Tweet Analytics and Tweet Summarization using Graph Mining

Apeksha P. Naik
PG Student, CMPN Department, SLRTCE
Mumbai University
Mumbai, India
apekshanaik06@gmail.com

Sachin Bojewar
Associate Professor, IT Department, VIT
Mumbai University
Mumbai, India
sachin.bojewar@vit.edu.in

Abstract – Social networking sites, like Twitter, have become the fastest means of communication. Millions of user tweet everyday discussing recent and important issues. Generally, the tweets are identical or similar in nature, which causes information overload on user's wall. This makes it difficult for the user to keep a track of all the events. The best solution for this is to summarize tweets that are similar, making it easier for user to understand and decide which tweets to follow. In this paper, we present a graph based clustering technique to generate summary for tweets that are similar or identical. In addition, the paper describes about the analytics performed on tweets. Analyzing tweets help in determining the popularity of a topic and in knowing user-interested topics. Twitter analytics is the key to measure the success of the tweets posted. The proposed system gives better results compared to other existing systems.

Keywords— Twitter, Tweets, Tweet Summarization, Graph Clustering, Twitter Analytics, Social Network.

I. INTRODUCTION

Twitter has become a popular online social networking site for sharing real-time information on recent and popular events. At present, lot of research is being conducted for efficiently utilizing the large amount of information posted on Twitter by different users. The research includes areas like detecting communities in social networks [1], [2], [9], methods for summarization [3], [4], [11], analysing tweets [5], [13] and so on.

Millions of tweets are posted everyday on Twitter, especially during important events like natural calamities, change in currency, elections, other political issues, etc. Tweets related to such events are generated at faster rates. Like most existing systems, Twitter also uses keyword-matching technique to search and retrieve relevant tweets. If the search keyword matches then it retrieves all the tweets related to that keyword. Many users tweet simultaneously, Twitter contains lot of redundant information. Practically, it is impossible for any user to go through all the tweets and keep a track of events. In addition, this redundancy leads to information overload for the user. In such cases, the best possible approach is to summarize tweets and provide user only with the overview about a topic. Summarization can be useful for quick understanding about any event [6], [9].

Analytics is an emerging trend in social media. It is important to realize that simply having a social media presence is no guarantee of success. Hence, analytics is a crucial part for social media success. The most common part of analytics for twitter is to mine customer sentiments, improve public opinion, get feedback on tweets, to determine locations of user, tweet counts for a particular topic or event, etc. Various tools have been developed in recent years to carry out analytics on tweets posted on daily basis [14].

In this paper, we use a well-established method for summarizing tweets that generates variable length summary for different tweets. This project proposes a graph-based approach for summarizing tweets, where a graph is first constructed considering the similarity among tweets and then a standard graph-clustering algorithm is used to identify similar tweets (nodes). The similarity among tweets is measured using various features including features based on WordNet synsets, which help to capture the semantic similarity among tweets. The proposed approach achieves better performance than other existing summarization technique SumBasic [1], [7].

Further to help user understand how the content shared on Twitter grows, Twitter Analytics is performed [5]. Analytics help to determine the success of information posted. It helps user to understand the trending topics, events or news-stories. The proposed system analyzes factors like number of hashtags, favorite tweets, retweets and location of tweets. A location based graph is generated which shows different cities/states having tweet to a particular event.

II. BACKGROUND STUDIES

The present work proposes the tweet summarization algorithm which finds similarity between different tweets to construct tweet-similarity graph and cluster similar tweets identified from that graph to generate summary. This section briefly explains how to identify similarity between tweets using WordNet [8], [10] and also discusses clustering technique [4], [12] applied on the graph to identify and group similar tweets. It further focuses on the parameters used for twitter analytics.

A. WordNet

WordNet [8] is a large lexical database of English language. Nouns, verbs, adjectives and adverbs are grouped together into sets of synonyms, called synsets. Each Synset expresses a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. Superficially, WordNet resembles a thesaurus as it groups words together based on their meanings. However, there lies some important distinctions. WordNet interlinks not only word forms—strings of letters—but also specific senses of words. As a result, words that are found in close proximity to one another in the network are also grouped together. For example, the words 'plant' and 'flora' are grouped together in common synset. Similarly, the words 'happy' and 'joy' are synonyms are stored in the same synset. Thus, WordNet can be called as a combination of dictionary and thesaurus. It also provides the hierarchical structure for related words [1].

