International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 6, Issue 9, September 2016)

Summarization and Timeline Generation for Social Stream with Categorization.

Snehal Chavan[1], Sanchayati Thorat[2], Tejaswini Surade[3], Priyanka Ghongade[4]
[1, 2,3,4]*Department of Computer Engineering, SND College of Engineering & RC, Yeola, India*

***Abstract*— as increasing popularity of social sites like Tweeter, Facebook and Instagram etc. We will get lot of tweets and 'N' no. of short messages being shared at unpredictable rate which is very high. As this data is large enough it will become critical to understand and analyze therefore redundancy and noisy data must be removed. To overcome these drawbacks of existing system we propose sumblr framework, in comparison with other regular approaches of summarization which depends on static data and small datasets where sumblr is dynamic and works on large data set. Firstly we have proposed tweet cluster vector algorithm for maintaining statistical data and compact cluster information to maintain dynamically in memory during stream processing, store and organize cluster snapshots of different moments. Generation of online and historical summaries with arbitrary time durations, we propose TCV rank summarization algorithm. We have proposed an evaluation method which generates timeline, categorization based on topic evaluation.**

***Keywords*—Tweet stream, continuous summarization, summary, timeline**

I. INTRODUCTION

Now a day a socially generated stream has become popular on WWW (World Wide Web). As rapid growth in an internet, use of social media also increases. There are many social sites like Twitter, Facebook, Instagram etc. in which twitter has become one of the most popular social site for users to share information like text, audio, video etc. Short messages are being created and shared at massive rate. Twitter receives thousands of tweets per hour. It is in raw form, the solution for this is summarization of tweets. Summarization represents a set of document which contain summary of related data. We have proposed Tweet Cluster Vector (TCV) algorithm which is used for making cluster of those retrieved tweets among which summarization will takes place. Tweet Cluster Vector (TCV) algorithm includes two data structure to keep important tweet information in cluster. These data structures are tweet cluster vector and pyramidal time frame. TCVs are considered as potential sub topic representative and maintained dynamically during stream processing in memory.

The second data structure pyramidal time frame which used for storing and organizing cluster snapshot. So historical and online tweets data extracted by any random time duration which will give more relevant in results. In the summarization we will adding category such as news, politics, entertainment etc. we can summarize the tweets as per category. In the tweet summarization many tweets are repeated so using summarization we can avoid redundancy. The summarization consists of four issues efficiency, flexibility, topic evolution, performance. Tweet streams or many messages of social site are very large in size so the summarization algorithm is very efficient. Performance of summarization is very effective. We have proposed TCV rank summarization algorithm which is used for generating historical and online summaries. This algorithm selects the top rank tweets from the Tweet Cluster Vector (TCV), to generate historical and online summaries where user specifies random time duration. We retrieve cluster snapshots from the Pyramidal Time Frame (PTF) with respect to beginning and ending of time duration, based on two clusters TCV rank summarization algorithm generates summaries. Also we proposed Topic Evolution Detection algorithm which takes the input of already generated summaries to produce timeline. Also we are working on other social stream which include clustering, timeline generation, Topic evolution etc.

II. LITERATURE SURVEY

We have studied the paper **"A framework for clustering evolving data streams" (**C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu**)** in which TCVs are considered as potential sub-topic; for stream clustering, Clustream method is used. It includes online and offline micro clustering component. For recalling historical micro cluster, pyramidal time frame also proposed for random time duration. [1]

For using function lexrank in TCV rank algorithm we have studied **"LexRank: Graph based lexical centrality as salience in text summarization" (**G. Erkan and D. R. Radev**)** in this paper lex ranking is calculated. Depending on the similar data graph is created; Lexrank is used for finding top ranked tweets among large data set. [2]

Also we referred, **"Text stream clustering based on adaptive feature selection"** (L. Gong, J. Zeng, and S. Zhang) worked on a various services on the Web such as

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 6, Issue 9, September 2016)

news filtering, text crawling, etc. It mainly focuses on topic detection and tracking (TDT). Clustering is used for analyzing text stream. [3]

Again we have studied paper **"Evolutionary timeline summarization A balanced optimization framework via iterative substitution"** (R. Yan, X. Wan, J. Otterbacher, L. Kong, X. Li, and Y.Zhang) evolutionary timeline summarization which consist of time stamped summaries which is used to generate timeline dynamically during the process of continuous summarization (Sumblr).[4]

For summarization we have studied **"Summarizing sporting events using twitter" (**J. Nichols, J. Mahmud, and C. Drews**)** in which Summarization algorithm creates sentence level summaries of important moments and then concatenated to generate an event summary of paragraphs. [5]

Lastly we have referred **"on summarization and timeline generation for evolutionary tweet stream"** we have referred Tweet Cluster Vector (TCV), TCV Rank algorithm, Topic evolution. In which TCV used for making effective clustering of tweet with the help of pyramidal time frame and tweet cluster vector, TCV rank summarization algorithm is used for generating online and historical summaries by evaluating top ranked function, depending upon top ranked tweets summarization is done. Topic evolution detection generates timeline by considering large variation of sub-topics in stream processing. [6]

III. PROPOSED SYSTEM

The system architecture is for "historical and online summarization of social stream". In today's world, summarization becomes necessity of social stream as millions of information posted on social sites. It is the simplest way to understand exact information using summarization by avoiding redundancy and noisy data.

