RESEARCH ARTICLE OPEN ACCESS

A Novel Approach for Extraction and Analysis of Tweets

Siddu P. Algur [1], Rashmi H. Patil [2], Prashant Bhat [3]
Department of Computer Science
Rani Channamma University
Belagavi - India

ABSTRACT

Online Social Media networks are being more popular nowadays where, we share rich and timely information about real world events such as sports, films, political issues etc. Twitter is one of the most popular online social media network which generates the up to date news or information throughout the world to the users. The information generated on twitter contains lot of irrelevant data. It is very difficult to extract the relevant data on day to day basis and to perform the analysis. In this paper, an attempt is made to demonstrate the extraction and analysis of tweets (on particular topic) by an extendible toolkit. The extracted data are analyzed using NodeXL which allows users to quickly generate useful network statistics, metrics and visualizations in the context of database.

Keywords: - Online Social Media, NodeXL, Degree of Centrality, Twitter.

I. INTRODUCTION

These roots of social media stretch very deeper. Interacting with family and friends across the long distances has been a concern of humans for centuries. By using Social Medias, people are happily and comfortably communicate to strengthen their relationships. The Social Media gives chance to users to upload a profile and make friends with other users. The first blogging site became popular which was creating a social media sensation that is still popular. Nowadays there are so many Social Medias; some of them are World Wide, Face book, Twitter, LinkedIn, Google+, MySpace etc. Among all these Social Medias, twitter is very popular and fast growing network.

Twitter is one of the popular social network which was created by the programmers (Jack Dorsey, Evan Williams and Biz Stone). On March 21 2006 Jack sent the first tweet as “just setting up my twttr” it would be the beginning of uprising. Now users state or express their feelings in 140 characters or less. One hundred and forty (ie 140) is the number of characters limit allowing users to post a tweet. Nowadays twitter has millions together users. Users share their opinions about any current issues like political, social, environmental, sports, educational, business, film industry etc. It is one of the Social Media which is spreading the news to all over the world.

In twitter, users can form tweet networks, they can follow one another, also the twitter allows to retweet. The twitter network connections are visible in the text of each tweet or by requesting lists of the users that follow the author of each tweet from Twitter.

Today, there is a wealth of Social Media data are coming to us at a steady stream in various format. Tweets contains a rich set of information like a unique numerical IDs which are attached to each tweet, IDs for all the replies, the URL of the author if a website is referenced, the number of followers and many other technical information which can be analyzed. In twitter, it is a challenging task to analyze trending topic and non-trending topics. Topic detection is a fundamental building block to monitor and summarize the information which is originating from social source.

The objective of the study is to extract and analyze topic from twitter data. In the first step it is necessary to define explicitly what constitutes a twitter topic. Then the next step is the extraction of tweets based on explicitly defined twitter topic. One of the most common searches includes this to fetch the term containing particular query. Further the tweets will be analyzed using degree of centrality approach and the results will be analyzed using NodeXL. Also analysis part contains, network graph generated based on the connection between topic to the tweets, with each topic consisting a representation of the tweets that are linked together to that topic i.e. clustering.

Centrality is the number of links incident upon a node (i.e. the number of ties that a node has). There are two types of centralities namely, Betweennes Centrality and Closeness Centrality which are discussed in this paper.

The paper is organized into five sections. The section 2 presents the related work. The section 3 presents with the design methodology. The section 4 presents with the experimental results. Section 5 presents with the conclusion.

II. RELATED WORK

There have been various efforts in the previous years to provide flexible, interactive and effective exploratory interfaces for network analysis [7]. The authors [1] have worked on the datasets extracted from the micro blogging service. They described how a dataset produced using the query term ‘Syria’ can be increased in size to include tweets on the topic of Syria that do not contain that query term. They compare three methods for this task, using the top hash tags from the set as search terms, using a hand selected set of hash tags as search terms and using LDA topic modeling to cluster tweets and selecting appropriate clusters. They described an evaluation method for accessing the relevance and accuracy of the tweets returned. The authors [2] have investigated the practice of sharing short messages (micro blogging) around live media events. They find that analysis of twitter usage patterns around this media event can yield significant insights into the semantic structure and content of the media object. Specifically, they find that the level of twitter activity serves as a predictor of changes in topics in the media event. Further they find that conversational cues can identify the key players in the media object and the content of the twitter posts can somewhat reflect the topics of discussion in the media object, but are mostly evaluative, in that they express the poster's reaction to the media. The key contribution of this work is an analysis of the practice of micro blogging live events. Finally, they offer suggestions on how their model of segmentation and node identification could apply towards any live, real-time arbitrary event.

