Time Based Tweet Summarization Using Bisect K-Means Clustering Algorithm

Shubhada Shimpi[1], Sambhaji Sarode[2]
[1, 2] Department of Computer Engineering
[1, 2] MIT, College of Engineering, Savitribai Phule Pune University Pune, India.

Abstract- *Twitter is the most popular micro blogging web site. More than millions of tweets are posted along twitter every day. This tweets includes various types of topics and includes descriptive natire. Also tweets contains huge amount of noisy and redundant data. It is very important to summarize the huge ammount of tweets by reducing the size of tweets and removing the noise, for improving the result accuracy. The operations over flood of tweets is not an easy task. There are so many tweets are unrelated, also arrival rate of tweets is fast. To handle these problems, there is a need of efficient and strong summarization algorithm. This algorithm should be flexible with random time duration. For topic evolution system should detect sub-topic and keeps track for any changes occur with the time. To achieve all these goals, proposed system performs three types of operations on tweets, named as clustering of tweets, summarization and topic evaluation over tweeter data. For clustering of tweets, Bisect K- means algorithm is used and proves that this algorithm is better than K-means algorithm. After this, tweets are summarized with greedy algorithm, which is more accuracy as compared to traditional summarization algorithm. Finally, the topic is detected for generated summary. Experimental results proves that the proposed system summarize the tweets more accurately and efficiently.*

Keywords- Tweet stream, continuous summarization, tweet clustering, summary, timeline

I. INTRODUCTION

Recently social network sites are using so much as we can say it is part of everyone's life. Social network sites are one of the modes of communication for the people all over the world. Any number of people from different part of the world can communicate over the internet. There are so many social networking sites such as Twitter, Facebook etc. There was survey presented by Facebook in 2012, on an average there are 3.2 billon interactions generated over social media which includes likes, comments, post update. Twitter is one of the famous social networking site, it also have huge number of interactions everyday in form of billions of comments, messages. All social media sites are very easy to use and convenient for expressing views on different topics. So popularity of such sites is very high among the people. These days celebrities, organizations, institutes, corporations have their own social pages to interact with people and to teach them as well as for advertisement because of the popularity of social network. Initialization cab be done with single message. User can review, express their feelings on that or also can simply forward it further, even one can like or leave comment on it. As the popularity of such social networking sites is more so number of such messages is very high with high generation rate. When any user wants to refer any certain message of comment, he has to refer them all which is impossible every time and note feasible. It will take lots of time of user in search of particular comment or review. But avoiding this is not possible because users are interested in what other people think about certain topic, or what is their opinion and discussion on certain topic. This is the main motivation of our work to summarize the content and easy to access of required content. There are two techniques can be used for summarization, they are extraction and abstraction. Extraction of summary means identifying relevant sentences among the whole document in short sentences. Abstraction of summary means identifying contents which present as summary and absorb them from whole document. Extraction of summary is the silent information which denotes the document as a whole in form of summary. Words, phrases involves in extraction are different the actual content of the document. Disadvantage for the extraction summary is there is lack of coherence be- tween actual document and summary generated. But extraction summarization is cost effective and easy to apply to any do- main. Abstractive summarization gives more coherency. They produce summary by rewriting and synthesizing actual textual content. Abstractive summarization perform deep analysis and language generation techniques. In our work before actual summarization we perform some preprocessing on data to refine its contents.

Extraction and abstraction there are two approaches for summarization. In extractive summarization method, automatically system retrieves objects from the entire document collection, without changing any of the its already present objects. In extraction method of summarization our main aim is to get different words that the original document which represent the document as a whole. But the meaning of

the sentence remains same, it gives summary in the form of short paragraph. This technique is not only use in textual summarization but also for the image summarization. It retrieves features of the picture without changing the actual picture. Where as abstraction summarization includes the paraphrasing the original docu- ment. By observing results by both the methods, abstraction based summarization can perform consolidation more firmly than extraction based summarization. But implementation of abstraction based summarization is harder because they used technology called natural language generation which is another developing field itself.

