Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 8, Issue, 12, pp.43599-43603, December, 2016

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

DYNAMICALLY EXTRACTING REPRESENTATIVE TWEETS OF A DOMAIN AS THEY EVOLVE

[1,*]Jonnalagadda Surya Kiran, [2]Dwarapu Suneetha and [3]Prof. Mogalla Shashi

[1]M.Tech CST with Artificial Intelligence and Robotics, Department of Computer Science and Systems Engineering, Andhra University College of Engineering (A), Visakhapatnam, Andhra Pradesh, India
[2]Research Scholar, Department of Computer Science and Engineering, Jawaharlal Nehru Technological University College of Engineering Kakinada, Kakinada, Andhra Pradesh, India
[3]Professor, Department of Computer Science and Systems Engineering, Andhra University College of Engineering (A), Visakhapatnam, Andhra Pradesh, India

ARTICLE INFO

Article History:
Received 23rd September, 2016
Received in revised form
12th October, 2016
Accepted 19th November, 2016
Published online 30th December, 2016

Key words:
Hash tag,
Tweet Stream,
Incremental Clustering,
Continuous Summarization,
Summary.

ABSTRACT

Twitter is a fabulous source for information to keep track of latest happenings and concerns in the world. Whenever something is happening, people around the world start tweeting away. Often they include hash tags, allowing us to selectively search for tweets about a certain event or thing. Many twitter users also engage in conversations, and looking at these conversations allows us to identify leaders and frequent actors. Tweets, in their raw arrangement, while being useful, can also be devastating. For both end-users and data experts, it is a terrible to cultivate complete millions of tweets which contain massive amount of noise and redundancy. This paper includes development of a frame work for extracting representative tweets on a given topic along a time line incrementally to address the problem. This frame work is capable of capturing changing dynamics on the topic in social media and tracks the trend as a series of snap shots. The proposed framework perform Real-Time summarization of selected domain for current trending topics like US President Elections 2016 and Demonetization from twitter stream. It proposes an approach that extracts representative tweets from given domain. It mainly focuses on summarizing current trending topics while identifying outdated topics to be discarded.

Citation: Jonnalagadda Surya Kiran, Dwarapu Suneetha and Prof. Mogalla Shashi. 2016. "Dynamically extracting representative tweets of a domain as they evolve", *International Journal of Current Research*, 8, (12), 43599-43603.

1. INTRODUCTION

Social Network Mining is the process of demonstrating, evaluating, and extracting actionable patterns from social media data. Social Network Mining introduces basic concepts and principal algorithms appropriate for investigating vast social media data; it cultivates a new kind of data scientist who is well versed in social and computational theories, specialized to analyze intractable social media data, and skilled to help associate the gap about the vast social media world with computational tools.Microblogging today has become a very predominant communication tool between Internet users. Masses of posts are found on social networking sites like twitter, Facebook, Tumblr which could be used for advertising and social studies. Study of the user statistics of social networks is one of the existing trends of the times. Big data can be used for group of large and unstructured data which is stimulating while using traditional database.

Corresponding author: Jonnalagadda Surya Kiran,
M.Tech CST with Artificial Intelligence and Robotics, Department of Computer Science and Systems Engineering, Andhra University College of Engineering (A), Visakhapatnam, Andhra Pradesh, India.

Handling the difficulty of big data it is possible to use it for understanding data pattern and use it for learning to envisage. The tweets that are done may be related to diverse topics and the business man want to know more about it. But the enormity of data over twitter confines them from undergoing outline. So the essential of summarization is there which will make available an improved solution. We use data collected from twitter which is in form of messages. The content of communications varies from individual to social views. Natural Language opinions are articulated in undemonstrative and diverse ways, which are challenging to resolve by elementary text handling methodologies. Identifying the sentiment and events correctly is more tedious due its unrestricted message format. Sentiment study is exposure of attitude enduring, affectivity colored beliefs, disposition towards individual or entity. Sentiment comprises of cause i.e. the container of sentiment, objective to whom it is going to affect and type of attitude.

