Analysis and Visualization of Twitter Data using k-means Clustering

Neha Garg
Department of Computer Science and Engineering
Thapar University, Patiala, India
nehagarg691@gmail.com

Rinkle Rani
Department of Computer Science and Engineering
Thapar University, Patiala, India
raggarwal@thapar.edu

***Abstract*— A social structure of individuals related directly or indirectly on the basis of some common factor like similar likings etc. is a social network. In order to understand the behavior and structure of a social network we need to study the network and this study is called social network analysis. There has been a rapid increment in the research and study of data mining community and social network analysis. There are various social networking sites available on internet like LinkedIn, Facebook, Instagram, Twitter, Google and many more. Interactions over such sites produces huge amount of data because billions of active users maintain their accounts. Hence, it is a tedious task to analyze the complex data. It is of great importance for academic and business to analyze such online social communities and predicting their behavior. In this paper, we are concentrating on Twitter data. R language is used for acquisition, preprocessing, analyzing and visualization of the twitter data. Twitter data is extracted, preprocessed and then clustered based on the geotagged information.**

Keywords—Twitter, K-means, Clustering, Analysis, Visualization.

I. INTRODUCTION

A. Social Network

In the past few years, there was an immense growth of the social networks. Such networks provide the platform to users to express, share or discuss their ideas, opinions with their friends in social graph and also communicate with them. Social networks also provide the platform to the internet users to interact with both the technology and other people. Social Networking is one of the primary reasons that many people have become avid Internet users [1]. Today, the significant quality of multimedia content is produce and consume by the user. A new Internet era is formed where multimedia content is shared via Social Networking Sites.

Social networks are oriented for both the professional and non professional users. The professional social networking sites allow establishing the professional connection between their users and business collaboration. These sites are LinkedIn, Viadeo and Xing. The non professional social networking sites are Facebook, Twitter, Instagram, Google+ etc. The most visited networking site in the world is Facebook. Facebook has more than 1000 million active users. These users are increasing about 85% annually since 2008 [6].

B. Microblogging

Social media provides a large amount of data. Automation is required for extracting the knowledge from large volume of data. It is a challenge for both the developer and algorithm to compute the data quickly. Recently, microblogging has become a popular trend which is responsible for a large amount of information dissemination. Microblogging websites are services which allow the user to post their ideas, opinions in defined number of words. It also allows the user to exchange content, images, video links and others. However, microblogging sites have become popular due to Twitter. Other microblogging sites are also available with the same functionality such as Tumblr, Pinterest, Flattr, Plurk etc. Tumblr provides the same functionality like twitter; however it focuses on the design and style. It is best for its simplicity in content management and posting. Each Tumblr blog is known as tumblelog. The important and powerful feature of social networking and microblogging platform is that user can post a message only to a selected friends or group of friends and not necessarily for all friends.

C. Twitter

Twitter is a social networking service which allows the user to send and read the short message of 140 characters called “tweets”. There are two types of users for twitter account. One is registered users who can only read the tweets and another are registered users who can read and post the tweets. It is a public platform for all the people of different age categories all over the world. Data generated by twitter is heterogeneous in terms of content because user can post a text, image, video and audio in any format. Data is also big in size because hundred of thousand of tweets per day is generated [2].

In the late 2009, twitter added a new feature which allows each tweet to be geo-tagged which is associated with

longitude and latitude of specific location [7]. Fig 1. shows some tweets extracted from twitter.

In this paper, tweets are extracted, refined, analyzed and visualized in a geospatial representation. The main goal of our research is to visualize the user's tweet in a particular area and visualize the clustered tweets according to the location information. The various packages and libraries are provided by R for extracting and processing the data and also for the visualization of clustered data.