Traditional document summarization process are not successful for large size tweets and additionally not reasonably pertinent for tweets which are arrived quick and continuously. To tackle this problem tweet summarization is requires which ought to have new usefulness fundamentally not the same as traditional summarization. Tweet summarization needs to think about the temporal feature of the arriving tweets.

Here are some examples of search engines in which summarization methods are used such as Twitter, Facebook, and Google etc. Other category involve document summarization, image collection summarization and video summarization. The main concept behind summarization is to evaluate a representative and common subset of the data, which exhibit unique data of the entire set. Document summarization, tries to automatically create a representative summary or abstract of the entire document, by finding the most informative sentences. Similarly, in image summarization the system finds the most representative and important (or salient) images. Document summarization technique is used mostly for tweet summarization.

Similarly, in image collection summarization, the system extracts images from the collection without modifying the images themselves. On the other hand, abstraction based summarization task, involves paraphrasing sections of the source document. In general, abstraction can condense a text more strongly than extraction, but the programs which can do this are harder to develop as they require the use of natural language generation technology, which itself is a growing field. Traditional document summarization approach are not effective for big size tweets and not suitably applicable for tweets which are arrived fast and continuously. To overcome this issue tweet summarization is requires which should have new functionality significantly different from traditional summarization. Tweet summarization has to take into consideration the temporal feature of the arriving tweets.

To understand the concept lets see the example of Apple tweets. Summarization algorithm which is designed for the tweet summarization will observe the tweets related to the Apple which are real-time generated on the timeline of the twitter. We can provide certain time range and use it as a document summarization. For the giver time duration, our system will create summary for that document considering the topics and subtopics. Results of such framework will give user output with regarding Apple tweet summarization without even going through all the document content with short amount of time effectively.

Stemming is the term used in linguistic morphology and information retrieval to describe the process for reducing inflected (or sometimes derived) words to their word stem, base or root form generally a written word form. The stem need not be exact to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root. Algorithms for stemming have been researched in computer science since the 1960s. Many search engines treat words with the same stem as synonyms as a kind of query expansion, a process called conflation.

In computing, stop words are words which are filtered out before or after processing of natural language data (text). Though stop words commonly refer to the most common words in a language, there is no single universal list of stop words used by all natural language processing tools, and indeed not all tools even use such a list. Some tools specifically avoid eliminating these stop words to support phrase search. Any group of words can be chosen as the stop words for a given purpose. For some search engines, these are some of the most common, short function words, such as the, is, at, which, and on. In this case, stop words can cause problems when searching for phrases that include them, particularly in names such as The Who, The The, or “Take That”. Other search engines remove some of the most common words including lexical words, such as want from a query in order to enhance performance.

In this paper we study about the related work done, in section II, the proposed approach modules description, mathematical modeling, algorithm and experimental setup in section III .and at final we provide a conclusion in section IV.

II. REVIEW OF LITERATURE

Zhenhua Wang et al. designs a summarization framework called Sumblr. Sumbler is the continuous summarization by stream clustering. Firstly they researched

about continuous tweet stream summarization. This schema consists of three main components, namely the Tweet Stream Clustering mod- ule, the High-level Summarization module and the Timeline Generation module. To work on dynamic, fast arriving, and large-scale tweet streams Sumblr is useful [1].

In paper [2] authors aims to generate digests of tweets from live trending also ongoing topics. The important goal is to group the tweets by significance or usefulness so that an end user can be given a sensible concentrate of the most vital substance from the Twitter stream. Summarization is refined utilizing a non-parametric Bayesian model connected to Hidden Markov Models and a novel observation model intended to permit ranking base.

In paper [3] authors developed a new application, named as sequential summarization for Twitter trending topics. The two proposed methods identify the subtopics and extract significant tweets to create sub-summaries. The evaluations as far as the three estimations, consisting scope, curiosity and relationship and in addition the human evaluation all show that the stream/semantic consolidation ST+SE-PA methodology is the best choice between all the implemented approaches.

