Volume 6, Issue 7, July 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Exploring Twitter for Large Data Analysis

Sandeep Ranjan*, Dr. Sumesh Sood
IK Gujral Punjab Technical University, Kapurthala,
Punjab, India

Abstract— Social network websites like Twitter, Facebook and Linkedin are very popular among Internet users. Billions of users have their account on these websites and use this medium to express themselves, comment on products, services, news and other events and form enormous graphs. Analyzing and understanding these graphs is important to sense the user's views and responses about products and services offered by different companies to get real time feedback. This paper presents different methods used to fetch data from Twitter for analysis. We used OAuth and APIs for crawling Twitter.

Keywords— Social network, analysis, graphs, open authentication, visualization.

I. INTRODUCTION

The Internet is rapidly evolving and impacting our lives. By 2015 the number of Internet users had already touched 3 Billion [1]. Recent developments like wifi, 3G and 4G mobiles, fiber optic and satellite connectivity; decreasing Internet access cost and location independence are adding millions of Internet users every year. Users stay connected at homes, in offices and also during traveling. They post Exabytes of data daily using their mobiles and tablets. Much of the traffic has been due to social network websites.

Online social media has come into existence as a means of information dissemination where people create their own content, share their views, pictures and videos with other users [2]. The most common social network websites include Twitter, Facebook, LinkedIn, Flicker, YouTube and Instagram. There is a very large amount of data on these social networking websites generated by the users. These websites allow users to keep track of huge collections of posts posted by hundreds of their friends or other users [3]. Users can create their own social circles (or lists) for filtering content or applying privacy constraints. In addition to text posts, users can upload and share pictures and videos, repost or share the content of other users or comment on the content of others.

Social networking websites are finding a great place in business [4]. Billions of users today communicate through laptops, desktops and mobiles using email, blogs, discussion forums and social network websites. A huge, complex and lively network of relations evolving using these technologies is of great importance to individuals and corporations. People often read reviews posted by connected friends or other user's reviews about a product or service and rely on them before buying them. Most of the companies have their pages on Facebook or are being talked about on Twitter or other websites. It helps them track customer behavior on real time and also advertisement of their new services and products at lesser costs compared to the conventional methods. Products like movies, songs, sports events create a buzz among the general public weeks or months before the release and this can be sensed through social network responses about them.

II. TWITTER

Among social network website, Twitter is one of the most popular. We selected Twitter for crawling user responses as it has a large number of registered users and an enormous amount of their content and also that Twitter provides restricted access to data (tweets) to the general public via APIs. We studied various methods to crawl Twitter data and listed the outline of these methods.

Twitter users generate tweets of the order of 400 million per day [5]. Some of these tweets are accessible to the general public through APIs working with authentication requests. Authentication requests coupled APIs are based on the Open Authentication (OAuth) standard, which Twitter uses to provide access to information [6]. It provides a three way handshake securing user passwords from being shared with third parties.

III. NODEXL BASIC

Nodexl Basic is an open source, freeware template for Microsoft Excel for network analysis [7]. Nodexl provides a GUI menu to fetch data from social networks like Twitter, Facebook, YouTube and Flickr etc. Nodexl fetches information from social networks to be used as an input for generating network graphs of relationships. This information is fetched into multiple worksheets in the Excel template. Edges of graph in network relationships are stored as an edge-list containing all vertex pairs that are connected in the network. Other worksheets contain information about each vertex representing the nodes and the evolved clusters. The visualization feature of Nodexl allows users to render network graph representations and link attributes to visual properties like location, size, shape, color and transparency.