Fig 3.1 mainly focuses on three module Stream Clustering, Summarization, and Timeline Generation. Here Categorization also done on summary generated.

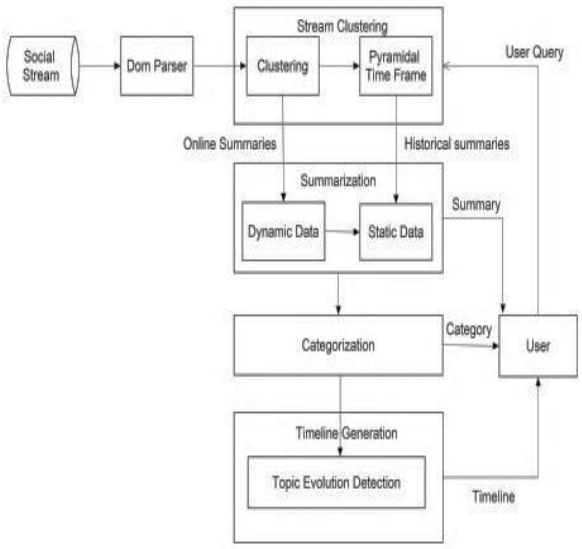

Fig 3.1 System Architecture

1. *Stream Clustering*

For making the cluster of social stream we use the clustering algorithm, by checking relevant data we are making cluster. [1]

1.1 Pyramidal Time Frame

Pyramidal time frame include time frame. Here user can provide the two points that is starting and ending time.

2. Summarization

After apply clustering on social data set, summarization is done by using TCV- Rank summarization algorithm. Static as well as dynamic data is summarized. [6]

3. *Categorization*

From generated summary, categorization is done like summary is relates to entertainment, sports, politics etc.

4. *Timeline Generation*

The base of the timeline generation is topic evolution detection algorithm which uses online summaries and generates timeline. Topic evolution describes changes in subtopics by monitoring variation in stream clustering. [6]

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 6, Issue 9, September 2016)

IV. ALGORITHM

A. Clustering algorithm:

Input: Cluster set

Output: Assigning cluster for new tweets

Step I- Collection of new tweet stream

Step II- Depending upon two attribute it create new cluster

1)$maxsim(maximum\ similarity)$

2)$MBS(minimum\ bounding\ similarity)$

Step III- If $maxsim$ is less than MBS then it creates new cluster.

Step IV-otherwise update new cluster

B. Summarization algorithm:

Input: Cluster Set

Output: Summarization according to rank

Step I- Building similarity graph for all tweet

Step II-Computing LexRank to know which tweets are top ranked

Step III-Adding tweets into summary according to equation

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

Step IV- Checking summary length till it reached to max size. Selecting tweet globally ($t_i \epsilon\ T - s$) based on above equation.

C. Topic Evolution Detection algorithm:

Input: A tweet stream binned by time units

Output: A timeline node set TN

Step I-Binning tweets by time.

Step II- Appending new timeline nodes whenever large variation detected. By using

$$while\ !\ stream.end()\ do$$
$$Bin\ c_i = stream.next()$$
$$if\ hasLargeVariation()Then$$
$$TN.add(i);$$

D. Web Extraction Algorithm

Step I- Recognizing peer node

Step II- It align nodes in peer matrix to get a list of aligned nodes

$$childList = matrixAlignment(M)$$

Step III- Repetition of pattern detected starting with length 1

$$childList = repeatMining(childList, 1)$$

Step IV- Optimal merging

V. CONCLUSION

We proposed a Sumblr which supports continuous tweet stream summarization. Sumblr uses a tweet stream clustering algorithm for compress tweets into TCV and manages them in an online way. Then, it uses a TCV (tweet cluster vector)-Rank summarization algorithm for generating online and historical summaries with random time durations. Also categorization will be done on summarized data. The topic evolution will be done automatically, permitting Sumblr to create dynamic timelines for tweet streams. We are also working same for other social site also.

REFERENCES

[1] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. "A framework for clustering evolving data streams." ACMSIGMOD Conference (2003): 81-92.

[2] Radev, G. Erkan and D. R. "LexRank: Graph-based lexical centrality as salience in text summarization." J. Artif. Int. Res. 22 (2004): 457–479.

[3] L. Gong, J. Zeng, and S. Zhang. "Text stream clustering algorithm based on adaptive feature selection." Expert Syst. Appl. 38 (2011): 1393–1399.

[4] R. Yan, X. Wan, J. Otterbacher, L. Kong, X. Li, and Y. Zhang,. "Evolutionary timeline summarization: A balanced optimization framework via iterative substitution." 34th Int. ACM SIGIR Conf. Res. 2011. 745–754.

[5] J. Nichols, J. Mahmud, and C. Drews,. "Summarizing sporting events using twitter." ACM Int. Conf. Intell (2012): 189–198.

[6] Zhenhua Wang, Lidan Shou, Ke Chen, Gang Chen and Sharad Mehrotra. "on summarization and timeline generation for evolutionary tweet stream." IEEE 27 (2015): 1301-1315.