B. Clustering Similar Tweets

Once the tweet similarity graph is created, the next step is detecting and clustering similar tweets from that graph. Similar types of tweets in a network forms a community. The edges that connects the nodes inside a community are called Intra-community edges and the edges connecting the nodes outside or in different communities are called Inter-community edges. Figure 1 shows two clusters having Inter-community and Intra-community edges. In the present work, we construct a graph where nodes are tweets and the edges represent the similarity among the tweets. We then cluster similar tweets from tweet similarity graph using the K-means algorithm. Once the similar tweets are clustered, a representative from each cluster is selected to be included in the summary.

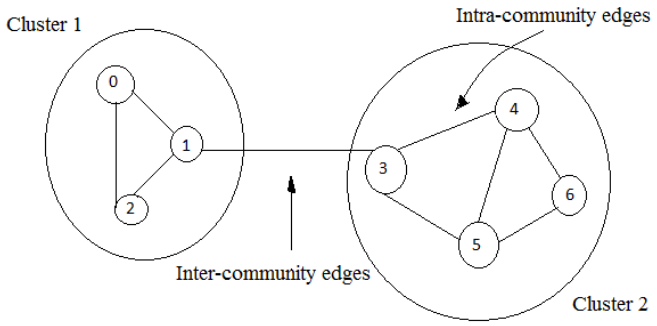

Figure 1. Graph showing intra-community and inter-community edges.

C. Twitter Analytics

Visual analytics on any tweet can be performed. Analytics makes it easy to monitor or keep a track of tweets from your favorite users, location, lists or search keyword. Twitter analytics also includes tweet reach count, engagement rate, etc. It is also possible to have a graph for tweet statistics, which includes tweets per hour, tweets per month, etc. Figure 2, shows the graph for tweet statistics. With the use of the newly, available Twitter analytics feature, users are able to measure Mentions, Followers, hashtag count and demographics of Followers like gender and location. The propose system takes into account various factors like tweet count, retweets, favorite tweets, number of hashtags, location of the tweet.

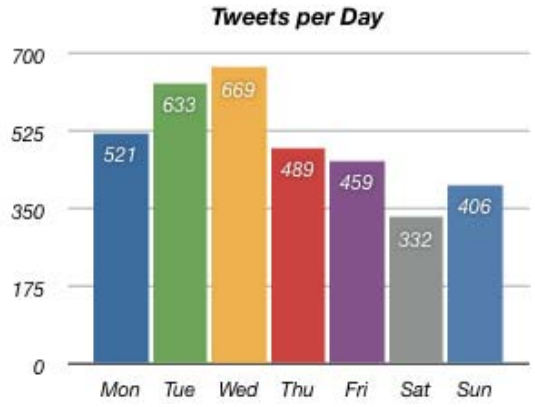

Figure 2. Graph Showing Tweet statistics.

III. PROPOSED SYSTEM

For a given set of tweets, the proposed methodology identifies a small subset of the tweets as a summary of the entire set of tweets. This section describes the proposed methodology in detail. Figure 3 shows the process for tweet summarization.

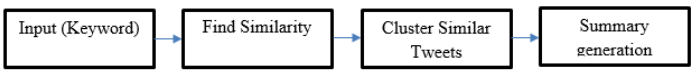

Figure 3. Block diagram for Tweet Summarization.

Consider a scenario, where we use the sample search keyword 'demonetization'. The proposed system accepts the input keyword, retrieves all the tweets related to that keyword. We attempt to find similarity among tweets in two ways: term-level similarity and semantic similarity.