Nodexl allows calculating graph parameters like vertex in degree, vertex out degree, betweenness, closeness centrality, eigenvector centrality, edge reciprocation, page rank and clustering coefficient. Nodexl also allows to change the color, opacity, edge width, visibility, vertex shape and vertex size in the graph. When any node of the graph is clicked, its corresponding spreadsheet values get highlighted helping the user to get more information and understanding about the node and its neighbours. Following table lists different overall metrics that get inserted into the Overall Metrics worksheet:

Table I Nodexl Overall Metrics Worksheet

Graph type	**Directed or undirected**
Vertices	The Number of vertices in the graph
Unique Edges	The number of edges that do not have duplicates.
Edges with Duplicates	The number of edges that do not have duplicates.
Total Edges	The number of edges in the graph
Self Loops	The number of edges that connect a vertex to itself.
Reciprocated Vertex Pair Ratio	The number of vertex pairs that have edges in both directions divided by the number of vertex pairs that are connected by any edge.
Reciprocated Edge Ratio	The number of edges that are reciprocated divided by the total number of edges.
Connected Components	The number of connected components in the graph
Single-Vertex Connected Components	The number of connected components that have only one vertex.
Maximum Vertices in a Connected Component	The number of vertices in the connected component that has the most vertices.
Maximum Edges in a Connected Component	The number of edges in the connected component that has the most edges.
Maximum Geodesic Distance (Diameter)	The maximum geodesic distance among all vertex pairs, where geodesic distance is the distance between two vertices along the shortest path between them.
Average Geodesic Distance	The average geodesic distance among all vertex pairs, where geodesic distance is the distance between two vertices along the shortest path between them.
Graph Density	This is a ratio that compares the number of edges in the graph with the maximum number of edges the graph would have if all the vertices were connected to each other.
Modularity	When the graph has groups, this is a measure of the "quality" of the grouping. Graphs with high modularity have dense connections among the vertices within the same group, but sparse connections between vertices in different groups.

We used Nodexl to fetch tweets for Freecharge, a popular Indian website/ smartphone app for online set top box and mobile recharge and got the following graph metrics:

Table 2 Freecharge Graph Metric Values

Graph Metric	**Value**
Graph Type	Directed
Vertices	562
Unique Edges	693
Edges With Duplicates	420

Total Edges	1113
Self-Loops	193
Reciprocated Vertex Pair Ratio	0.020864
Reciprocated Edge Ratio	0.040876
Connected Components	156
Single-Vertex Connected Components	131
Maximum Vertices in a Connected Component	370
Maximum Edges in a Connected Component	902
Maximum Geodesic Distance (Diameter)	6
Average Geodesic Distance	2.661366
Graph Density	0.002173
Modularity	Not Applicable
NodeXL Version	1.0.1.350

In addition to all these metrics, Nodexl can fetch tweets about a particular hashtag, twitter IDs as vertices, relationship (replies to, mentions or tweet), relationship date (UTC), URLs in tweet, domains in tweet, Twitter page for tweet and tweet ID. Nodexl has very powerful virtualization features to generate graphs, choose shapes for vertices, color option for edges and also user profile picture can be fetched for vertices.