- Tweets are very undersized in length with the message length being about 140 characters. Such a small piece of

text offers very few contextual clues for applying machine learning techniques

- Twitter is a wonderful source for information. Whenever something is trendy, people around the world start tweeting away. Frequently they include hash tags, letting us to selectively search for tweets about a certain event or object. Several twitter handlers moreover involve in discussions, and observing at these discussions allows us to classify leaders and frequent actors.
- Tweets are active in nature and they have a status update minute to minute, if we miss tweets of famous person for certain period for example one week, though we searched with the same time line, the current system in twitter results wanted and unwanted all the tweets are presented, among these many tweets some sensitive and important and topic related tweets searching may take much time and much risk also. Even though we apply filtering criteria it is difficult to identify the required tweet among many tweets because of redundancy and noisy.
- In recent years, researchers have concentrated on problems such as the summarization and detection of topics for twitter messages as well as corpus clustering of tweets
- Several Product companies need this tweet information for making and improving their product based on tweets posted, further all these tweets are used for sentiment analysis. A Data analyst searches and analyses tweets posted by the customer product wise.

In this paper, introduce a novel summarization framework which involves two methods, specifically the Dynamically Clustering of tweets as they evolve and the Summarization i.e., Identification of illustrative tweets. In the first module, design an effective tweet stream clustering algorithm, an online algorithm allowing for active clustering of tweets with only one pass over the data. The second component provisions generation of summaries includes illustrative tweets. To implement continuous or endless tweet stream summarization is however a very tough task, since ample of tweet data are not meaningful, and hence useless in nature. Furthermore, tweets are firmly related with their time that is posted and new tweets arrive at a very faster rate. Finally, a better solution for continuous tweet summarization has to pay heed on these two issues namely:

- Effectiveness
- Elasticity

II. RELATED WORK

Here, we define some of the related existing research work and discuss how our work differs from it. Summarization of tweets is a two-step process. In the first step, tweet data is clustered organized and in the second step actual summarization is done.

A. Microblogging and Twitter

There has been much recent interest on determining and then tracking the evolution of events on Twitter and other social media websites, detecting new events which are also called first stories in the tweet-stream (He *et al*., 2007), visualizing the evolution of tags (Aggarwal *et al*., 2003) and other events on Flickr, YouTube, and Face book (Zhong, 2005). The problem has also been approached from the point of view of efficiency: (Aggarwal and Yu, 2010) propose indexing and compression techniques to speed up event detection without sacrificing detection accuracy. Assume that the event recognition has already been achieved, perhaps using one of the above mentioned techniques; our goal is to collate all the information in the tweets and present a summarized timeline of the event.

B. Stream Data Clustering

Stream data clustering has been extensively studied in the literature. BIRCH (Zhang *et al*., 1996) clusters the data based on an in-memory structure called CF-tree as another of the creative huge data set. Bradley *et al*. (1998) proposed a scalable clustering framework which selectively deliveries important shares of the data, and compresses or discards other shares. CluStream (Aggarwal *et al*., 2003) is one of the best characteristic stream clustering methods. It involves of an online micro-clustering constituent and an offline macro-clustering constituent. The pyramidal time frame was also offered in (Aggarwal *et al*., 2003) to recall historical micro clusters for diverse time durations.

In (Aggarwal and Yu, 2010), the authors extended CluStream to produce duration- based clustering results for text and categorical data streams. However, this algorithm depends on an online phase to produce a huge amount of "micro-clusters" and an offline phase to re-cluster them. In disparity, our tweet stream clustering algorithm is an online procedure deprived of extra offline clustering. And in the framework of tweet summarization, we acquaint the online clustering phase by including the new structure TCV, and restricting the number of clusters to guarantee efficiency and the quality of TCVs.

C. Document/Microblog Summarization

Document summarization can be considered as extractive and abstractive. The earlier selects sentences from the documents, while the later may produce axioms and verdicts that do not appear in the original documents. In this paper, we focus on extractive summarization. While document summarization has been studied for years, microblog summarization is still in its early stages. Sharifi *et al*. proposed the Phrase Reinforcement algorithm to summarize tweet posts exhausting a single tweet (Sharifi *et al*., 2010). Later, Inouye and Kalita proposed a Hybrid TF-IDF algorithm and a Cluster-based algorithm to produce multiple post summaries (Inouye and Kalita, 2011). In (Harabagiu and Hickl, 2011), Harabagiu and Hickl leveraged two relevance models for microblog summarization: an event structure model and a user behavior model. Takamura *et al*. (18) proposed a microblog summarization method based on the median problem, which takes posted time of microblogs into consideration.Unfortunately, almost existing document/ microblog summarization methods mainly deal with small and static data sets, and rarely pay attention to efficiency and evolution issues.