III. DESIGN METHODOLOGY

AllThe system model of the proposed work is represented in Fig.1 and is consists of the following modules:

1. Tweets extraction module
2. Pre-processing and Refinement module
3. Analysis of tweets module
4. Knowledge Discovery module.

Each of these modules may contain sub modules and they are discussed in the following sections.

3.1 Tweet Extraction Module

This module of the proposed system allows to extract large scale tweets on a particular topic using twitter search network. The extracted tweets contains long information such as name of a person, who tweet, date, tweets, hash tags, import id, follower,URL addresses etc are all extracted and stored in the dataset in the form of xls or csv file. The snapshot of extracted tweets are shown in Fig.2.The data used in this experiment is collected from the official website of the twitter. The size of the dataset is 18000 tweets for the query "modi". The Fig.2 shows the snapshot of the extracted tweets with related fields such as username, date of the tweets, URLs etc. In the Fig.2 each "edge" represents a connection event between two people who tweeted within the data sample period i.e. vertex1 and vertex2. The third column represents relationship edges which represents the various kinds of relationships. The proposed system constructs three different types of twitter edges from the data viz: mentions, replies and tweets. A "mentions" edge is created when one user creates a tweet that contains the name of another user (indicated with a proceeding "@" character, ex: "just spoke about social media with @marc_smith").

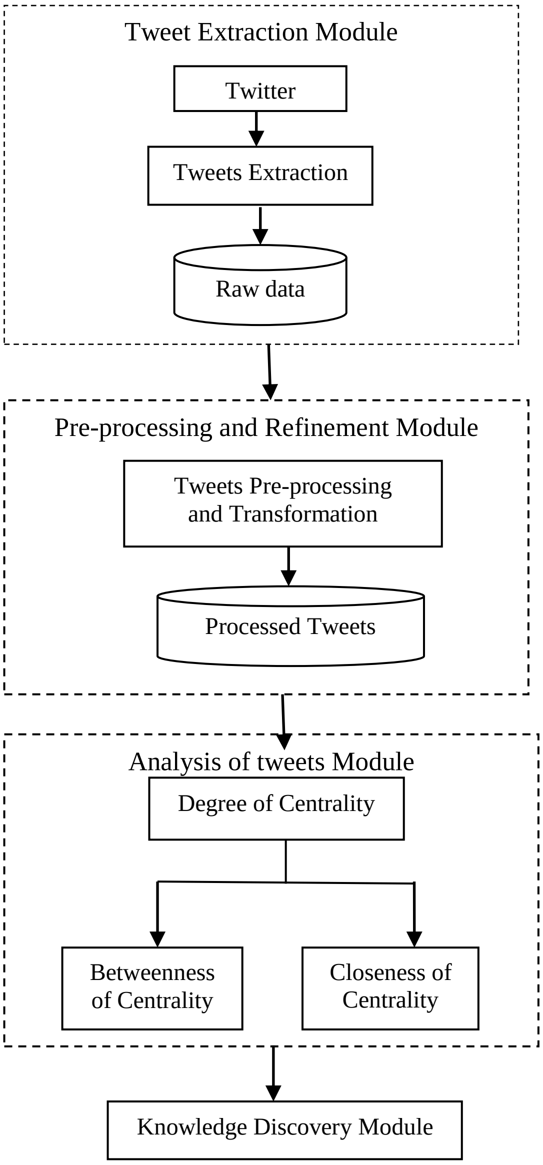
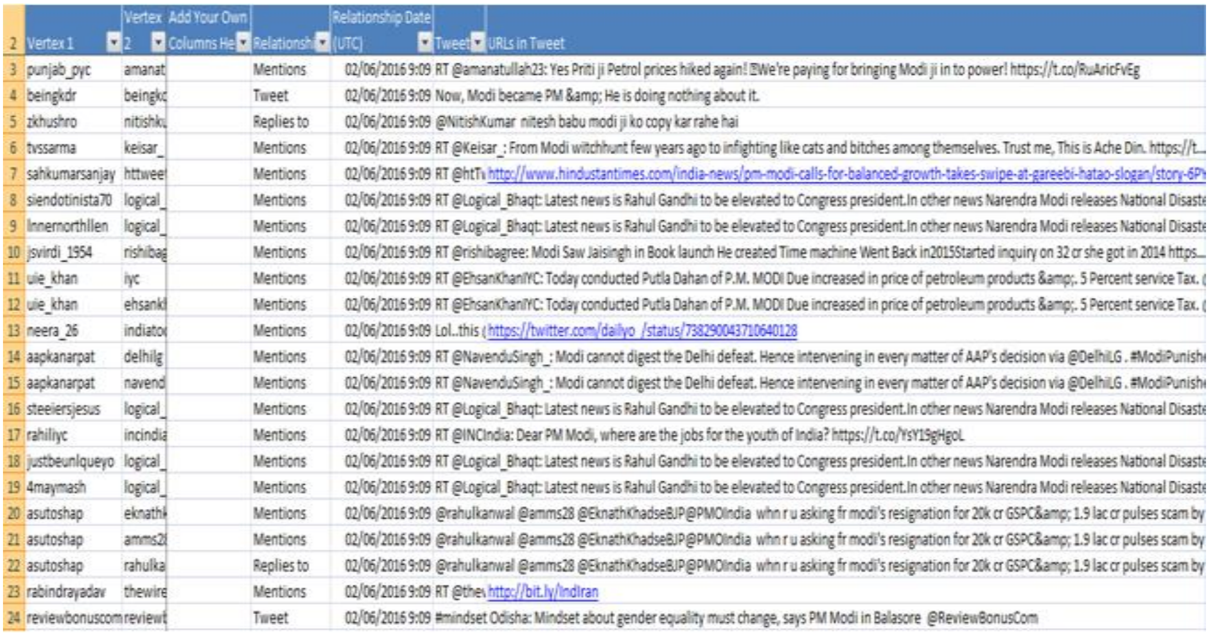