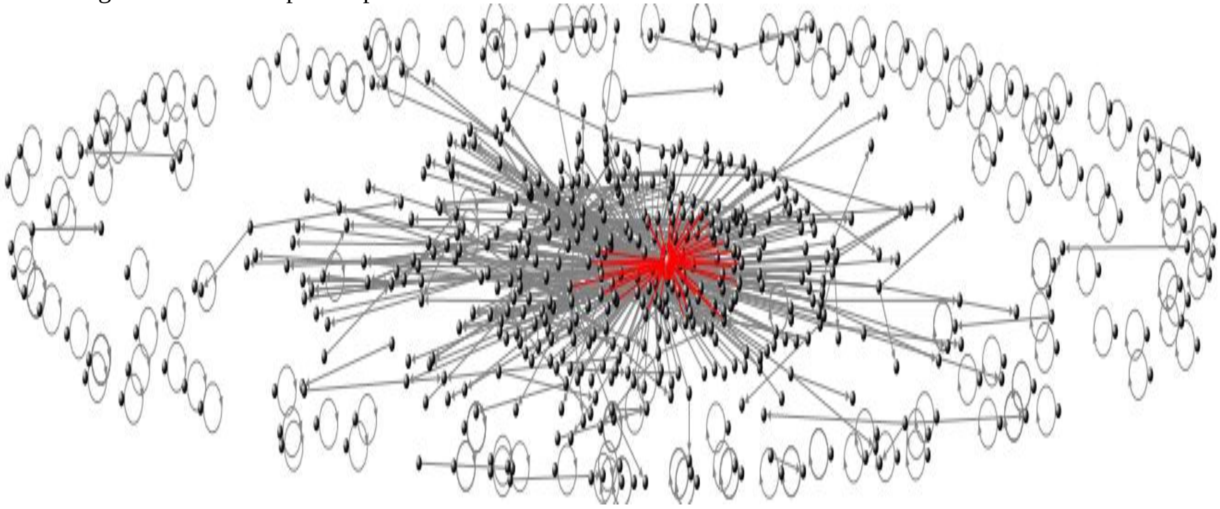


Fig 1. Nodexl Graph generated for #freecharge tweets

Fig 1 shows the graph for #freecharge related tweets from Twitter. This graph can be subgrouped into subgraph for finer analysis of smaller communities.

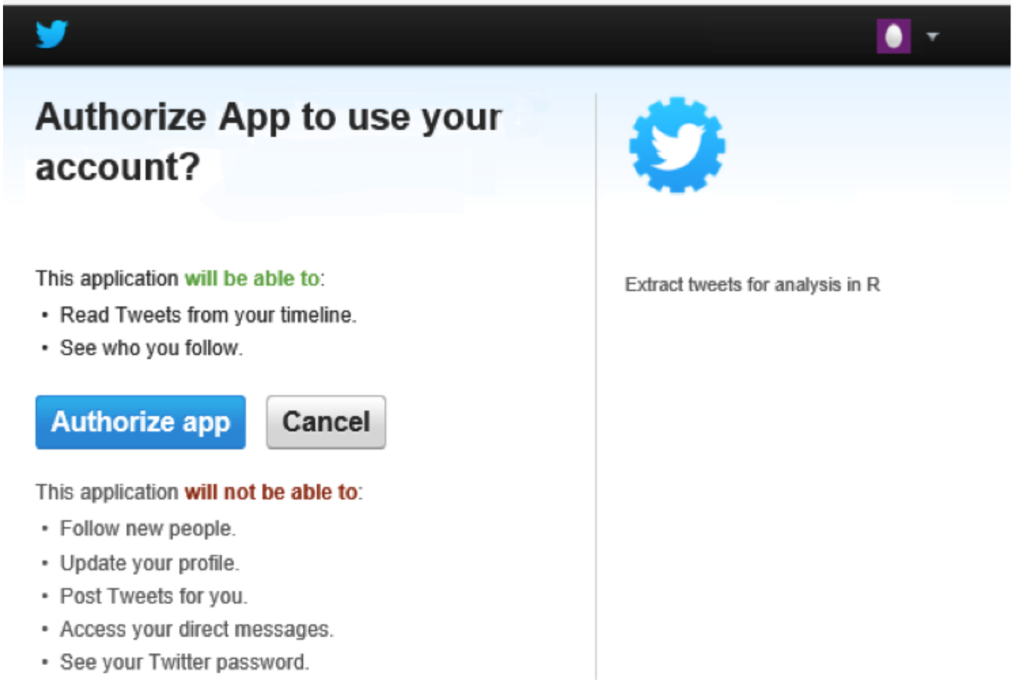
IV. R PROGRAMMING & SOFTWARE SUITE

R is an open source programming language cum software suite for data computation, manipulation and visualization by R Foundation for Statistical Computing [8]. The R environment consists of numerous classical and modern statistical techniques implemented in it. Some of these are already embedded into the R environment while others can be installed as packages from CRAN websites or any other compatible source. Some of the packages have dependencies on other packages so one needs to install a set of packages for a particular functionality.