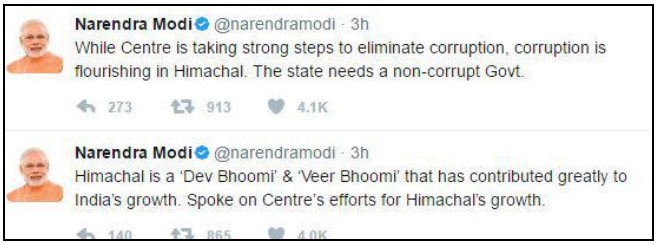

Fig 1. Example of tweets from twitter

The paper is structured as follows: section II discusses about OAuth, k-means Clustering Algorithm and Packages of R. Section III describes implementation of the proposed model. Section IV presents the results of clustering and visualization and section V conclude our work.

II. PREMILINARIES AND BACKGROUND

A. *OAuth*

API stands for Application Programming Interface which defines the set of methods to communicate between various software components. The twitter data such as followers, tweets, retweets, latitude, longitude etc are extracted by using different twitter APIs. The *REST APIs* helps the developer or programmer to read, write and access the twitter data. The *Streaming APIs* provides the uninterrupted access to the twitter data for a search query. It runs continuously until the internet connection interrupts or they kill. For accessing the twitter data, twitter provides different streaming endpoints [8].

Public Streams: Twitter data which is publically available can be accessed by using these streams. These streams are suitable for particular user, particular research interest or mining the data.

User Streams: The data which is related to specific user can be accessed by using these streams. By using these streams we can access followers, tweets or retweets.

Site Streams: These streams are used for those servers which are linked to twitter on favor of many users.

Twitter uses the OAuth to allow the secure authorization access to its API. To access server resources on the behalf of client or end-users OAuth provides a method for client. Twitter platform uses OAuth 2.0 for authentication and authorization in mobile applications, web applications and desktop applications. The OAuth 2.0 authorization framework enables a third-party application to obtain limited access to an HTTP service [9].

A process is provided by OAuth for end-users to enable third party-access to their server resources without giving their credentials (Consumer Key, Consumer Secret, Access Token and Access Token Secret).

Data format: by default twitter data is extracted and presented in JSON format [4]. But for our analysis, we convert the JSON format into data frame format because handling of data is easier in this format. Fig 2. explains the structure of tweet in data frame format. The information contains tweet related data like tweet, latitude, longitude, date of creation, etc. This information also carries information of tweet creator, retweet.

	text	favorited	favoriteCount	replyToSN	created	trunc
1	RT @Madan_Chikna: People from Mumbai expressing...	FALSE	0	NA	2017-04-26 08:33:02	FALSE
2	RT @Abhina_Prakash: That Delhi chose Dengue &am...	FALSE	0	NA	2017-04-26 08:33:02	FALSE
3	RT @smritiirani: Congratulations to karyakartas &a...	FALSE	0	NA	2017-04-26 08:33:02	FALSE
4	RT @Republicofindia: #Sensex Hits Record High 301...	FALSE	0	NA	2017-04-26 08:33:01	FALSE
5	RT @vivekshettym: Once again !! #MCDresults #MCDE...	FALSE	0	NA	2017-04-26 08:32:57	FALSE
6	RT @MODIfyingBHARAT: Thanks for the trust shown ...	FALSE	0	NA	2017-04-26 08:32:57	FALSE
7	On a day when his party has won a landslide victory ...	FALSE	0	NA	2017-04-26 08:32:55	TRUE
8	RT @himantabiswa: Trends in #MCDresults yet again...	FALSE	0	NA	2017-04-26 08:32:54	FALSE
9	RT @Republicofindia: BJP winning in Jama masjid , is ...	FALSE	0	NA	2017-04-26 08:32:54	FALSE

Fig 2. Tweet in data frame format

B. *K-means Clustering Algorithm*

Clustering is defined as collection of objects in one group which are similar between them and dissimilar to other group. It is extensively used in various fields such as text mining, machine learning, image analysis, image processing, web cluster engines, bioinformatics, weather report analysis, etc [5]. There are various methods of the clustering such as model based method, density based method, hierarchical method, grid based method, partitioned method etc.