Tweet Extraction Module
Twitter
Tweets Extraction
Raw data
Pre-processing and Refinement Module
Tweets Pre-processing and Transformation
Processed Tweets
Analysis of tweets Module
Degree of Centrality
Betweenness of Centrality
Closeness of Centrality
Knowledge Discovery Module

Fig.1 System Model of the proposed work

A "reply" relationship is a special form of "mention" that occurs when the user's name is at very start of a tweet (ex:"@itaih just spoke about social media"). A "tweet" is a post (message) which includes text, links, photos or videos etc which is represented in fifth column in dataset. The fourth column represents the date of the relationship. The remaining columns are irrelevant for this experiment and hence ignored.



		Vertex	Add Your Own		Relationship Date		
2	Vertex 1	2	Columns He	Relationshi	(UTC)	Tweet	URLs in Tweet
3	punjab_pyc	amanat		Mentions	02/06/2016 9:09	RT @amanatullah23: Yes Priti ji Petrol prices hiked again! □We're paying for bringing Modi ji in to power! https://t.co/RuAricFvEg	
4	beingkdr	beingkc		Tweet	02/06/2016 9:09	Now, Modi became PM &amp; He is doing nothing about it.	
5	zkhushro	nitishku		Replies to	02/06/2016 9:09	@NitishKumar nitesh babu modi ji ko copy kar rahe hai	
6	tvssarma	keisar_		Mentions	02/06/2016 9:09	RT @Keisar_: From Modi witchhunt few years ago to infighting like cats and bitches among themselves. Trust me, This is Ache Din. https://t...	
7	sahkumarsanjay	httweet		Mentions	02/06/2016 9:09	RT @htTv	http://www.hindustantimes.com/india-news/pm-modi-calls-for-balanced-growth-takes-swipe-at-gareebi-hatao-slogan/story-6P
8	siendotinista70	logical_		Mentions	02/06/2016 9:09	RT @Logical_Bhaqt: Latest news is Rahul Gandhi to be elevated to Congress president.In other news Narendra Modi releases National Disaste	
9	Innernorthllen	logical_		Mentions	02/06/2016 9:09	RT @Logical_Bhaqt: Latest news is Rahul Gandhi to be elevated to Congress president.In other news Narendra Modi releases National Disaste	
10	jsvirdi_1954	rishibag		Mentions	02/06/2016 9:09	RT @rishibagree: Modi Saw Jaisingh in Book launch He created Time machine Went Back in2015Started inquiry on 32 cr she got in 2014 https...	
11	uie_khan	iyc		Mentions	02/06/2016 9:09	RT @EhsanKhanIYC: Today conducted Putla Dahan of P.M. MODI Due increased in price of petroleum products &amp;. 5 Percent service Tax. (	
12	uie_khan	ehsankl		Mentions	02/06/2016 9:09	RT @EhsanKhanIYC: Today conducted Putla Dahan of P.M. MODI Due increased in price of petroleum products &amp;. 5 Percent service Tax. (	
13	neera_26	indiatoc		Mentions	02/06/2016 9:09	Lol..this (	https://twitter.com/dailyo_/status/738290043710640128
14	aapkanarpat	delhilg		Mentions	02/06/2016 9:09	RT @NavenduSingh_: Modi cannot digest the Delhi defeat. Hence intervening in every matter of AAP's decision via @DelhiLG . #ModiPunishe	
15	aapkanarpat	navend		Mentions	02/06/2016 9:09	RT @NavenduSingh_: Modi cannot digest the Delhi defeat. Hence intervening in every matter of AAP's decision via @DelhiLG . #ModiPunishe	