D. LexRank and Continuous LexRank Approaches

LexRank and Continuous LexRank methods which are developed based on variation of the most prevalent page ranking algorithms considered for web link analysis (19). Such ranking models have been effectively demoralized for multi document summarization by making use of the link

relationships between sentences in the document set. A link between two sentences is considered as a vote cast from one sentence to the other sentence.

In sentence mining process all the words in a sentence cannot be preserved as equal prominence, hence we execute necessary preprocessing like removal of stop words and stemming (Porter, 1980). It is also; found from our previous work that IDF would definitely improve the performance of the system (Erkan *et al*., 2004). Equations 1 and 2 give the LexRank and Continuous LexRank for the given document as proposed by Erkan and Radev (Hariharan and Srinivasan, 2009).

$$LexRank[i] = \frac{d}{N} + (1-d) * \sum_{j \in S[i]} \frac{LexRank[j]}{deg[j]} \quad (1)$$

$$Continuous\ LexRank[i] = \frac{d}{N} + (1-d) * \sum_{j \in S[i]} \frac{idf\_modified\_Cosine(i,j) * PR[j]}{\sum_{\kappa \in S[j]} idf\_modified\_Cosine(j,k)} \quad (2)$$

Where N is the aggregate total of sentences in the document, d is the damping factor which is characteristically chosen in the interval (0 to 1), PR (j) indicates the centrality of node j, and S (i) indicates the set of nodes that are adjacent to 'u' and deg (j) is the degree of the node j.A document can be reflected as a network of sentences that are associated to each other. The similarity between the two pairs of sentences x and y is calculated is done after pre-processing. Though there exists several choices of measures to measure the similarity, cosine is superior (Hariharan and Srinivasan, 2008) and is chosen to calculate the importance among the two sentence vectors as reformed by the inverse document frequency given by equation 3:

$$idf\_modified\_Cosine(x,y) = \frac{\sum_{W \in x,y} tf_{w,x} * tf_{w,y} * (idf_w)^2}{\sqrt{\sum_{x_i \in x} (tf_{x_i,x} * idf_{x_i})^2} * \sqrt{\sum_{y_i \in y} (tf_{y_i,y} * idf_{y_i})^2}} \quad (3)$$

Where $tf_{w,x}$ denote the number of occurrences of word 'w' in sentence. A cluster of 'n' sentences in the document can thus be denoted by an n×n symmetric cosine-similarity matrix.

III. METHODOLOGY

A. Clustering of tweets

In this step a minor chunk of tweets are collected. A tweet cluster vector (TCV) is formed which comprises the twee*t al*ong with time stamps. Then by using K-means algorithm initial clusters are formed, the detailed process as described is shown in Fig 1.

1) Incremental Clustering

In this, any tweet that arrives at a time t, is subjected to the MBS algorithm and is decided whether that it is added to a clusters or created. A new cluster. It chooses whether to include t into one of the current cluster or promote t as a new cluster. It finds the cluster whose centroid is closest to t. The modernizing process is executed upon the arrival of each new tweet.

2) Deleting Outdated Clusters and Merging Clusters

The clusters which those are time-bounded and rarely conferred are deleted. An upper limit for the no of clusters as N (max) when the limit is extended, a merging process starts. The most comparable pairs are merged composed. When both clusters are individual clusters which have not been combined with other clusters, they are combined into different composite clusters. Merging Clusters includes aggregate and subtract operations

B. Identification of Representative Tweets

In this step, it provides summaries. Historical summarization offers the useful information to the user. Online summary define what is currently discussed among public. A historical summary helps to recognize the main happening during specific period. The tweet summarization module excludes the tweet that contains undesirable data, outside of that period. In summarization we use Tweet Cluster Lex Rank algorithm for create summaries of tweet clusters for extracting representative tweets, the detailed process as described is shown in Fig 2.