In paper [4] authors address the challenges of designing algorithm to group direction stream upon the sliding window model, including variable inspecting rate, information instability, constrained assets, advancing property, and the impact of the obsolete tuples. In perspective of such problem, they have propose a system for trajectory stream clustering, including three sections, the information preprocessing part, the online part that separating summary statistics of trajectory stream segment over sliding window, and the offline part that reclustering micro-clusters depend on such statistical data. furthermore, cluster features can be kept up viably when new trajectory line segments consistently comes in, though the impact of the lapsed records can be expelled securely to keep away from performance degradation with negligible damage to result quality.

In paper [5] authors given solution on a realistic problem of stream mining with activity recognition. The technique consolidates active and incremental learning technique for identifying numbers of activities. They also incorporate super- vised, unsupervised and active learning to assemble a hearty and effective recognition framework. Past methodologies for stream classification did not address this crucial issue. Authors tried given process on genuine datasets and talked about the framework performance contrasted with other classification systems.

Color continues to be an essential topic and the cultural identification plays a significant role in society. This paper [6] research aimed on consolidating known facts related to cultural responses to colors by data-mining social media. To divide the utilization of 11 fundamental color terms in Japanese and German Twitter sustains, word clusters and co-occurrences are studied.

In paper [7] authors given distinct approaches for opinion mining those are aimed on collecting data from twitter on specified topic or keyword. In the wake of gathering informa- tion the information is changed into required format. This data is preprocessed and subjected to find out the opinion mining score utilizing different techniques. Such an analysis would be useful for analyzation. Just a couple of the techniques can achieve to some high level of precision. Hence, the answers for Opinion Mining still have far to go before achieving the certainty level requested by down to specific applications.

In paper [8] authors designed STREAMCUBE to support hierarchical spatio-temporal hash tag clustering, in that, case users can see twitter data interactively with different time and space granularity. It was the first framework to support such application. This system has three components: (1) a spatio-temporal hierarchy influenced by the quad-tree and by data cube. Hashtag clustering is done based on a divide-and- conquer technique at the lowest level of the hierarchy. Then the results of clustering are combined incrementally in a bottom- up manner. (2) A single pass hashtag clustering algorithm. Unique in relation to existing clustering procedures, they are managing content-evolving hashtags. (3) Event ranking, which is intended to help users identify local events and burst events. In paper [9] authors developed simultaneous visualization with a stream graph and relational graph with a spring model for a set of tweets. The test results established the flow and currency of associated topic words, also demonstrated modification in trends in the relational graph. Tweets have data which is temporal which has users trends as well as the relevance of every topic, and customize in group interests. However, they need to inspected singular tweets to comprehend why these phenomena happen or why people are tweeting at a specific time. Contrasting existing examination, the exploration is all the more centering a brief timeframe of specific. The reason that charts have social diagram. So we can see short purpose of connections.

The main goal of [10] is to find out and summarize useful information from the tweets taken at the moment of natural disasters and afterwards, and to provide information sources to aid units. First important tweets selected using classification method then from these important tweets a

subset of tweets which summarizes situation selected as summary. For this, a similarity graph was created by looking at the term and semantic similarities between the tweets. Tweets similar to each other on the graph were clustering in the same cluster. Afterward, the most weighted tweet from each cluster was selected and the summary was created.

The six automatic summarization algorithms are imple- mented in [11], for finding similar Thai tweets. The exper- imental results showed that TextRank algorithm performed the best because this algorithm selected the tweets with the highest scores. On the other hand, Hybrid TF-IDF algorithm could detect similar tweets the least because this algorithm calculated the score by taking the sum frequency of words in a tweet instead of considering the similarity in the level of sentences.

the paper [12] introduce a new tweet summarization approach where the decision of selecting an incoming tweet is made immediately when a tweet is vailable. Tweet selection is based upon three criterion namely informativeness, novelty and relevance with regards of the user's interest which are combined as conjunctive condition. Only tweets having an informativeness and novelty scores above a parametric-free threshold are added to the summary.

The graph-based approach for summarizing tweets is pro- posed in [13], where a graph is first constructed considering the similarity among tweets, and community detection tech- niques are then used on the graph to cluster similar tweets. Finally, a representative tweet is chosen from each cluster to be included into the summary.