16	steeiersjesus	logical_		Mentions	02/06/2016 9:09	RT @Logical_Bhaqt: Latest news is Rahul Gandhi to be elevated to Congress president.In other news Narendra Modi releases National Disaste	
17	rahiliyc	incindia		Mentions	02/06/2016 9:09	RT @INCIndia: Dear PM Modi, where are the jobs for the youth of India? https://t.co/YsY19gHgoL	
18	justbeunlqueyo	logical_		Mentions	02/06/2016 9:09	RT @Logical_Bhaqt: Latest news is Rahul Gandhi to be elevated to Congress president.In other news Narendra Modi releases National Disaste	
19	4maymash	logical_		Mentions	02/06/2016 9:09	RT @Logical_Bhaqt: Latest news is Rahul Gandhi to be elevated to Congress president.In other news Narendra Modi releases National Disaste	
20	asutoshap	eknathk		Mentions	02/06/2016 9:09	@rahulkanwal @amms28 @EknathKhadseBJP@PMOIndia whn r u asking fr modi's resignation for 20k cr GSPC&amp; 1.9 lac cr pulses scam by	
21	asutoshap	amms28		Mentions	02/06/2016 9:09	@rahulkanwal @amms28 @EknathKhadseBJP@PMOIndia whn r u asking fr modi's resignation for 20k cr GSPC&amp; 1.9 lac cr pulses scam by	
22	asutoshap	rahulka		Replies to	02/06/2016 9:09	@rahulkanwal @amms28 @EknathKhadseBJP@PMOIndia whn r u asking fr modi's resignation for 20k cr GSPC&amp; 1.9 lac cr pulses scam by	
23	rabindrayadav	thewire		Mentions	02/06/2016 9:09	RT @thev	http://bit.ly/IndIran
24	reviewbonuscom	reviewt		Tweet	02/06/2016 9:09	#mindset Odisha: Mindset about gender equality must change, says PM Modi in Balasore @ReviewBonusCom	

Fig.2 Extracted tweets

3.2 *Preprocessing and refinement Module*

In this module the extracted raw data will be preprocessed. Because in Twitter, users post real time messages and opinions (tweets). Due to the nature of this micro blogging service (quick and short messages), usually people make spelling mistakes, use emoticons and other characters which gives special meanings. There exists a brief terminology associated with tweets. Emoticons means facial expressions pictorially represented, they express the user's mood. Users of Twitter use the "@" symbol to address the other users on the micro blog. Users usually use hash tags to mark topics, URLs etc. In the proposed model, the raw data i.e., extracted tweets are given as an input and it will be preprocessed and stored in a database as processed tweets. These terminologies are irrelevant. So these terminologies have to work on preprocessing of data for better knowledge discovery. Such preprocessing steps are Tokenization, Stop word removal etc. In the proposed model, the raw data i.e., extracted tweets are given as an input and they will be preprocessed and stored in a database as refined tweets.

3.3 *Analysis of tweets Module*

This module of the proposed system analysis the collected data by applying network metrics and visualization the network. In analysis part:

1. Calculate the network graph and work with groups of vertices.
2. Group the graphs vertices using the values in a column on the vertices worksheet
3. Group the set of vertices connected in a specific way (group the repeating patterns in network).