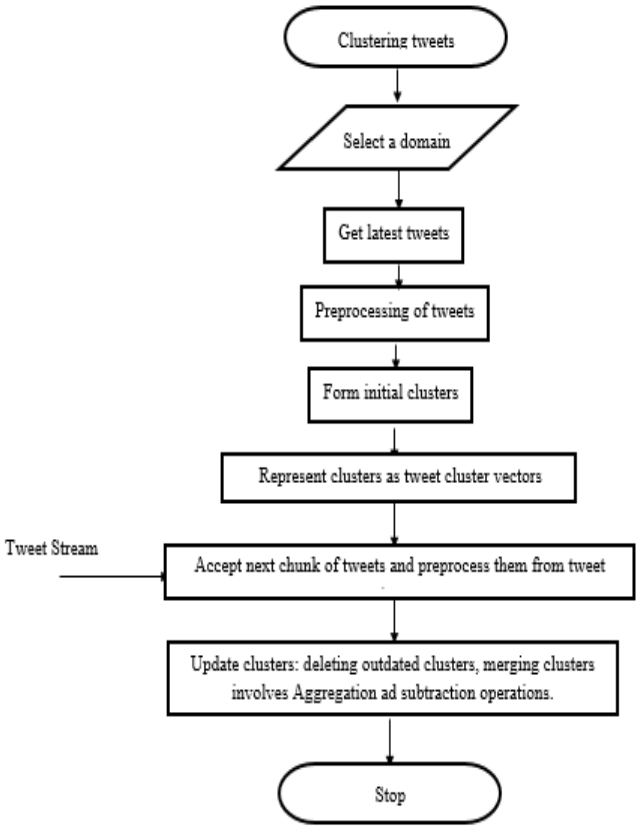

Fig. 1. Clustering Tweets

IV. EXPERIMENTAL SET UP AND RESULTS

(A) DATASETS

For input data, we considered the tweets of current trending topics while identifying outdated topics to be discarded like US President Elections 2016, Demonetization and so on. We crawled about 300 tweets from October 1st, 2016 through December 10th, 2016, from Twitter API using python. The time span totals 111 days, which we split into 2664 time periods (hours). From this set, we generate more than 100

representative tweets within the designated time period. Finally we tabulate the representative tweets along with cluster size and influencing scores.

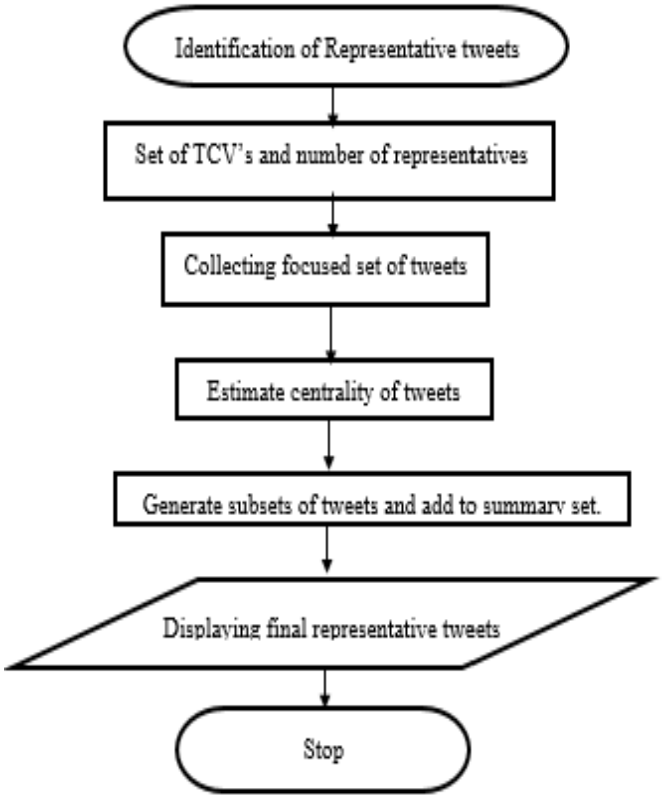

Fig. 2. Identification of Representative tweets

(B) EXPERIMENTS

(1) Clustering of Tweets

Step 1: Selected one of the topics from the list contains US President Elections 2016, Demonetization etc. (considered US President Elections 2016)

Step 2: Specified the time duration from October 1st, 2016 to December 1st 2016

Step 3: Extracted totally 250 tweets after preprocessing like stemming, POS etc.