K-means [3] clustering algorithm is an unsupervised learning algorithm which is used for the unlabelled data i.e. data are not labeled into any group or cluster. The objective of this algorithm is to find the clusters in the data with the already given number of clusters. The number of clusters and the dataset are the inputs of the algorithm. The dataset is the collection of data for each data point. The number of

clusters can either be randomly selected or randomly generated from the dataset. The algorithm works as: firstly initialize the number of clusters and the set the centroid of the clusters. Each data point is assigned to a cluster based on the smallest distance between the centroid and the data point. The centroids are updated or recomputed by taking the average (mean) of the data point assigned to the cluster. The process continues until the stopping criterion is met. The stopping criterion is any one of them which are data points is not changing the clusters, the sum of distances is minimized or the number of iterations are reached maximum.

C. *Packages in R*

In our research work for analysis and clustering of twitter data, we are using R language. R is a popular language used for retrieve, clean, analyze and visualize the data. The various packages are available in R to analyze and visualize the twitter data. We can download the different packages in R using CRAN mirror. R framework [10] is needed with R studio [11]. Table I. describes the various packages in R which is used for our analysis and visualization.

Table I. Various Packages in R for Analysis and Visualization

S.No.	R Packages	Description
1	twitterR	Provides an interface to the Twitter web API
2	data.table	Fast aggregation of large data
3	ROAuth	R OAuth Provides interface to OAuth 1.0
4	leaflet	Create and customize interactive maps
5	rtweets	An implementation of calls designed to extract and organize Twitter data via Twitter's REST and stream APIs
6	maps	Display of maps
7	ggplot2	Improve the quality of graph
8	gdata	Used for data manipulation
9	ggmap	Spatial visualization of data from various online sources
10	rio	Import and export the data
11	plotly	Translate the graph into interactive web based version
12	RColorBrewer	Provides the color scheme for graphs and maps
13	MASS	Provides functions and datasets to support Venables and Ripley
14	RCurl	Used for composing HTTP request
15	tm	Provides a framework for text mining
16	wordcloud	Used for making a cloud of words based on the count of words

III. IMPLEMENTATION OF PROPOSED MODEL

The overall methodology is shown in Fig 3. It consists of data acquisition, data preprocessing, clustering, visualization of clusters which is discussed below:

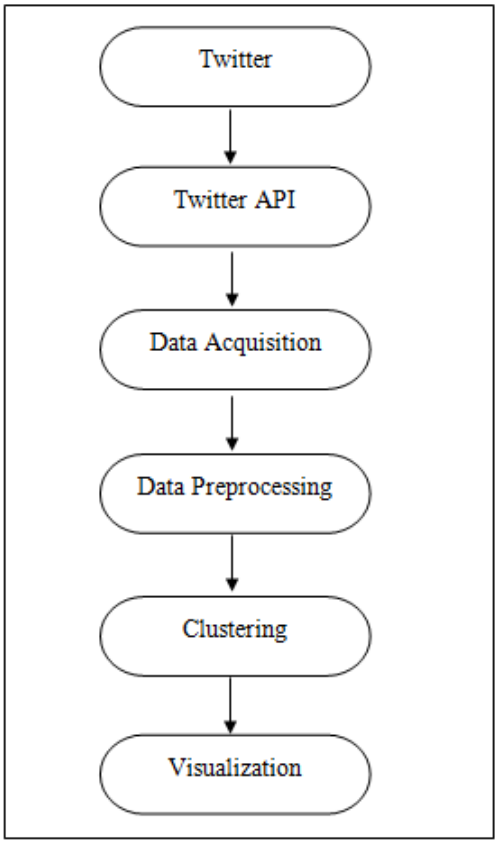

Fig 3. Methodology of Proposed Model

Data Acquisition- For any analysis which is the one of the most important aspect, we require data. Here we are using the Twitter data. In order to extract the twitter data, we are using the *Twitter API* (Application Program Interface). For

extracting the data, one must have a twitter account and the authentication is required for extracting the data. For authentication, first create a twitter developer account for creating an application using *Twitter API*. The credentials such as Consumer Key (API Key), Consumer Secret (API Secret), Access Token and Access Token Secret are generated. With the help of these credentials, we access the Twitter API to access the data.