3.3.1 Degree of Centrality:

Centrality is defined as the number of links incident upon a vertex. The degree can be interpreted in terms of the immediate risk of a vertex for catching whatever is flowing through the network. In the case of directed network usually define two separate measures of degree of centrality, namely out degree and in degree. Accordingly, in degree is a count of the number of ties directed to the node and out degree is the number of ties that node directs to others. When ties are associated to some positive aspects such as friendship or collaboration, in degree is often interpreted as a form of popularity and out degree as sociability. This paper discusses two types of centralities: Betweenness Centrality and Closeness Centrality.

3.3.2 Betweenness Centrality:

Betweenness centrality is a centrality measure of a vertex within a graph and edge betweenness. Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other vertices. It was introduced as a measure for quantifying the control of a human on the communication between other humans in social network vertices. There is a high probability to occur on randomly chosen shortest path between two randomly chosen vertices have high betweenness.

The betweenness can be represented as

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Where σ_{st} is total number of shortest paths from node s to node t and σ_{st} (v) is the number of those paths that pass through v. The betweenness may be normalized by dividing through the number of pairs of vertices not including *v*, which for directed graph is (n-1) (n-2) and for undirected graphs is (n-1) (n-2)/2.

3.3.3Closeness Centrality:

In connected graphs there is a natural distance metric between all pairs of vertices defined by the length of their shortest path. The **farness** of a node *x* is defined as the sum of its distances from all other nodes.

$$C(x) = \frac{1}{\sum_y d(y,x)}.$$

Thus, the more central a node is the lower its total distance from all other nodes. Taking distances *from* or *to* all other nodes is irrelevant in undirected graphs, whereas in directed graphs distances *to* a node are considered a more meaningful measure of centrality, as in general a node has little control over its incoming links. Closeness centrality of a node u is the reciprocal of the sum of the shortest path distances from u to all $n-1$ other nodes. But the sum of distances depends on the number of nodes in the graph; closeness is normalized by the sum of minimum possible distances$n-1$.

$$"C(x) = \frac{n-1}{\sum_{y=1}^{n-1} d(y,x)}"$$

Where d(y,x) is the shortest-path distance between y and x, and n is the number of nodes in the graph.

3.4 Knowledge Discovery Module:

This component of the proposed model explores useful knowledge from twitter database. The knowledge discovery process includes data mining intelligent techniques and models.

IV. EXPERIMENTAL RESULTS

This section represents the graph metrics and grouping the graph's vertices using the values in the column on the vertices worksheet. The groups are usually represented by vertex colour and shape.

4.1 Twitter overall network:

In this part, it describes networks in terms of their number of components and the length of paths in those networks. Here the extracted dataset contains 18000 tweets for the query "modi". Users are heavily interconnected in a network. A range of measurements exist that capture the size and internal connectivity of a network as well as attributes of each node

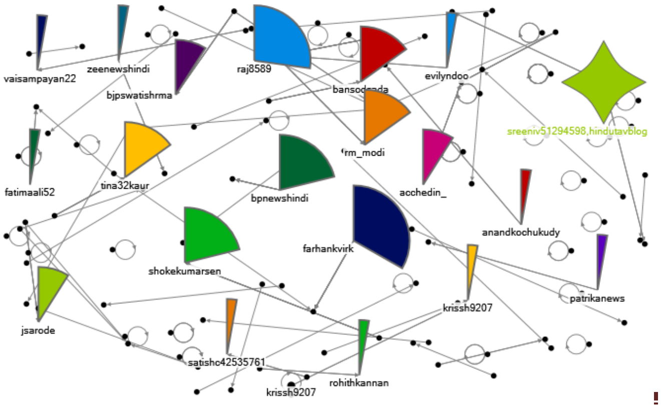
Sl.no	Graph Metrics	Value
1	Graph Type	Directed
2	Vertices	245
3	Unique Edges	184
4	Edges With Duplicates	38
5	Total Edges	222
6	Self-Loops	47
7	Reciprocated Vertex Pair Ratio	0
8	Reciprocated Edge Ratio	0
9	Connected Components	84
10	Single-Vertex Connected Components	30
11	Maximum Vertices in a Connected Component	32
12	Maximum Edges in a Connected Component	33
13	Maximum Geodesic Distance (Diameter)	7
14	Average Geodesic Distance	1.935946
15	Graph Density	0.002709936

TABLE I. Visual representation of network metrics

. The proposed work supports a minimal set of the most crucial network measures for individual vertices: in degree and out degree, clustering coefficient, betweenness centrality, Closeness centrality, network reciprocity, self loops, connected components, graph density and so on. Each of the network metrics has a different dimension of the size and shape of the graph. The Table1. Shows the summarization of key properties of complete graph.