The term-level similarity [1] between a pair of tweets is measured based on common terms in the tweets. Specifically we consider similarity measures like: similar URL count, similar Hashtag count, similar Username count, Cosine similarity and finally we find the Levenshtein distance. The term level similarity for the sample keyword, demonetization, is shown Figure 4.

sId	tIdFrom	tIdTo	sUrl	sHashTag	sUsernam	sCosine	Levenshte	Synset	PathSimila	TotalWeig
1	1	2	0	0	0	0.9485...	94			
2	1	3	0	0	0	0.9532...	109			
3	1	4	0	0	0	0.9546...	96			
4	1	5	0	0	0	0.9273...	103			
5	1	6	0	0	0	0.9774...	74			
6	1	7	0	0	0	0.9610...	104			
7	1	8	0	0	0	0.9533...	107			
8	1	9	0	0	0	0.9583...	107			
9	1	10	0	0	0	0.9580...	104			
10	1	11	0	0	0	0.9821...	66			
11	1	12	0	0	0	0.9579...	101			
12	1	13	0	0	0	0.9442...	93			
13	1	14	0	0	0	0.9652...	96			
14	1	15	0	0	0	0.9571...	100			
15	2	3	0	0	0	0.9541...	103			
16	2	4	0	0	0	0.9545...	102			
17	2	5	0	0	0	0.9181...	114			
18	2	6	0	0	0	0.9554...	86			
19	2	7	0	0	0	0.9599...	107			
20	2	8	0	0	0	0.9608...	108			
21	2	9	0	0	0	0.9619...	107			

Figure 4. Figure showing Term-level similarity

The *cosine similarity* [1], cos(Θ), for two vectors of attributes A and B is given as:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \qquad (1)$$

We consider that the vectors are comprised of distinct words in the string while computing cosine similarity between two strings.

Levenshtein distance (LD) is a measure of the similarity between two strings, which is referred as the source string (s) and the target string (t). The distance is the number of deletions, insertions, or substitutions required to transform s into t For instance, given the two strings 'tent' and 'test', the Levenshtein Distance is 1, whereas for the strings 'kitten' and 'sitting', the Levenshtein Distance is 3. The greater the Levenshtein distance, the more different the strings are.

The *semantic similarity* is also determined along with term-level similarity The WordNet [8] tool, described in previous section, is used to identify semantically similar terms. As mentioned, two terms are said to be semantically similar if they belong to a common WordNet synset. The semantic similarity also finds the path similarity [1] between tweets. If two tweets contain more than two semantically similar words we consider two tweets to be semantically similar. Finally, the similarity score between two tweets is computed by simple summation of the parameters of term-level and semantic similarities. Figure 5 shows the semantic similarity for the sample keyword demonetization.

sId	tIdFrom	tIdTo	sUrl	sHashTag	sUsernam	sCosine	Levenshtn	Synset	PathSimila	TotalWeig
1	1	2	0	0	0	0.9485...	94	1	0.6089...	96.56
2	1	3	0	0	0	0.9532...	109	1	0.5902...	111.54
3	1	4	0	0	0	0.9546...	96	1	0.5136...	98.47
4	1	5	0	0	0	0.9273...	103	1	0.3436...	105.27
5	1	6	0	0	0	0.9774...	74	1	0.8373...	76.81
6	1	7	0	0	0	0.9610...	104	1	0.5406...	106.5
7	1	8	0	0	0	0.9533...	107	1	0.5982...	109.55
8	1	9	0	0	0	0.9583...	107	1	0.6360...	109.59
9	1	10	0	0	0	0.9580...	104	1	0.5475...	106.51
10	1	11	0	0	0	0.9821...	66	1	0.8286...	68.81
11	1	12	0	0	0	0.9579...	101	1	0.6533...	103.61
12	1	13	0	0	0	0.9442...	93	1	0.5231...	95.47
13	1	14	0	0	0	0.9652...	96	1	0.6434...	98.61
14	1	15	0	0	0	0.9571...	100	1	0.6922...	102.65
15	2	3	0	0	0	0.9541...	103	1	0.6316...	105.59
16	2	4	0	0	0	0.9545...	102	1	0.5456...	104.5
17	2	5	0	0	0	0.9181...	114	0	0.2702...	115.19
18	2	6	0	0	0	0.9554...	86	1	0.6277...	88.58
19	2	7	0	0	0	0.9599...	107	1	0.5163...	109.48
20	2	8	0	0	0	0.9608...	108	1	0.7014...	110.66
21	2	9	0	0	0	0.9619...	107	1	0.5236...	109.49

Figure 5. Figure showing Semantic similarity

We then generate a weighted graph based on similarity scores among the tweets. This graph is referred as tweet-similarity graph.
The *clustering technique* is then applied to group similar tweets. The k-means algorithm is used for clustering. First, we normalize the raw data. Once normalization is done the code results the clusters created. Based on the clustering, different types of summary are generated.