Data Preprocessing- The fetched data contains some noise and the data needs to be clean for the further analysis. To start with re-tweets, remove the duplicate tweets because they already convey the same information numerous times. Using classical tokenization method split the raw data into the separated words separate by space, comma, special character, etc. Also transform the words into the lower case version and remove the punctuation mark, stop words and URL's. The data is transformed and contains word from the dictionary which is used for the analysis.

Clustering- The clustering algorithm is to be applied on the fetched data based on the information of location. The tweets which are unique and contain only the geographical information are considered for the clustering. The tweets with geographical information means we consider only those tweets which contains latitude and longitude information. The k-means clustering algorithm is applied on the data to classify the twitter tweets into the geographic clusters.

Visualization- The geo clustering can be visualized using *ggplot* package in R. The objective is to visualize on real world geographical map information about the number of tweets originating in a specific region [2].

IV. VISUALIZATION AND DISCUSSION

One way to analyze the data is through data visualization. Data visualization is the representation of data in a graphical or pictorial format. It helps the decision makers to recognize new patterns or grasp difficult concepts to see analytics visually. Using technology we can take the concept ahead with interactive visualization in order to make charts and graphs leading to move detailed description of how we visualize and how the data is processed. Word Clouds, maps, bar chart, pie chart, histogram etc. can be used for visualization.

We first extracted the twitter data in the data frame format from the #MCD results hash tags using the *Twitter Graph API* in R. Then we have considered the unique tweets which contains the geo tagged information using the Google Map *Geocoding* API in the *ggmap* package in R. Fig 4. shows the spatial distribution of tweets. Every dot represents a single tweet and we restricted the area to India.

Apply the k-means clustering algorithm to cluster the tweets. In k-means clustering algorithm, the number of clusters is predefined and selected randomly or generated from the data set using *Elbow method*. The idea behind the elbow method is to run the k-means clustering algorithm on the dataset for a range of values and calculate the sum of squared error (SSE) for each value. Then plot a line chart for each value of the SSE. If the chart looks like an arm, then the value of the elbow of the arm is the best.

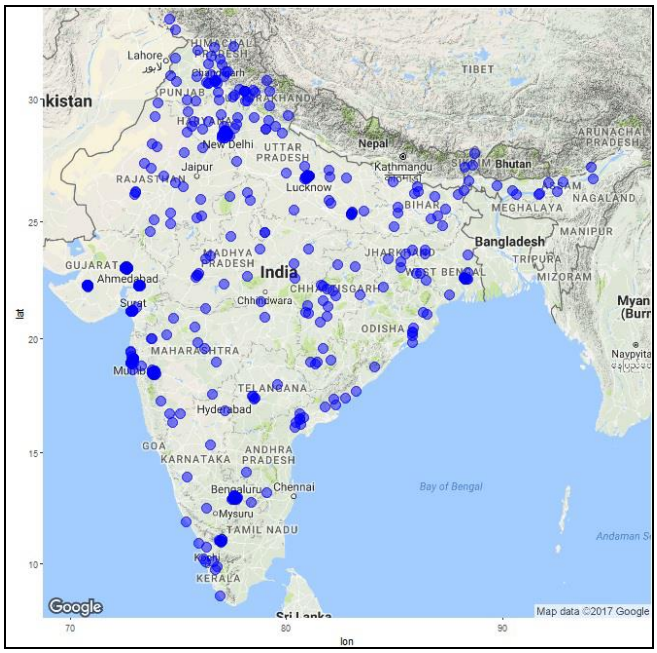

Fig 4. Location of each tweet

The idea is that we want a small SSE, but as we increase the value, the SSE tends to 0. So choose a small value that still has a low SSE. Fig 5. shows the number of clusters found using *Elbow Method.*