4.2 Group by Motif (Group the repeated patterns in network)

A Tweet Motif, an exploratory search application for Twitter. Unlike traditional approaches to information retrieval, which present a simple list of messages, Tweet Motif group messages by frequent significant term result set's subtopics - which facilitate navigation and drilldown through a faceted search interface. Tweet Motif's subtopic groupings make it easy to obtain both an overview and specific examples of what people are saying;



Created with NodeXL (http://nodexl.codeplex.com)

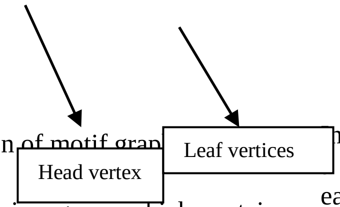
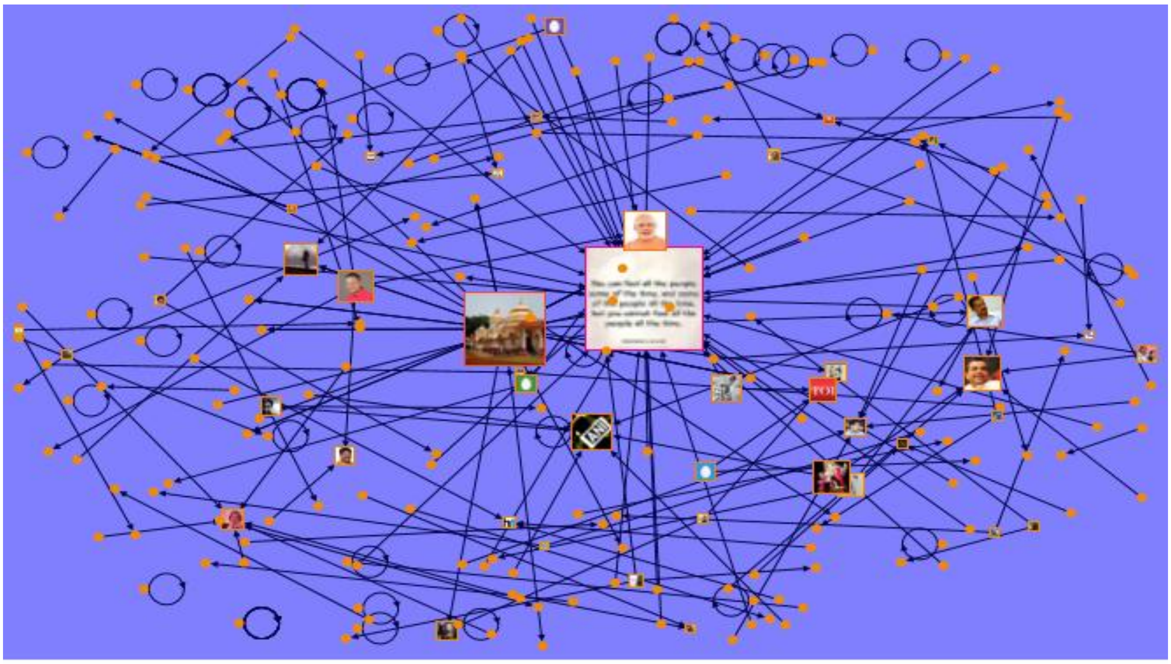


Fig.3Visual representation of motif grap

The Fig.3 shows that each sector is a group which contains many users (leaf vertex) tweet the same content of words to a same person (head vertex).

4.3 Betweenness Centrality:

n the Fig.3 the blue colour sector contains 61 leaf vertices to head vertex farhankvirk (name of the head vertex) like this each size and shape of the sector increases depending upon the how many leaf vertices it contains.

