. The RMSE for the geometric pmf fit for the number of number of re-

(a) Probability density function (pdf)

(b) Cumulative distribution function (cdf)

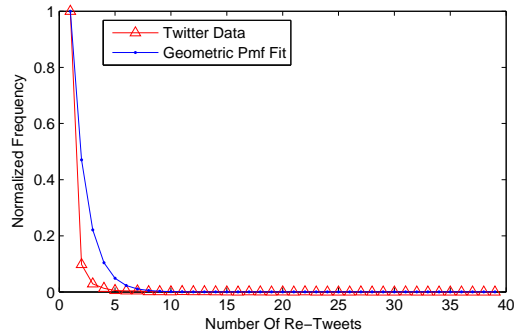Fig. 9. Exponential pdf/cdf fit to model the inter-arrival time between tweets

Fig. 10. Geometric pmf fit to model the number of re-tweets

[9] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthwauke shakes Twitter users: Real time event detection by social sensors," *ACM Intl. Conf. on World Wide Web*, pp. 851–860, April 2010.

[10] Python home page. [Online]. Available: http://python.org

[11] Mysql home page. [Online]. Available: http://www.mysql.com/

[12] H. D. Mitsuhiro Nakamura, "Cognitive-costed agent model of the microblogging network," in *Proceedings of the AESCS The Sixth International Workshop on Agent-based Approaches in Economic and Social Complex Systems*, Taipei, Taiwan, 2009.

[13] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in Twitter: The million follower falacy," *Proc. AAAI Intl. Conf. on Web Blogs and Social Media (ICWSM'2010)*, May 2010.

[14] Representational state transfer. [Online]. Available: http://en.wikipedia.org/wiki/Representational_State_Transfer

[15] Yahoo! maps web services - geocoding api. [Online]. Available: http://developer.yahoo.com/maps/rest/V1/geocode.html