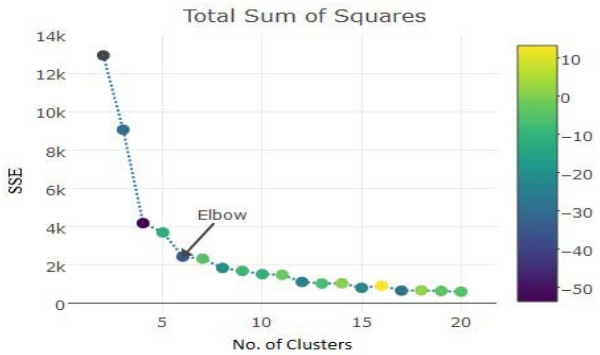

Fig 5. Finding number of clusters using Elbow Method

After finding the number of clusters required for the dataset using *Elbow method*, apply the k-means clustering algorithm based on the geotagged information. The result of the clustering algorithm is represented in Fig 6. In the figure below, it can be clearly observed that in the northern parts

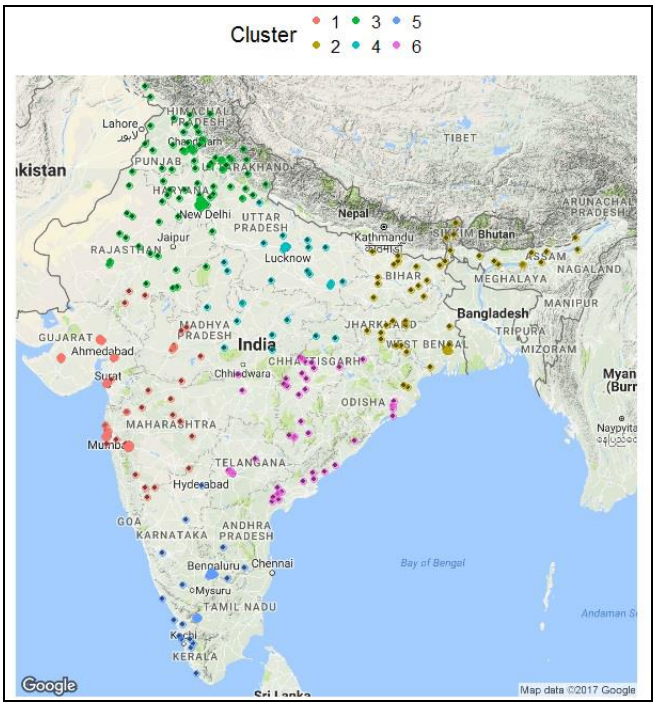

Fig 6. Clustering the tweets

of India there is maximum number of tweets regarding the Delhi MCD results. Specifically, Delhi itself has the maximum number of tweets in northern part of India as shown in the Fig 6. On the other hand, when compared to northern parts of India, the western parts have comparatively less tweets. The number of tweets in central, southern and eastern parts of India is similar in number.

Fig 7. further shows the overlay of tweets represented in Fig 6. *MASS* package and *2-D Kernel density estimation* function have been used to estimate the overlay onto plot. In case there are new tweets with the help of overlay plot we can classify them into specified clusters.