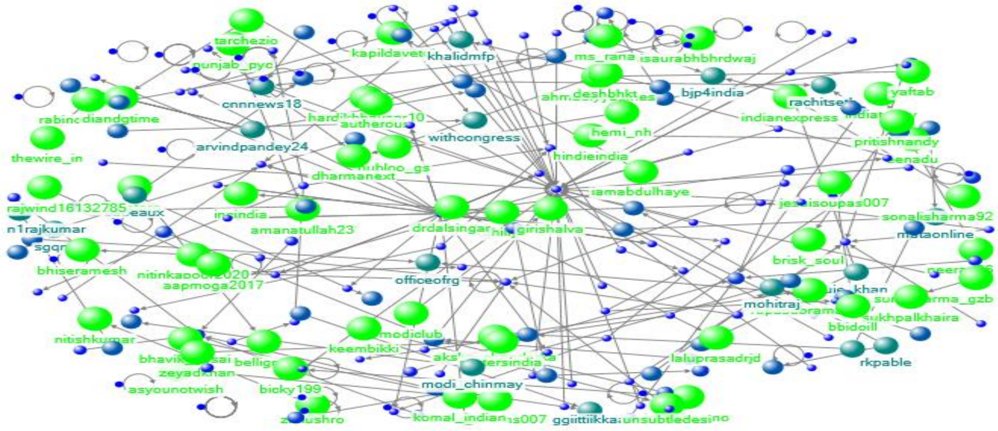


Created with NodeXL (http://nodexl.codeplex.com)

Fig.4 Network graph of betweenness centrality

The Fig.4 shows that graph size of images represents the value of betweenness centrality. The dots are representing user who are not involve in betweenness centrality. In the Fig.4 the image of modi is a centrality with the tweet “This Pigeon is the Biggest Threat to PM Modi: Indian Media: Apparently pigeons are a credible threat to nation”. There are 25 tweet vertices connected with modi vertex.

4.3.1 Closeness Centrality:



Created with NodeXL (http://nodexl.codeplex.com)

Fig.5 Network graph of closeness centrality

The Fig.5 shows that graph disks are representing vertices, size of disks are representing value of closeness centrality and colour represents in which manner they are closed in in degree or in out degree

V. CONCLUSIONS

Network structures are important in many disciplines and professions. Interest in these structures is growing more common as the world of social networks and computer-mediated social content becomes more main stream. In this paper aims to make extraction, analysis and visualization of network data easier by combining the common analysis and visualization functions with the familiar spreadsheet. Extracting the tweets by entering a particular query in the import option. In analysis part, calculate the overall graph metrics for the visualized graph and also calculate the centralities of network graphs. Then work with grouping of some related vertices. The tool enables essential network analysis tasks and thus supports a wide audience of users in a broad range of network analysis scenarios.

ACKNOWLEDGMENT

I am very much thankful to Mr. Basavraj G. Research Scholar, Department of Computer Science, Rani Channamma University, Belagavi for his suggestions.

REFERENCES

[1] Clare Llewellyn, Claire Grover, Beatrice Alex, Jon Oberlander, and Richard Tobin were “Extracting a Topic Specific Dataset from a Twitter Archive”. School of Informatics, University of Edinburgh, United Kingdom. Springer **2015.**

[2] David A. Shamma ,Lyndon Kennedy ,Elizabeth F. Churchill “Understanding Community Annotation of Uncollected Sources” Tweet the Debates Internet Experiences Yahoo! Research 4301 Great America Parkway, 2GA-2615 Santa Clara, CA, 95054 USA.

[3] Sterne, JohnWiley, “Social Media Metrics: How to Measure and Optimize Your Marketing Investment”, 2010.

[4] Heer, J. and boyd, d. Vizster: “Visualizing online social networks, IEEE Symposium on Information Visualization” 2005.

[5] Marc A. Smith, Ben Shneiderman, Natasa Milic-Frayling, Eduarda Mendes Rodrigues,Vladimir Barash, Cody Dunne, Tony Capone, Adam Perer, Eric Gleave “Analyzing (Social Media) Networks with NodeXL”C&T’09, June 25–27, 2009, University Park, Pennsylvania, USA.ACM 978-1-60558-601-4/09/06.

[6] R. Cross , R.J. Thomas , Driving Results through Social Networks: How Top Organizations Leverage Networks for Performance and Growth , John Wiley & Sons , San Francisco, CA , 2009 .

[7] J. Heer, D. Boyd, Vizster: Visualizing Online Social Networks, in: Proc. 2005 IEEE Symposium on Information Visualization (October 23-25, 2005), INFOVIS. IEEE Computer Society, Washington, DC, 2005.

[8] L.C. Freeman , Centrality in social networks: conceptual clarifi cation , Social Networks 1 (1979) 35 – 41 .