Step 4: Calculated TF-IDF Scores with respect to each word for each tweet and then we tabulated the details including tweet id, tweet content and TFIDF scores of each word with respect to each tweet

Step 5: Formed initial 20 clusters for small chunk of tweets (i.e., 150) using Means Clustering

Step 6: Created TCV (Tweet Cluster Vector) Information for initial clusters (i.e., 20).

Step 7: Formed Clusters from remaining tweets (i.e., 100) using incremental clustering, from this 5 clusters are created newly

Step 8: Now, Updated TCV (Tweet Cluster Vector) Information for all clusters (i.e., 25).

Step 9: Finally Updated clusters including deleting outdated clusters, merging clusters.

(2) Identification of Representative tweets

Step 1: Considered focused tweets (totally 75) of each cluster into one set T.

Step 2: Generated Cosine Similarity Matrix sim(i)(j) of Size n*n (i.e., 75*75) and Degree vector degree(i) of size n (i.e., 75), for degree calculation we considered t (similarity threshold, 0<t>1) as 0.5 and repeated the same for 0.7 and 0.9 also.

Step 3: Calculated Final Matrix M of size n*n (i.e., 75*75) as sim(i)(j)/ degree(i).

Step 4: Calculated LexRank (LR) Scores for each tweet present ft_set (focus set) of each cluster where LR=PowerMethod (M, n, ε)

Step 5: Consider Set T_C which contains each tweet of ft_set (focus set) with highest LR Score.

Step 6: Added each tweet "t" to Summary for (T-S) and (T_C-S) sets with condition that S(current summary set size)<L(defined summary set size) respectively using below formula, where Summary Set S initialized to null set and L is considered as 50.

$$t = \underset{t_i}{\operatorname{argmax}} \left[\lambda \frac{n_{t_i}}{n_{max}} LR(t_i) - (1-\lambda) \underset{t_j \in S}{\operatorname{avg}} Sim(t_i, t_j) \right]$$

Here considered λ (damping factor or weight parameter, 0≤ λ ≥1) as 0.85 and repeated the same for 0.75 and 0.9 also.

Step7: Finally, tabulated Summary set (contains 35 representative tweets) including cluster size along with influencing score or representative score.

(C) RESULTS

$$\text{Precision} = \frac{\text{No of relevant tweets retrieved}}{\text{No of retrieved (L)}}$$

Where L = defined summary set size

$$\text{Recall} = \frac{\text{No of relevant tweets retrieved}}{\text{No of relevant tweets}}$$

$$\text{F – Score} = \frac{2 * precision * recall}{precision + recall}$$

The Results obtained showed in below table

Table 1: Values of Precision, Recall, F-score with respect to T, λ for given value of L=50

t	λ	S	Precision	Recall	F-score
0.5	0.75	40	0.8	0.762	0.780
0.7	0.85	35	0.7	0.571	0.638
0.9	0.9	30	0.6	0.432	0.504

From obtained results, observed that the size of summary set contains representative tweets increases when λ value decrease

and t value increase. We suggested among three values of λ, t consider λ=0.75 and t=0.5

V. CONCLUSTION AND FUTURE SCOPE

In this, a continuous tweet stream summarization employs an incremental tweet stream clustering algorithm to compress tweets into TCVs and maintains them in an online fashion. Then, it uses a TCV-Rank summarization algorithm for producing online summaries and historical summaries with arbitrary time durations to identify representative tweets. This framework generates representative or most influencing tweets for topics like US President Elections 2016, Demonetization and so on. The framework can be seamlessly extended to a large collection of topics of user's interest.As a future scope the framework can be extended to opinion mining on current topics which can be automatically performed on twitter data related to the topics without explicitly collecting public opinions.

REFERENCES

Aggarwal, C. C. and Yu, P. S. 2010."On clustering massive text and categorical data streams," *Knowl. Inf. Syst*., vol. 24, no. 2, pp. 171–196.

Aggarwal, C. C., Han, J., Wang, J. and Yu, P. S. 2003. "A framework for clustering evolving data streams," in Proc. 29th Int. Conf. Very Large Data Bases, pp. 81–92.