R is a case sensitive expression language with a simple syntax. Elementary commands are either expressions or assignments. If an expression is executed as a command, it is evaluated and printed and the resultant value is not saved. An assignment also performs the same for an expression, saves the value to a variable but does not print the result automatically. R has both a command line console and a rich graphical environment. Commands are executed onto the console and the graphical output if required can be viewed in the graphical interface. For fetching tweets from Twitter, we need to include “twitteR”, and “ROAuth” packages by using the following commands:

```
install.packages("twitteR")
install.packages("ROAuth")
library("twitteR")
library("ROAuth")
```

You may get a prompt to get “cacert.pem” file. This file can be downloaded from the specified URL and stor stored in the working directory. Now you need to enter the consumerKey and consumerSecret which is unique for each user connecting to Twitter through the API. On the completion of the handshake, it will direct the user to a hyperlink in the console window. Navigate to the link to authorize the app by clicking on “Authorize App”.



Authorize App to use your account?

This application will be able to:

- Read Tweets from your timeline.
- See who you follow.

This application will not be able to:

- Follow new people.
- Update your profile.
- Post Tweets for you.
- Access your direct messages.
- See your Twitter password.

Extract tweets for analysis in R

Fig 2: Twitter APP authorization

We can create an R script using the filterStream function of the streamR package. This function takes the following parameters [9] [10]:

Table III R filestream & streamR PARAMETERS

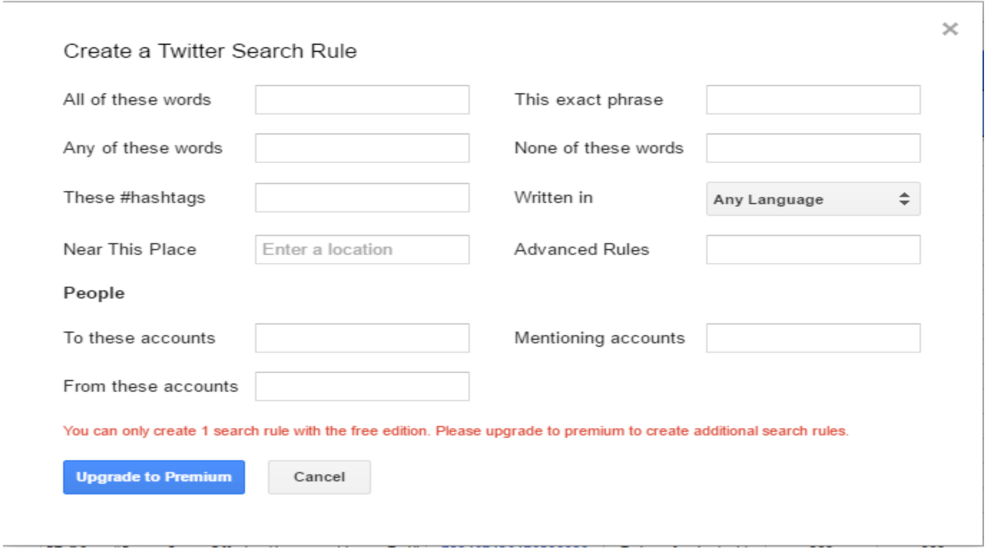
Parameter	Description
file.name	Filename where tweets will be saved
track	A string containing keywords to be tracked
follow	Twitter user IDs if one wants to only track tweets of specific twitter users.
locations	A vector containing latitude and longitude pairs specifying a set of bounding boxes to filter incoming tweets.
language	A list of BCP 47 language identifiers.
timeout	Time in seconds, the max length of time to connect to the stream
tweets	Number of tweets to be collected
oauth	Object where one specifies oauth setup (my_oauth).
verbose	Can be TRUE or FALSE, generates output to the R console with information about the capturing process.

All the tweets within the Twitter rate limit will be saved in a file which can be used in R or any other tool to be analyzed statistically and graphically using R or any other software tool.

