© 2018 IJSRSET | Volume 4 | Issue 8 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Summarization Method and Timeline Generation of the Tweet

Pooja Patil, Nilamvhatte, Srushti Rajput, Ujjwala Panhalkar, K. V. Deshpande

Department of Computer Engineering, RSCOE, Tathawade, Savitribai Phule Pune University, Pune, Maharashtra, India

ABSTRACT

Twitter is the most popular micro blogging web site. More than millions of tweets are posted along twitter every day. Tweets contains huge amount of noisy and redundant data. It is very important to summarize the huge amount of tweets by reducing the size of tweets and removing the noise, for improving the result accuracy. The operations over flood of tweets are 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. Framework has component is data duplication checking using SHA1 hashing strategy. Framework used clustering procedure it uses EM clustering and compare the EM clustering algorithm with K-means clustering 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 every day 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 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 avoidingthis 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 coherencies. They produce summary by rewriting and synthesizing actual textual content. Abstractive summarization performs 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 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. Whereas abstraction summarization includes the paraphrasing the original document. 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. 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 involves 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. To understand the concept let's see the example of Apple tweets. Summarization algorithm which is designed for the tweet summarization will observe the tweets related to the Apple which is 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.

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

In paper [1], Zhenhua Wang et al. proposed a technique for 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 known as Tweet Stream Clustering module, High-level Summarization module and Timeline Generation module.

In paper [2] authors proposed a system for generate digests of tweets from live trending also ongoing topics. Goal of system 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.

In paper [3] authors proposed a techniques named as Sequential Summarization for Twitter trending topics. These two methods identify the subtopics and extract significant tweets to create sub-summaries.

In paper [4] 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 supervised, unsupervised and active learning to assemble a hearty and effective recognition framework.

In paper [6] they proposed a Color continues to be an essential topic and the cultural identification plays a significant role in society. Research aimed onconsolidating known facts related to cultural responses to colors by data-mining social media.

In paper [7] authors proposed distinct approaches for opinion mining those are aimed on collecting data from twitter on specified topic or keyword. In the wake of gathering information the information is changed into required format. This data is preprocessed and subjected to find out the opinion mining score utilizing different techniques.

The main goal of [8] 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 implemented in [9], for finding similar Thai tweets. The experimental results showed that Text Rank 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.

III. SYSTEM ARCHITECTURE

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

Develop 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 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 thesummarization 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 frameworks 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 summaries 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.

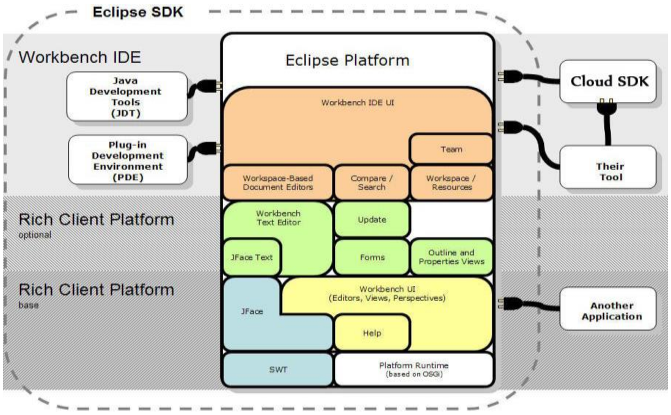

Fig. 1. Proposed System Architecture

- 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. Mathematical Model

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

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

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

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

$$Cosien\ Coefficient = \frac{X \cap Y}{|X|^{\frac{1}{2}}|Y|^{\frac{1}{2}}}$$

D. Algorithm Used

Algorithm 1: Bisecting K-means Clustering

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

IV. RESULT AND DISSCUSIION

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. Thesystem doesn't require any specific hardware to run; any standard machine is capable of running the application.

B. Data Set:

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 I 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 2 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 I
TIME COMPARISON

System	Time Required
Existing system	28000ms
Proposed system	13000 ms

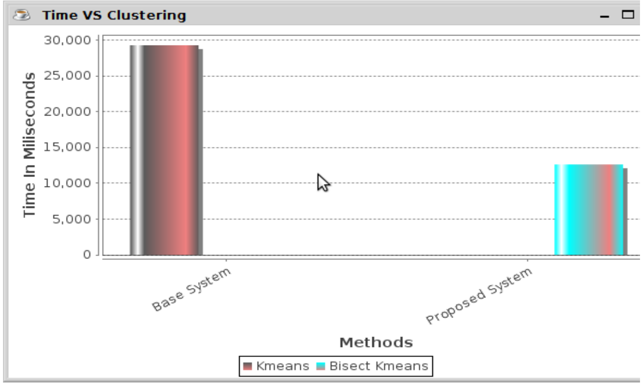

Fig. 2.Time ComparisionGraph

Table II 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 accurate by identify the attributes of all users, which is very important to find out the relevant friends.