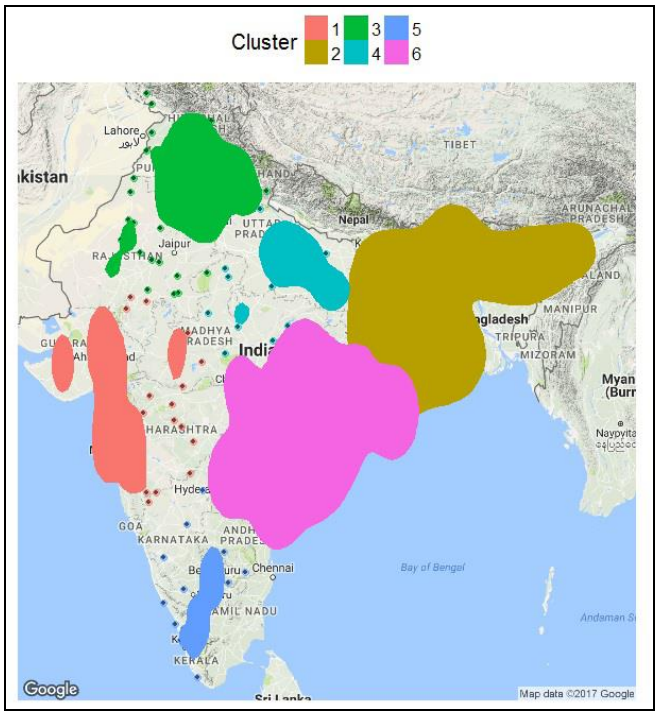

Fig 7. Overlay visualization of Clusters

A word cloud is a graphical representation of text data, which is used to depict keyword metadata (tags) on websites, or to visualize free form text [12]. It is pictorial representation of word frequency. The *tm* and *wordcloud* package is used in R to represent the most used words associated with the hashtag in a pictorial representation. Fig 8. shows the word cloud for word “MCDresults”.

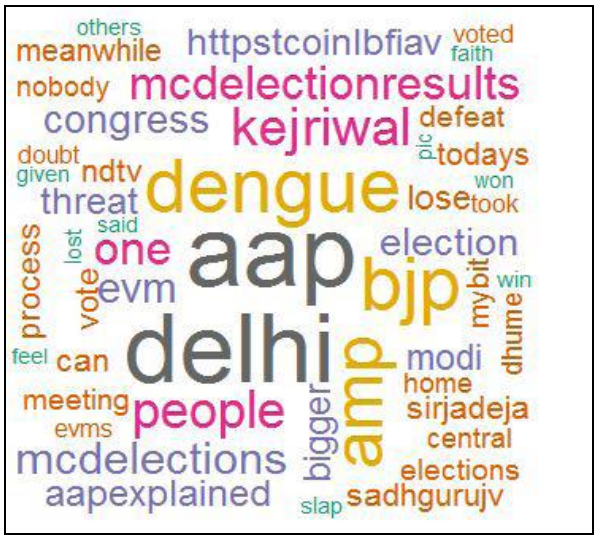

Fig 8. The word cloud

V. CONCLUSION

In this paper, twitter *REST API* is used for the data extraction. R language is used for the data acquisition, preprocessing and visualization of the clustered data. As R is a popular language used for retrieving, cleaning, analyzing and visualizing the data. Therefore, features of R are utilized for sentimental analysis and visualization. We consider the tweets which contain the geotagged information and then proceed to the geo clustering of tweets using k means clustering algorithm. These clustered tweets are then visualized on the Map using the *ggmap* of R. Web 2.0 introduced the concept of word cloud, which helps to visualize the content of a website, the larger the words appeared on a widget, the more frequently they appeared on the site.

VI. REFERENCES

[1] C. Nextmedia, "Social networks overview: Current trends and research challenges," European Commission Information Society and Media, 2010.

[2] A. Sechelea, T. Do Huu, E. Zimos, and N. Deligiannis, "Twitter data clustering and visualization," in Telecommunications (ICT), 2016 23rd International Conference on. IEEE, 16 May 2016, pp. 1–5.

[3] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, no. 14, Oakland, CA, USA., 21 June 1967, pp. 281–297.

[4] A. Sharma and R. Rani, "Community detection and analysis of Twitter social data," in the proceedings of 1st International IEEE Conference 'INBUSH ERA 2015' Theme: Futuristic trends in computational analysis and knowledge management, 25-27 February, 2015.

[5] L. Bijuraj, "Clustering and its applications," in Proceedings of National Conference on New Horizons in IT-NCNHIT, 2013, pp. 169-172.

[6] P. Garg, R. Rani, and S. Miglani, "Analysis and visualization of professionals Linkedin data," in the proceedings of International Conference on Emerging Research in Computing, Information, Communication and Applications, Springer, 31 July - 01 August 2015, pp. 1–9.

[7] https://blog.twitter.com/2009/location-location-location

[8] https://dev.twitter.com/streaming/overview

[9] https://oauth.net/

[10] http://cran.rproject.org/bin/windows/base/

[11]http://ftp.iitm.ac.in/cran

[12] https://en.wikipedia.org/wiki/Tag_cloud