V. GOOOGLE SPREADSHEETS AND BLOCKSPRING

Google Spreadsheets is a part of the very popular free and web based office suite by Google [11]. Google office suite is available both as a web app and offline app for Chrome and is tightly integrated with Google Drive. The files created are saved on Google Drive and can also be remotely accessed. The Google Sheets API v3 supports developing client applications to read and modify worksheets and data in Google Sheets [12].

One needs to login into chrome with Google ID, go to Google spreadsheets and create a new spreadsheet. Authorize your Google account to connect to Twitter. Create a search rule and specify the search criteria. The basic version of spreadsheet allows creation and execution of only one rule at a time. User can select from multiple search criteria available for search rule.



Create a Twitter Search Rule

All of these words		This exact phrase	
Any of these words		None of these words	
These #hashtags		Written in	Any Language
Near This Place	Enter a location	Advanced Rules	

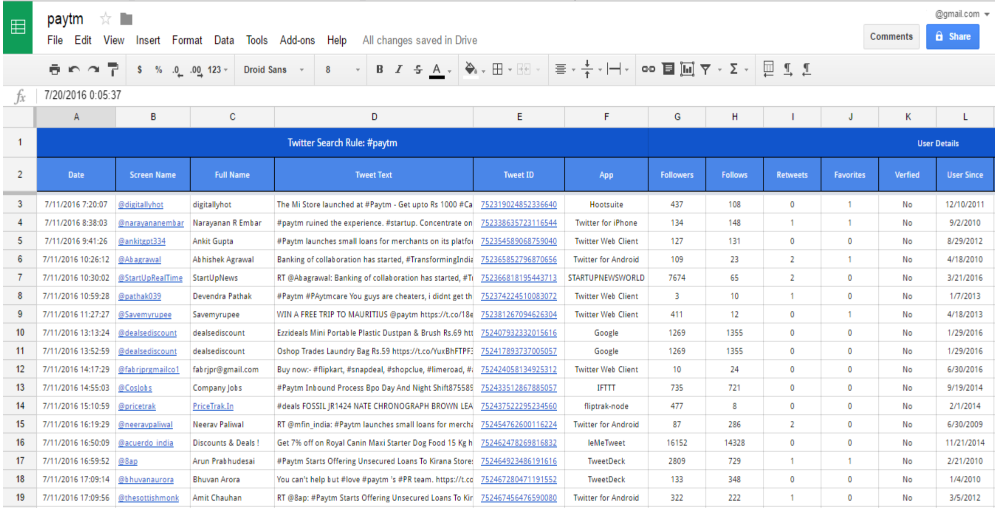
People

To these accounts		Mentioning accounts	
From these accounts			

You can only create 1 search rule with the free edition. Please upgrade to premium to create additional search rules.

Fig 3 Google spreadsheet search rule creation

One of the best features of this tool is that it keeps on fetching tweets within the Twitter rate limit in the background even if the PC is shut down. If the user has existing spreadsheets in the linked Google Drive, Google Spreadsheet's import data feature can import data from these spreadsheets for analysis.



paytm
File Edit View Insert Format Data Tools Add-ons Help
All changes saved in Drive
@gmail.com
Comments
Share
Droid Sans
8