Proposed Technique for Summarization:

1. Accept search keyword from the user.
2. Retrieve the tweets related to the search keyword.
3. Calculate the term-level similarity.
4. Calculate the semantic similarity.
5. Cluster similar tweets.
    a. Enter the number of clusters.
    b. Normalize the raw- data.
    c. The total weight parameter of the tweet similarity graph is used for clustering.
    d. Clustering is completed when there are no more tweets to move.
6. After clustering, generate summary for the retrieved tweets.

Tweet Analytics

The proposed system takes into account various analytical factors like: tweet count, favorite tweets, retweets, hashtag count, username, location etc. When the user enters the search keyword, all the information related to the tweets are saved in background. This information helps a lot while analyzing tweets. The system maintains a log table, which stores the information for the above-mentioned factors for various tweets. The log table maintained by the system is shown in Figure 6.

tw_id	keyword	tweet_text	hashtag	username	location	retweet_count	fav_count
1	demonetisation	RT @AITCoffici...	#DeMonetisati...	muntaziraapka		2	4
2	demonetisation	RT @dobozapp...	#DeMonetisati...	DhandreGaurav	Nagpur, Mahar...	3	3
3	demonetisation	81 officers repo...		darpans	New Delhi, India	0	0
4	demonetisation	#FX16NEWS #N...	#FX16NEWS #...	FX16NEWS	Chennai, India	0	0
5	demonetisation	RT @panku_: S...	#DeMonetisati...	singh__avijit	?????, ????	13	6
6	demonetisation	RT @goyalsanj...		SickularLibtard	???? ???? ??????? !!	137	102
7	demonetisation	#Demonetisati...	#Demonetisati...	ahmdebasir	srinagar	0	0
8	demonetisation	RT @kamaalrkh...	#DeMonetisati...	ShehzadaSalim		94	151
9	demonetisation	Third try , was a...	#DeMonetisati...	shantanugupta	Bangalore , India	0	0
10	demonetisation	RT @myvoteto...	#politicians #D...	AnuSinghSolanki	Patna	6	8
11	demonetisation	RT @bsindia: #...	#DeMonetisati...	Vikas_chirag	Haryana, India	9	2
12	demonetisation	"Slowdown for ...		ashkronos	Mars	0	0
13	demonetisation	RT @SriramNY:...	#Transforming...	BroadWit	Cyberia	1	1
14	demonetisation	RT @AskThePa...		pariraja1	Mumbai, India	9	17
15	demonetisation	RT @kamaalrkh...	#DeMonetisati...	chuzpaa		94	151
16	attacks of 26/11	RT @shubh888...	#NeverForget ...	libran1304	??????	37	6
17	attacks of 26/11	RT @sanjaybaf...	#Mumbai	vinodjain220220		19	19
18	attacks of 26/11	RT @sanjaybaf...	#Mumbai	VickyKanungo	India	19	19
19	attacks of 26/11	RT @sanjaybaf...	#Mumbai	tusharshTT	New Delhi	19	19
20	attacks of 26/11	RT @sanjaybaf...	#Mumbai	sanjaybohrajj	Mumbai	19	19
21	attacks of 26/11	RT @sanjaybaf...	#Mumbai	sureshdoshi42	Mumbai	19	19
22	attacks of 26/11	RT @sanjaybaf...	#Mumbai	sureshdoshi	Mumbai	19	19
23	attacks of 26/11	RT @sanjaybaf...	#Mumbai	riteshpunatar	Mumbai | India	19	19
24	attacks of 26/11	RT @sanjaybaf...	#Mumbai	Priyanka_Ji	Mumbai, India	19	19

Figure 6. Figure showing Log Table

The table above shows the log maintained for the sample search keyword. The Block diagram for the analysis process is shown below in Figure 7.

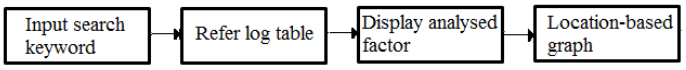

Figure 7. Block diagram for Tweet Analytics

Finally a location based graph is generated which shows tweets from geographically diverse areas.

The Proposed methodology for Analytics:

1. Accept search keyword from user.
2. Refer the log table.
3. Display the factors accordingly.
4. Generate location-based graph.