TABLE II. MEMORY COMPARISON

System	Memory Required
Existing system	13425000 kb
Proposed system	61000000 kb

Following figure 3 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 %.

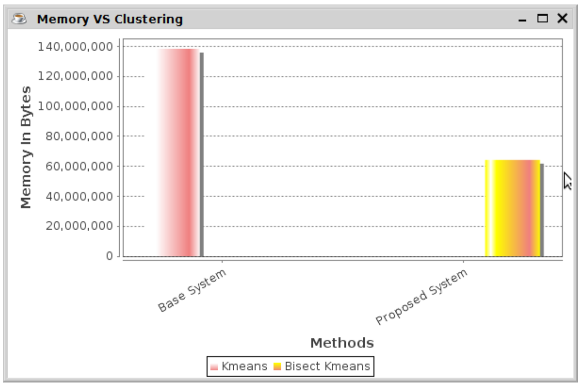

Fig. 3. Memory Graph

Table III 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 II

ACCURACY COMPARISON

System	Accuracy
Existing system	56%
Proposed system	59 %

Following figure 4 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 iterations, accuracy is increases in proposed system.

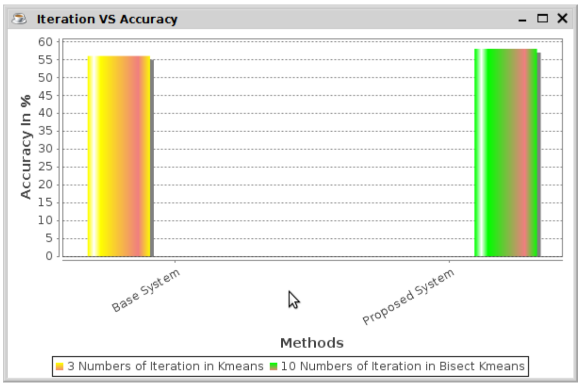

Fig. 4. Accuracy Comparison Graph

The figure 5 shows the time required for execution of number of iterations in existing and proposed system. In k- means, for three numbers of iterations, 29000miliseconds are required and for bisect k-means algorithm, 140000 miliseconds are required to execute 10 numbers of iterations.

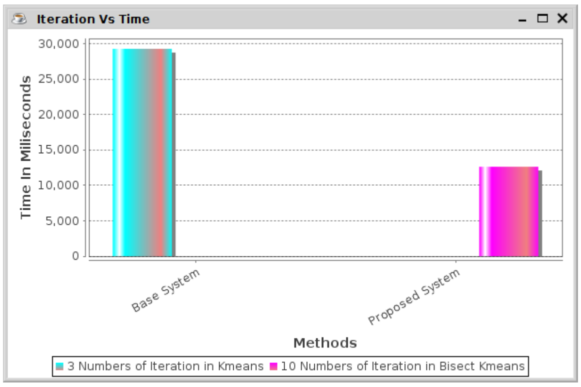

Fig. 5. Time comparison for number of itterations

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

VI. REFERENCES

[1] Zhenhua Wang and Ke Chen, "on summarization and timeline generation for evolutionary Tweet Streams", IEEE Transaction [2015].

[2] D. Wen, G. Marshall, "Automatic twitter topic summarization" Computational Science and Engineering (CSE), IEEE 17th Inter- national Conference on, Chengdu, [2014].

[3] D. Gao, R. Zhang and Y. Ouyang, "Sequential summarization: a full view of twitter trending topics," IEEE/ACM Transactions on Audio, Speech, and Language Processing, Feb. 2014.

[4] J. Mao, X. Wang and A. Zhou, "Challenges and issues in trajectory streams clustering upon a Sliding-Window Model", 12th Web Information System and Application Conference (WISA), [2015].

[5] M. M. Gaber, B. and S. Krishnaswamy, "Stream AR incremental and active learning with evolving sensory data for activity recognition," IEEE 24th International Conference on Tools with Artificial Intelligence, Athens, [2012]

[6] S. Krysanova, D. M. Marutschke, and H. Ogawa, "Clustering word co- occurrences with color keywords based on twitter feeds in Japanese and German culture," International Conference on Culture and Computing (Culture Computing), Kyoto, 2015, pp. 191-192, [2015].

[7] V. Sindhura , Y. Sandeep, "Medical data opinion retrieval on Twitter streaming data," IEEE International Conference on, Coimbatore, 2015, pp. 1-6 [2015].

[8] W. Feng et al., "STREAMCUBE: Hierarchical Spatio Temporal Hashtag Clustering for Event Exploration Over The Twitter stream," IEEE 31st International Conference on Data Engineering, Seoul, 2015.

[9] 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), 15th International Symposium on, Kitakyushu, 2014, pp. 761-764[2014].