A novel framework to identify and summarize tweets is proposed in [14], that are specific to a location. First, pro- pose a weighting scheme called Location Centric Word Co- occurrence (LCWC) that uses the content of the tweets and the network information of the twitters to identify tweets that are location-specific. This paper reports three key findings: (a) top trending tweets from a location are poor descriptors of location-specific tweets, (b) ranking tweets purely based on users' geo-location cannot ascertain the location specificity of tweets, and (c) users' network information plays an important role in determining the location-specific characteristics of the tweets. Finally, we train a topic model based on Latent Dirichlet Allocation (LDA) using a large collection of local news database and tweet-based Urls to predict the topics from the location-specific tweets and present them using an interactive web-based interface.

III. PROPOSED APPROACH

A. Problem Definition

For given real time and historical tweets, apply pre-processing techniques, Bisect K-means is used for incremental cluster formation, ranking for tweet sorting and finally evaluate the topic with timeline and summary generation.

B. Proposed System Overview

Developing continuous tweet stream summarization is a hard task to perform, since countless number of tweets is useless, noisy as well as irrelevant in nature, because of the social way of tweeting. Tweets are firmly associated with their posted time and new tweets have a tendency to touch base at a quick rate. Tweet streams are constantly extensive in scale, henceforth the summarization algorithm ought to be very proficient. It ought to give tweet summaries of subjective time spans. It ought to naturally recognize sub-topic changes and the minutes that they happen. In this paper we are going to build up a multi-point variant of a constant tweet stream summarization system, in particular Sumbler to produce summaries and timelines of events with regards to streams, which will likewise reasonable in distributed frame-works and evaluate it on more finish and extensive scale data sets. The past variant of sumbler was not viable in distributed range.

Proposed system in figure 1 comprises of three principle modules: the tweet stream clustering module, the high-level summarization module and the timeline generation module. The tweet stream clustering module keeps up the online statistical information. The topic-based tweet stream is given; it can proficiently cluster the tweets and keep up minimal cluster data. The high-level summarization module gives two sorts of summaries: online and historical summaries. An online rundown depicts what is as of now talked about among the general population. Hence, the input for creating online sum- maries is recovered straightforwardly from the present clusters kept up in memory. Then again, a historical summary helps people groups comprehend the principle happenings amid a particular period, which implies we have to dispense with the impact of tweet substance from the outside of that period. Therefore, recovery of the required data for creating historical summaries is more confounded. The center of the timeline generation module is a topic evolution detection algorithm which delivers real-time and range timelines also.

The proposed overview system contains the following points:

- We are using and enhancing a continuous tweet stream summarization framework, namely Sumblr, to generate summaries and timelines in the context of streams.

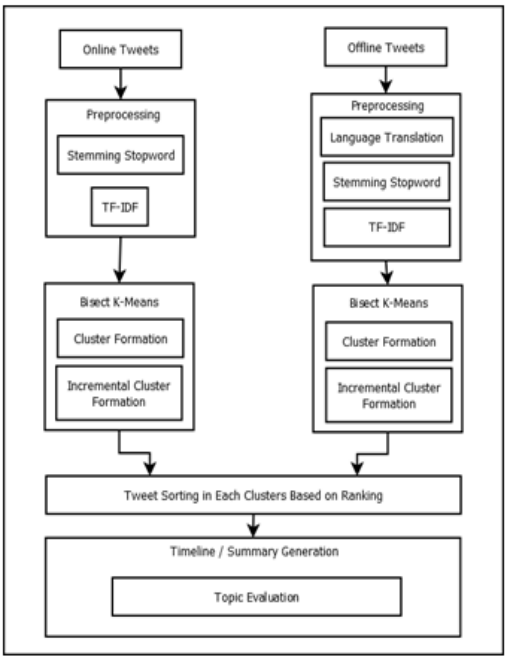

Figure 1. Proposed System Architecture

- Two types inputs are used such as online and offline tweets.
- Offline tweets are translated into English language.
- We are using a novel data structure called TCV for stream processing, and propose the TCV-Rank algorithm for online and historical summarization.
- We are using a topic evolution detection algorithm which produces timelines by monitoring three kinds of variations.
- Extensive experiments on real Twitter data sets demonstrate the efficiency and effectiveness of our framework.
- Produce multi topic summarization.