fx 7/20/2016 0:05:37

	A	B	C	D	E	F	G	H	I	J	K	L
1	Twitter Search Rule: #paytm						User Details					
2	Date	Screen Name	Full Name	Tweet Text	Tweet ID	App	Followers	Follows	Retweets	Favorites	Verfied	User Since
3	7/11/2016 7:20:07	@digitallyhot	digitallyhot	The Mi Store launched at #Paytm - Get upto Rs 1000 #Ca	752319024852336640	Hootsuite	437	108	0	1	No	12/10/2011
4	7/11/2016 8:38:03	@narayananembar	Narayanan R Embar	#paytm ruined the experience. #startup. Concentrate on	752338635723116544	Twitter for iPhone	134	148	1	1	No	9/2/2010
5	7/11/2016 9:41:26	@ankitgpt334	Ankit Gupta	#Paytm launches small loans for merchants on its platfo	752354589068759040	Twitter Web Client	127	131	0	0	No	8/29/2012
6	7/11/2016 10:26:12	@Abagrawal	Abhishek Agrawal	Banking of collaboration has started, #TransformingIndia	752365852796870656	Twitter for Android	109	23	2	1	No	4/18/2010
7	7/11/2016 10:30:02	@StartUpRealTime	StartUpNews	RT @Abagrawal: Banking of collaboration has started, #Tı	752366818195443713	STARTUPNEWSWORLD	7674	65	2	0	No	3/21/2016
8	7/11/2016 10:59:28	@pathak039	Devendra Pathak	#Paytm #PAytmcare You guys are cheaters, i didnt get th	752374224510083072	Twitter Web Client	3	10	1	0	No	1/7/2013
9	7/11/2016 11:27:27	@Savemyrupee	Savemyrupee	WIN A FREE TRIP TO MAURITIUS @paytm https://t.co/18e	752381267094626304	Twitter Web Client	411	12	0	1	No	4/18/2013
10	7/11/2016 13:13:24	@dealsediscount	dealsediscount	Ezzideals Mini Portable Plastic Dustpan & Brush Rs.69 htt	752407932332015616	Google	1269	1355	0	0	No	1/29/2016
11	7/11/2016 13:52:59	@dealsediscount	dealsediscount	Oshop Trades Laundry Bag Rs.59 https://t.co/YuxBhFTPF:	752417893737005057	Google	1269	1355	0	0	No	1/29/2016
12	7/11/2016 14:17:29	@fabrjprgmailco1	fabrjpr@gmail.com	Buy now:- #flipkart, #snapdeal, #shopclue, #limeroad, #:	752424058134925312	Twitter Web Client	10	24	0	0	No	6/30/2016
13	7/11/2016 14:55:03	@CosJobs	Company Jobs	#Paytm Inbound Process Bpo Day And Night Shift875589	752433512867885057	IFTTT	735	721	0	0	No	9/19/2014
14	7/11/2016 15:10:59	@pricetrak	PriceTrak.In	#deals FOSSIL JR1424 NATE CHRONOGRAPH BROWN LEA	752437522295234560	fliptrak-node	477	8	0	0	No	2/1/2014
15	7/11/2016 16:19:29	@neeravpaliwal	Neerav Paliwal	RT @mfin_india: #Paytm launches small loans for mercha	752454762600116224	Twitter for Android	87	286	2	0	No	6/30/2009
16	7/11/2016 16:50:09	@acuerdo_india	Discounts & Deals !	Get 7% off on Royal Canin Maxi Starter Dog Food 15 Kg h	752462478269816832	IeMeTweet	16152	14328	0	0	No	11/21/2014
17	7/11/2016 16:59:52	@8ap	Arun Prabhudesai	#Paytm Starts Offering Unsecured Loans To Kirana Store:	752464923486191616	TweetDeck	2809	729	1	1	No	2/21/2010
18	7/11/2016 17:09:14	@bhuvanaurora	Bhuvan Arora	You can't help but #love #paytm 's #PR team. https://t.co	752467280471191552	TweetDeck	133	348	0	0	No	1/4/2010
19	7/11/2016 17:09:56	@thesottishmonk	Amit Chauhan	RT @8ap: #Paytm Starts Offering Unsecured Loans To Kir	752467456476590080	Twitter for Android	322	222	1	0	No	3/5/2012

Fig 4 Google spreadsheet for #paytm tweets

The data in Google Spreadsheet can be accessed via Google Credentials on any system and can be used for further analysis through addons like Blockspring or exported to desktop data analysis applications

Blockspring is a Google Sheets template used to connect the spreadsheets to the Internet [13]. It has more than 1000 functions that can be run from Google Sheets. Blockspring can run algorithms, automate tweets and emails, visualize data and call APIs. Blockspring is available as an add-on for chrome. When data is fetched into a spreadsheet, a GUI can be used to activate Blockspring and choose one of the numerous functions for analysis. Below are graphical outputs of two of the functions for data analysis to compute "number of followers" and "count of app" related to the search criteria.