Barzilay, R. and Elhadad, M. 1997. "Using lexical chains for text summarization," in Proc. ACL Workshop Intell. Scalable Text Summarization, pp. 10–17.

Bradley, P. S., Fayyad, U. M. and Reina, C. 1998. "Scaling clustering algorithms to large databases," in Proc. Knowl. Discovery Data Mining, pp. 9–15.

Erkan, G. and Radev, D. 2004. "LexRank: Graph-Based Lexical Centrality as Salience in Text Summarization," *Journal of ArtificialIntelligence Research*, vol. 22, pp. 457-479.

Erkan, G. and Radev,D. R. 2004. "LexRank: Graph-based lexical centrality as salience in text summarization," J. Artif. Int. Res., vol. 22, no. 1, pp. 457–479.

Gong, L., Zeng, J. and Zhang, S. 2011."Text stream clustering algorithm based on adaptive feature selection," Expert Syst. Appl., vol. 38, no. 3, pp. 1393–1399.

Harabagiu, S. M. and Hickl, A. 2011. "Relevance modeling for microblog summarization," in Proc. 5th Int. Conf. Weblogs Social Media, pp. 514–517.

Hariharan, S. and Srinivasan, R. 2009. "Enhancements to Graph Based Approaches for Multi Document Summarizations," *International Journal of Applied Computer Science and Mathematics*, vol. 3, no. 6, pp. 66-72.

He, Q., Chang, K., Lim, E.P. and Zhang, J. 2007. "Bursty feature representation for clustering text streams," in Proc. SIAM Int. Conf. Data Mining, pp. 491–496.

He, Z., Chen, C., Bu, J., Wang, C., Zhang, D. Cai, and X. He, "Document summarization based on data reconstruction," in Proc. 26th AAAI Conf. Artif. Intell, 2012, pp. 620–626.

Inouye, D. and Kalita, J. K. 2011."Comparing twitter summarization algorithms for multiple post summaries," in Proc. IEEE 3rd Int. Conf. Social Comput., pp. 298–306.

Mihalcea, R. and Tarau, P. 2005. "A Language Independent Algorithm for Single and Multiple Document Summarization," *in Proceedings ofInternational Joint Conference on NaturalLanguage Processing*, pp. 1-6.

Page L., Brin S., Motwani R., and Winograd T. 1998. "The PageRank Citation Ranking: Bringing Order to the Web," Technical Report, Stanford InfoLab.

Porter M. 1980. "An Algorithm for Suffix Stripping," Program: Electronic Library and InformationSystems, vol. 14, no. 3, pp.130-137.

Sharifi, B., Hutton, M.A.and Kalita, J. 2010. "Summarizing microblogs automatically," in Proc. Human Lang. Technol. Annu.Conf. North Amer. Chapter Assoc. Comput. Linguistics, pp. 685–688.

Takamura, H., Yokono, H. and Okumura, M. 2011. "Summarizing a document stream," in Proc. 33rd Eur. Conf. Adv. Inf. Retrieval, pp. 177–188.

Wang, D., Li, T., Zhu, S. and Ding, C. 2008. "Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization," in Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, pp. 307–314.

Xu, J., Kalashnikov,D. V. and Mehrotra, S. 2014. "Efficient summarization framework for multi-attribute uncertain data," in Proc. ACM SIGMOD Int. Conf. Manage., 2014, pp. 421–432.

Yih, W.T., Goodman, J., Vanderwende, L. and Suzuki, H. 2007. "Multidocument summarization by maximizing informative contentwords," in Proc. 20th Int. Joint Conf. Artif. Intell., pp. 1776–1782.

Zhang, J., Ghahramani, Z. and Yang, Y. 2004. "A probabilistic model for online document clustering with application to novelty detection," in Proc. Adv. Neural Inf. Process. Syst., pp. 1617–1624.

Zhang, T., Ramakrishnan, R. and Livny, M. 1996. "BIRCH: An efficient data clustering method for very large databases," in Proc. ACM SIGMOD Int. Conf. Manage. Data, pp. 103–114.

Zhong, S. 2005. "Efficient streaming text clustering," Neural Netw., vol. 18, nos. 5/6, pp. 790–798.

****