IV. RESULTS AND DISCUSIONS

As described in the previous section, we have different types of summaries generated for different tweets. The summary generated by the proposed method is accurate. We generate a summary for each topic entered as the search keyword.

This paper broadly focuses on the details of the Tweet Summarization using Graph Mining and the Tweet Analytics. The clustering technique is studied and applied properly to ensure appropriate grouping of similar tweets take place giving the final summary in output. We have tested our work on several recent events or topics. The detailed performance of the proposed method is given in this paper. The community detection technique is used to cluster similar tweets thereby generating appropriate summary for different events or news-story. Figure 8 shows the summary is generated using clustering technique for search-keyword (demonetization). Later part gives results for the analytics performed on searched keyword. It includes retweets, favorite tweets and number of hashtags in the given tweet. It also gives a location-based graph representing tweets in diverse areas. The location-based graph results tweets from various areas. While fetching the tweets for the search keyword, the location of that tweet along with other factors, previously mentioned, is stored in the log table in background. The log table is then referred while generating the graph based on location. Thus, the graph shows the popularity of tweets in different states, cities, etc. Figure 9 and Figure 10 shows the analytics performed for the sample keyword used.

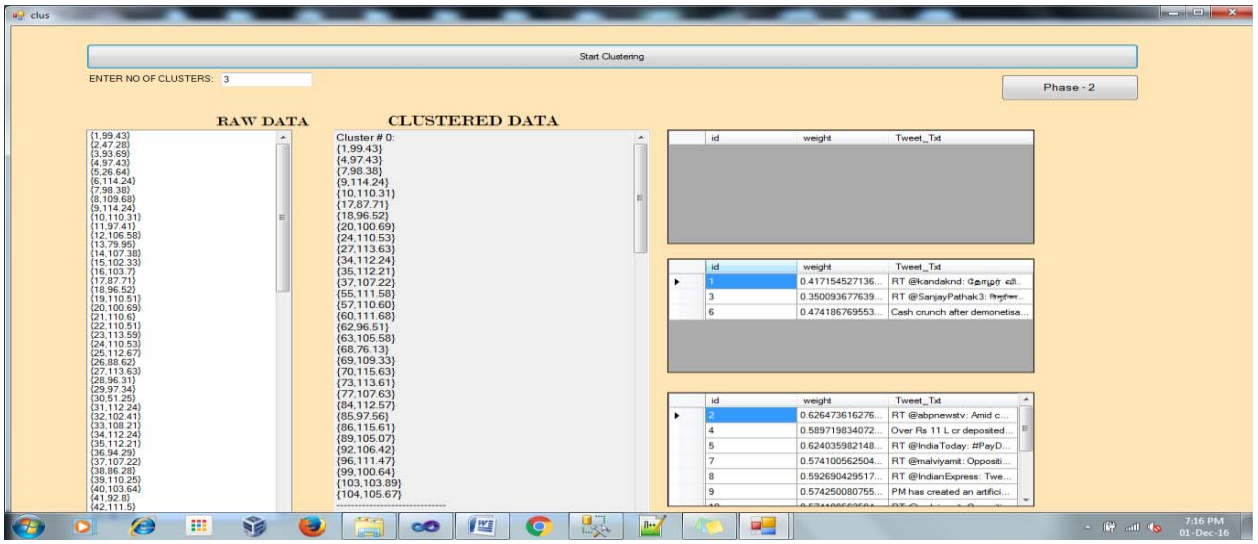

Figure 8. Results for Tweet summarization

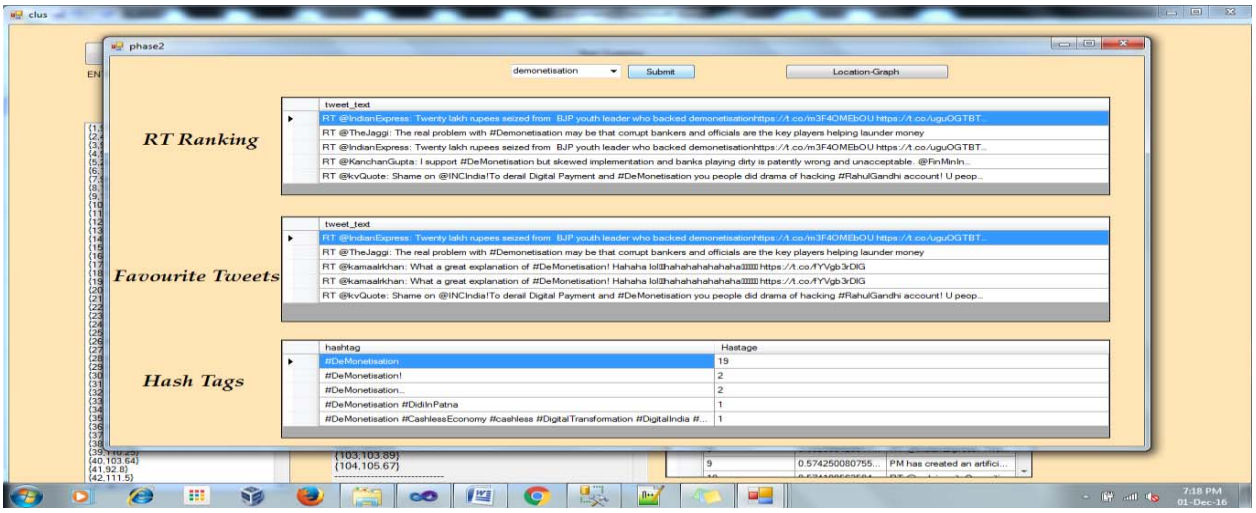

Figure 9. Results for Tweet Analytics

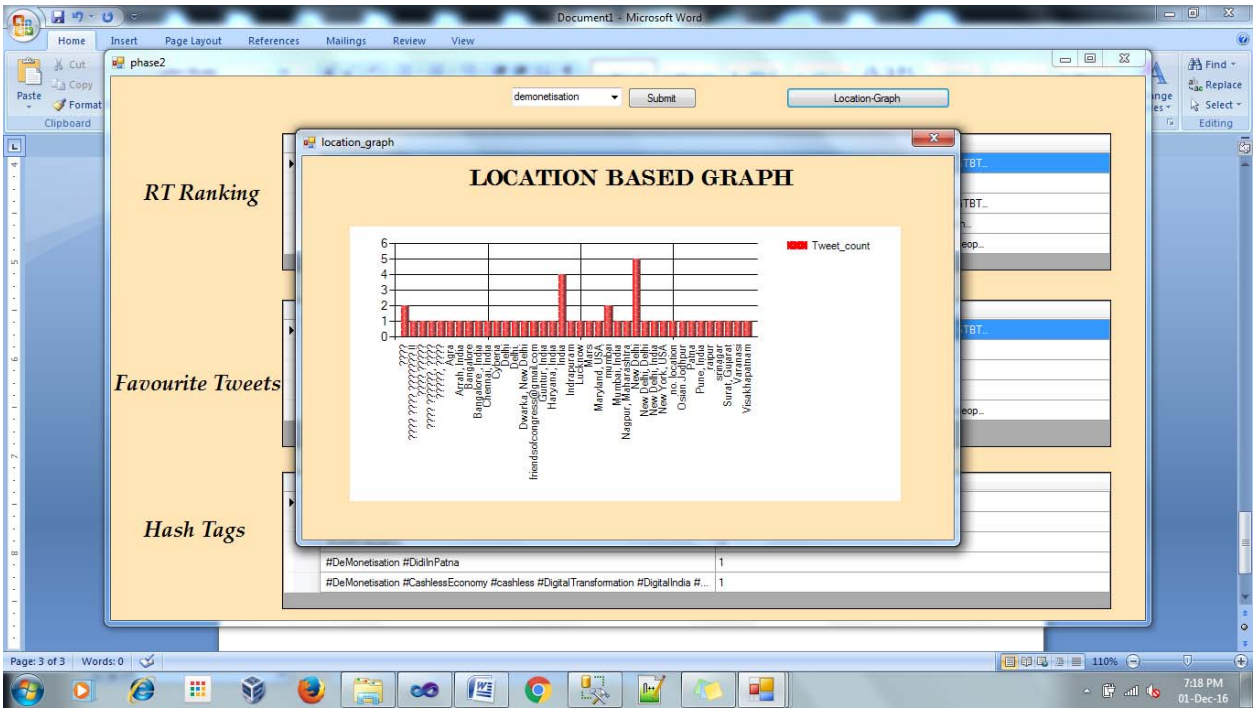

Figure 10. Location-based graph for the search keyword

V. CONCLUSION

This work proposed an efficient method for tweet summarization and tweet analytics. First, a tweet similarity graph is constructed, then a standard graph-clustering algorithm is used to identify similar tweets and finally we generate summary. Twitter Analytics help user to understand the valuable content. The metrics from Twitter are valuable for determining the success and importance of the tweets posted. Analytics provides a clear picture to understand whether the content produced is interesting to the audience or not.