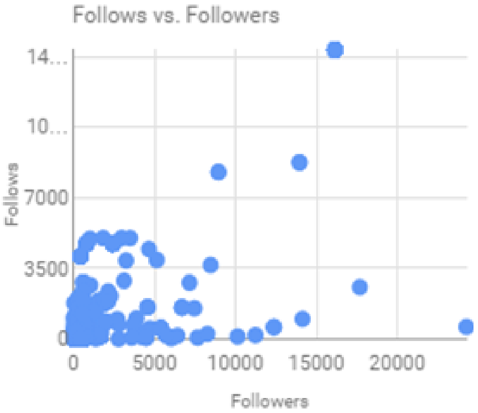
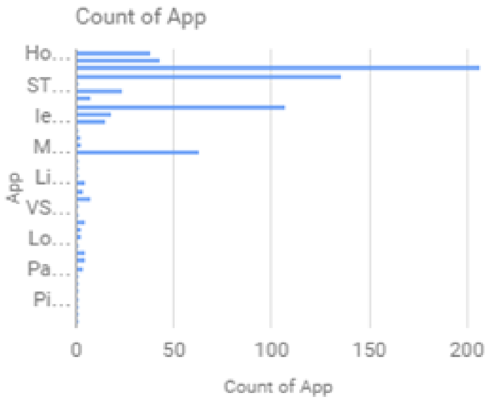


Fig 5. Blockspring data visualizations for Followers



"Twitter Web Client" has the highest value for "App" (207).

Fig 6. Blockspring data visualizations for Count of App

VI. CONCLUSION

There are many methods and tools available on the Internet to fetch tweets from Twitter within the prescribed rate limit. One can use these methods to fetch tweets about a particular hashtag (#) into a spreadsheet for analysis. The graphs generated through this data give in depth knowledge about the Twitter user, date and time of tweet and his connected users who have something in common with the user. This community can be selected as a target for futures predictions, reviews, feedbacks etc to generate valuable information for research and business promotion.

REFERENCES

[1] Internet Society Global Internet Report 2015

[2] A. Mislove, M. Marcon, K.P. Gumandi, P. Druschel, and B. Bhattacharjee, "Measurement and Analysis of Online Social Networks", *Proc. 7th ACM SIGCOMM conference on Internet measurement*, pp. 29-42, 2007

[3] J. McAuley, and J. Leskovec, "Learning to Discover Social Circles in Ego Networks", *Advances in neural information processing systems*, pp. 539-547, 2012

[4] L. Harris, and A. Rae, "Social networks: the future of marketing for small business", *Journal of business strategy*, 30(5), pp. 24-31, 2009

[5] http://articles.washingtonpost.com/2013-03-21/business/37889387_1_tweets-jack-dorsey-twitter

[6] S. Kumar, F. Morstatter, and H. Liu. "*Twitter data analytics", New York: Springer*, 2014

[7] D. Hansen, B. Shneiderman, and M. Smith, "Analyzing social media networks: Learning by doing with NodeXL." *Computing* 28.4: 1-47, 2009

[8] W. N. Venables, and D. M. Smith, "An introduction to R. version 1.9.1", 2004

[9] http://bogdanrau.com/blog/collecting-tweets-using-r-and-the-twitter-streaming-api/

[10] E. Borra, and B. Rieder, "Programmed method: developing a toolset for capturing and analyzing tweets." *Aslib Journal of Information Management* 66.3 : 262-278, 2014

[11] https://www.google.co.in/sheets/about/

[12] https://developers.google.com/google-apps/spreadsheets/

[13] https://www.blockspring.com/blog/blockspring-for-google-sheets