C. Algorithm

Algorithm 1: Bisecting K-means Clustering

Input: Document Vectors DV Number of Clusters k
Number of iterations of k-means ITER

Output: KClusters

1) Select a cluster to split (split the largest)
2) Find two sub-clusters by using the basic K-means algorithm
3) Repeat step 2
4) the bisecting step is doing for ITER times and take the split process that generate clustering with the highest overall similarity
5) Repeat steps 1, 2 and 3 till the desired number of clusters k are generated.

Algorithm 2: Proposed System Algorithm

Input: Online tweet streams and Historical Tweeter dataset.

Output: Summary generation, timeline generation and topic detection.

1) Read offline dataset
2) Perform language translation, to convert all tweets in english language.
3) Apply preprocessing with stemming stop word removal and TF-IDF computation.
4) Apply bisect K-means for tweet stream clustering
5) Apply TCV rank summarization algorithm for high level summarization with online and historical summaries.
6) Timeline generation with topic detection evolution algorithm.

D. Mathematical Formulation

Term Frequency tf(d) of term t in document d. The number of times that t occurs in d.

Inverse Document Frequency estimate the rarity of a term in the whole document collection

$$\text{idf} = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

Where |D|= Total no: of documents j = no: of documents containing the term ti

$$\text{Cosine Coefficient} = \frac{X \cap Y}{|X|^2 |Y|^2}$$

IV. RESULTS AND DISCUSSION

A. Experimental Setup

The system is built using Java framework(version jdk 8)on Windows platform. The Netbeans (version 8.0.2) is used as a development tool. The system doesn't require any specific hardware to run; any standard machine is capable of running the application.

B. Dataset

The proposed system used tweeter dataset as an input in which content tweet text, user id, share tweet and like tweet. The system used tweeter API file to extract dataset and filter the data by applying preprocessing method.

C. Evaluation Results

Table 13.1 depicts the comparison of existing and proposed system on the basis of time efficiency. Proposed system with bisect k-means is more efficient than existing system with K-means, to find out the social coordinates. Bisect k-means identify the social coordinates that is attributes of users, very fastly.

Following figure 13.1 depicts the time efficiency comparison graph of the proposed system with the existing system. Time required to identify sequential patterns in existing system by using apriori is more than the time required for proposed system with FP-Growth algorithm.

Table 13.2 depicts the accuracy in % of recommended friend list. It is clearly shown that the proposed system is more accurate than existing system. Because MLP more accurately identify the attributes of all users, which is very important to find out the relevant friends.

Table 1. TIME COMPARISON

System	Time Required
Existing system using K-Means	78000 ms
Proposed system using Bisect K-Means	30329 ms

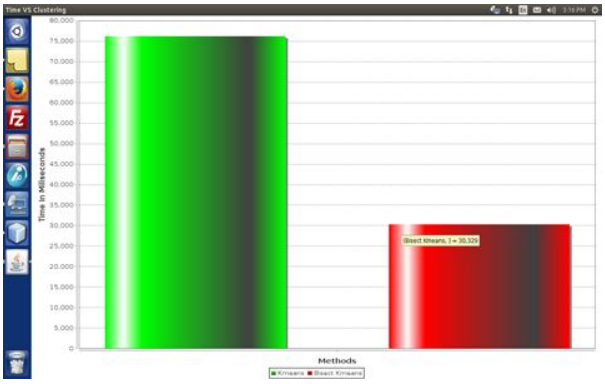

Figure 2. Time Efficiency Comparison

Table 2. MEMORY COMPARISON

System	Memory Required
Existing system using K-Means	18425000 kb
Proposed system using Bisect K-Means	9154628 kb

Following figure 2 shows the memory comparison of the proposed system with the existing system. Proposed system is more accurate. X-axis represent the system names and Y-axis represent the accuracy if friend recommendation in %.