The length of the summary generated by any of the proposed method is not fixed, since it depends on the number of clusters identified by the algorithm. In addition only those tweets are retrieved which includes the search keyword. The proposed system does not consider the synonym for the search keyword while retrieving tweets.

As future work, a study can be carried out to develop approaches, which can give a summary of a specified length, and to develop a method that will also consider the synonym for the search keyword while retrieving the tweets. Also other factors like engagement rate and tweet reach count can be added in the current system for analytics purpose.

REFERENCES

[1] Soumi Dutta, Sujata Ghatak, Moumita Roy, Saptarshi Ghosh, Asit Kumar Das, 'A Graph Based Clustering Technique for Tweet Summarization', 978-1-4673-7231-2/15/$31.00 ©2015 IEEE

[2] Bapuji Rao, Anirban Mitra "A New Approach for Detection of Common Communities in a Social Network using Graph Mining Techniques", *High Performance Computing and Applications (ICHPCA), 2014 International Conference_*, 978-1-4799-5958-7/14/$31.00 ©2014 IEEE.

[3] F. Ahmed, J. Erman, Z. Ge, A. X. Liu, J. Wang, and H. Yan, "Detecting and localizing end-to-end performance degradation for cellular data services", *in Proc. 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, Portland, OR, USA*, June 15-19, 2015, pp. 459-460, 2015.

[4] Nikhil Garg, Benoit Favre, KorbinianReidhammer, DilekHakkani-T¨ur "ClusterRank: A Graph Based Method for Meeting Summarization", *Idiap-RR publication,* Idiap-RR-09-2009. June 2009.

[5] Xintian Yang, AmolGhoting, YiyeRuan "A Framework for Summarizing and Analyzing Twitter Feeds" *KDD'12*, August 12–16, 2012, Beijing, China.

[6] Marina Litvak, Mark Last, Menahem Friedman "A new Approach to Improving Multilingual Summarization using a Genetic Algorithm", *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 927–936, Uppsala, Sweden*, 11-16 July 2010. c 2010 Association for Computational Linguistics.

[7] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova, "Beyond sumbasic: Task-focused summarization with sentence simpliffcation and lexical expansion," *Inf. Process. Manage*., vol. 43, no. 6, pp. 1606- 1618, November 2007

[8] Banshider Majhi, Y Santhosh Reddy, D Prasanna Babu, "Novel Features for Off-line Signature Verification", International Journal of Computers, Communications & Control Vol. I (2006), No. 1, pp. 17-24

[9] LIU X iaoHua1, LI Y iTong, WEI FuRu, ZHOU Ming "Graph-based Multi-tweet Summarization Using Social Signals" *Proceedings of COLING 2012: Technical Papers, pages 1699–1714, OLING 2012*, Mumbai, December 2012.

[10] Georgr A. Miller "Wordnet - a lexical database for English" *describes the use of WordNet tool*, showing comparisons with other techniques and determining the benefits of WordNet tool.

[11] Rao, Bapuji and Mitra, A., "An approach to study properties and behavior of Social Network using Graph Mining Techniques", *DlGNATE 2014: ETEECT 2014*, India, 2014.

[12] Mitra A., Satpathy S. R. Paul S. "Clustering analysis in social network using Covering Based Rough Set", *Advance Computing Conference (IACC)*, *2013 IEEE 3rd International*, India, 2013/2/22, 476-481,2013

[13] K. Sparck Jones, "Automatic summarising: The state of the art. Inf. Process. Manage.", vol. 43, no. 6, pp.1449-1481, November 2007

[14] Rohan D.W Perera, S. Anand, K. P. Subbalakshmi and R. Chandramouli "Twitter Analytics: Architecture, Tools and Analysis", *The 2010 Military Communications conference* - Unclassified Program - Cyber Security and Network Management.