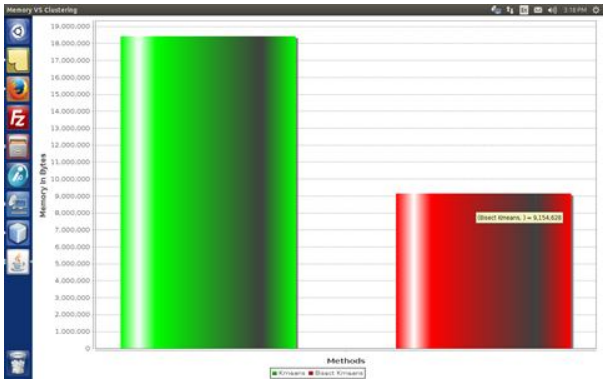

Figure 3. Memory Comparison

Table 13.3 shows the accuracy of the proposed system and existing system. The following table shows the recall value of existing system is less than the proposed system.

Table 3. RECALL COMPARISON

System	Recall Value
Existing system using K-Means	58.5%
Proposed system using Bisect K-Means	60%

Following figure 13.3 shows the accuracy comparison graph of the proposed system with the existing system. Recall by the proposed system is more than the memory required for existing system. As the bisect k-means has maximum number of itterations, accuracy is increases in proposed system.

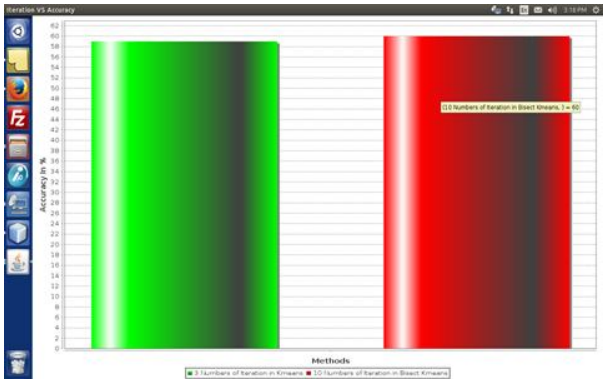

Figure 4. Accuracy Comparison Graph

The figure 13.4 shows the time required for execution of number of iterations in existing and proposed system. In k-

means , for three number of iterations, 76000 miliseconds are required and for bisect k-means algorithm, 30000 miliseconds are required to execute 10 number of iterations.

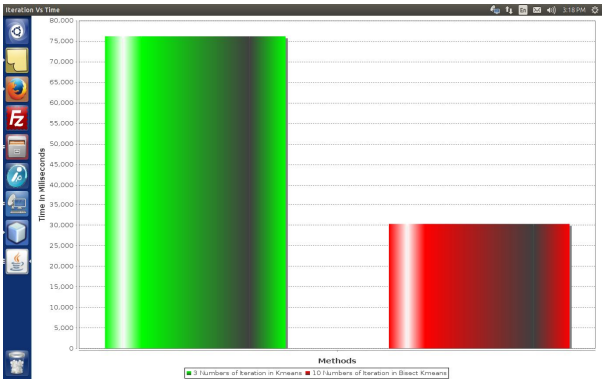

Figure 5. Time comparison for number of itterations

Figure 13.5 shows the number of clusters formed for K- means and bisect k-means algorithm. For 103 tweets 5 and 3 clusters are formed in K-means and Bisect K-means clustering approach.

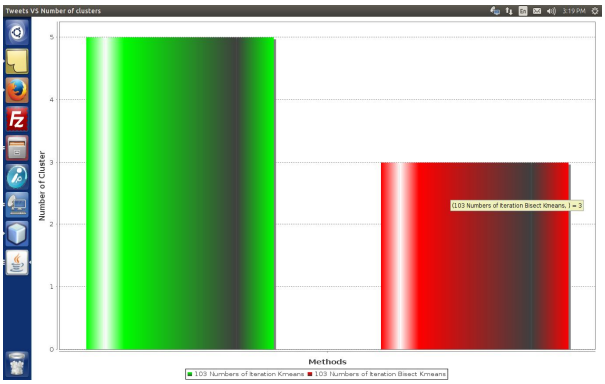

Figure 6. Number of iterations

V. CONCLUSION

In this paper we studied various problems related to tweet data. This data is affected by the noise and redundancy data, which affect the performance of the tweet summarization algorithm. We have studied various document summarization techniques such as filtering, tweet summarization etc. To avoid the problems and improve the performance there is a need of dynamic methodology to summarize the tweet feeds. The proposed algorithm is named as multi topic summarization and it makes use of online and offline tweet streams as an input. This paper proves that the bisect k means algorithm accuracy and efficiency of proposed system.

REFERENCES

[1] Zhenhua Wang, Lidan Shou, Ke Chen, "On Summarization and Timeline Generation for Evolutionary Tweet Streams, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 5, MAY 2015.

[2] D. Wen and G. Marshall, "Automatic Twitter Topic Summarization," Computational Science and Engineering (CSE), 2014 IEEE 17th Inter- national Conference on, Chengdu, 2014, pp. 207-212.

[3] D. Gao, W. Li, X. Cai, R. Zhang and Y. Ouyang, "Sequential Summariza- tion: A Full View of Twitter Trending Topics," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 2, pp. 293-302, Feb. 2014.

[4] J. Mao, C. Jin, X. Wang and A. Zhou, "Challenges and Issues in Trajectory Streams Clustering upon a Sliding-Window Model," 2015 12th Web Information System and Application Conference (WISA), Jinan, 2015, pp. 303-308

[5] Z. S. Abdallah, M. M. Gaber, B. Srinivasan and S. Krishnaswamy, "StreamAR: Incremental and Active Learning with Evolving Sensory Data for Activity Recognition," 2012 IEEE 24th International Conference on Tools with Artificial Intelligence, Athens, 2012, pp. 1163-1170.

[6] D. M. Marutschke, S. Krysanova and H. Ogawa, "Clustering Word Co- occurrences with Color Keywords Based on Twitter Feeds in Japanese and German Culture," 2015 International Conference on Culture and Computing (Culture Computing), Kyoto, 2015, pp. 191-192.

[7] V. Sindhura and Y. Sandeep, "Medical data Opinion retrieval on Twitter streaming data," Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on, Coimbatore, 2015, pp. 1-6.

[8] W. Feng et al., "STREAMCUBE: Hierarchical spatio-temporal hashtag clustering for event exploration over the Twitter stream," 2015 IEEE 31st International Conference on Data Engineering, Seoul, 2015, pp. 1561- 1572.

[9] K. Amma, S. Wada, K. Nakayama, Y. Akamatsu, Y. Yaguchi and K. Naruse, "Visualization of spread of topic words on Twitter using stream graphs and relational graphs," Soft Computing and Intelligent Systems (SCIS), 2014 Joint 7th International Conference on and Advanced In- telligent Systems (ISIS), 15th International

Symposium on, Kitakyushu, 2014, pp. 761-764.

[10] I. Hseynli and M. E. Karsligil, ”Determination and summarization of important tweets after natural disasters,” 2017 25th Signal Processing and Communications Applications Conference (SIU), Antalya, Turkey, 2017, pp. 1-4.

[11] P. Boonchaisuk and K. R. Saikaew, ”Efficient algorithms for Thai tweet summarization,” 2016 International Computer Science and Engineering Conference (ICSEC), Chiang Mai, 2016, pp. 1-5.

[12] [A. Chellal, M. Boughanem and B. Dousset, ”Multi-criterion Real Time Tweet Summarization Based upon Adaptive Threshold,” 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Omaha, NE, 2016, pp. 264-271.

[13] S. Dutta, S. Ghatak, M. Roy, S. Ghosh and A. K. Das, ”A graph based clustering technique for tweet summarization,” 2015 4th Interna- tional Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Noida, 2015, pp. 1-6.

[14] V. Rakesh, C. K. Reddy, D. Singh and M. Ramachandran, ”Location- specific tweet detection and topic summarization in Twitter,” 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013), Niagara Falls, ON, 2013, pp. 1441-1444.