Survey On Data Stream Clustering And Summarization
Dhanshri A. Nevase*1,Prof. Amrit Priyadarshi*2
*1(PG Student, DGOI, Pune, Department of Computer Engineering, Savitribai Phule Pune University, Maharashtra, India)
*2(Asst. Professor DGOI, Pune, Department of Computer Engineering, Savitribai Phule Pune University, Maharashtra, India)
dhanashri.nevase@gmail.com*1

Abstract - With the explosive growth of micro blogging services, short-text messages (also known as tweets) are being created and shared at an unpredictable rate. Tweets are present in raw form and tweets gives information but it is large in amount. That tweets are used by both end-users and data analyst, but walking through that lots of tweets which have noise & redundancy is impossible. To avoid this problem continuous tweet stream generation is used. The previous methods focus on selecting several posts from given sets manually and consider small as well as static data.

In this paper fast arriving, dynamic and big scale tweet streams are considered. Use Sumblr algorithm (SUMmarization By stream CLusteRing) for tweet streams. Here, present two methods for summarization. First algorithm cluster the tweets that are generated & maintain these statistics in a data structure called as tweet cluster vector. Second algorithm then rank that cluster for generating dynamic summaries & summaries that based on fixed any time duration. Our methods gives efficiency & effectiveness on large scale data sets.

Index Terms—Dynamic summarization, summary, tweet stream.

1. INTRODUCTION:

Now a day's people use Twitter, face book such a social site to post streams. So, it receives lots of tweets per day from various sources as news, opinions, debates, blogs & so on. Tweets are present in raw form & it contains duplication & unnecessary data. It’s difficult to generate Generating summaries on that tweets so filtering is required. But filtering method must apply continuously on every tweet stream but that tweet streams occur at an undecidedly rate. Information overload problem is occur. Solution for information overload is summarization. Summarization is nothing but minimizing the size of document such that document must maintain important points of original document. In tweet summarization, tweets occur in fast & continuously manner. So, tweet summarization consider temporal feature of arriving tweets.

In temporal feature two methods drill down & roll up are used. In drill down method summaries are depends on short period of time (e.g. summary between 11am to 11pm on 29th August). Opposite of that in roll up method summaries contains larger period of time (e.g. Summary between 15th August to 30 August).

In continuous tweet stream summarization three important concepts are considered as efficiency, flexibility & topic detection. But this is not an easy task because of arriving tweet stream in continuous & unwanted manner.

In existing system they consider static & small scale data. Second they perform iterative summarization for every possible given time duration. And third the result of summaries is not depending on temporal feature.

In current system we introduce a novel summarization framework called Sumblr (continuous summarization By stream clustering). The framework consists of three main components, as clustering of tweet streams module, document Summarization module and the Timeline Generation module. In the clustering of tweet streams module, we design an efficient tweet stream clustering algorithm, an online algorithm allowing for effective clustering of tweets with only one pass over the data. The document summarization module supports two kinds of summaries as online and historical summaries.
The system consists of two main functions as follows:
1. Generate summaries & timelines based on tweet streams.
2. Design a novel data structure called as Tweet Cluster

2. LITERATURE SURVEY ON DATA STREAM CLUSTERING AND SUMMARIZATION:

2.1 Data Stream Clustering:
Data stream clustering is defined as the clustering of data that arrive continuously such as telephone records, multimedia data, financial transactions, etc. Data stream clustering is usually studied as a streaming algorithm and the objective is, given a sequence of points, to construct a good clustering of the stream, using a small amount of memory and time. Lots of algorithms are used for clustering data stream.

- **Title**: Clustering data stream [2].
**Author**: O'Callaghan, Nina Mishra, Adam Meyerson, Sudipto Guha, Rajeev Motwan.
This paper introduces the STREAM algorithm to show that clustering can take place in small space. Small-Space is a divide-and-conquer algorithm that divides the data, S, into some pieces, clusters each one of them, and then clusters the centers obtained. The STREAM algorithm performs a bottom-up approach.

But disadvantage of this algorithm is streams are scans recursively number of times.

- **Title**: An efficient data clustering method for very large databases BIRCH [3].
**Author**: T. Zhang, R. Ramakrishnan, and M. Livny.
In this paper, we build a hierarchical data structure to incrementally cluster the incoming points using the available memory and minimizing the amount of I/O required.

It is local in that each clustering decision is made without scanning all data points and currently existing clusters. It exploits the observation that data space is not usually uniformly occupied and not every data point is equally important. It makes full use of available memory to derive the finest possible sub-clusters while minimizing I/O costs. It is also an incremental method that does not require the whole data set in advance. The BIRCH algorithm performs a top-down approach.

- **Title**: Iterative Optimization and Simplification of Hierarchical Clusterings [4].
**Author**: Fisher, D. H.
COWEB is an incremental clustering technique that keeps a hierarchical clustering model in the form of a classification tree. But this algorithm does not provide cost and quality benefits to applying optimization strategies at inter-mittent points during hierarchical sorting.

- **Title**: Incremental Clustering for Dynamic Information Processing [5].
**Author**: Fazli Can.
In this paper, C2ICM builds a flat partitioning clustering structure by selecting some objects as cluster seeds/initiators and a non-seed is assigned to the seed that provides the highest coverage.

The algorithm creates an effective and efficient retrieval environment, and it is cost effective with respect to re-clustering and can be used for many increments of various sizes.
• Title: A framework for clustering evolving data streams [6].
Author: C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu.
This paper gives CluStream algorithm. CluStream algorithm is divided into a statistical data collection component and an on-line analytical component based on a pyramidal time window. All these techniques fail to provide effective clustering.

• Title: On clustering massive text and categorical data streams [7].
Author: C. C. Aggarwal and P. S. Yu.
In categorical data stream, it provides an online analytical processing approach to stream clustering. The proposed algorithms is based on text and categorical data stream clustering. We will propose a condensation based approach for stream clustering which summarizes the stream into a number of fine grained cluster droplets.

• Title: On Summarization and Timeline Generation for Evolutionary Tweet Streams [1].
Author: Zhenhua Wang, Lidan Shou, Ke Chen, Gang Chen, and Sharad Mehrotra.
In this paper the tweet stream algorithm is an online process without requiring any offline clustering. TCV data structure is used for this clustering algorithm which keeps information of tweet cluster.

2.2 Document Summarization
Document summarization is classified as extractive & abstractive summarization. Abstraction involves paraphrasing sections of the source document. Extractive performs the automatic system extracts objects from the entire collection, without modifying the objects themselves.

• Title: Multi-Document Summarization by Maximizing Informative Content-Words [8]
Author: W.-T. Yih, J. Goodman, L. Vanderwende, and H. Suzuki.
In this paper two components are used for multi-document summarization. First component generate score for each word in the set of document using machine learning. Second component find set of sentences from document cluster for maximize the scores.

• Title: Document summarization based on data reconstruction [9].
Author: Z. He, C. Chen, J. Bu, C. Wang, L. Zhang, D. Cai, and X. He.
This paper gives data reconstruction using document summarization based on data reconstruction (DSDR) method. They use two approaches as linear construction & non negative linear construction. In linear construction take document by linear combination of selected sentences. In non negative linear reconstruction only additive but not subtractive linear combination is used.

• Title: A participant-based approach for event summarization using twitter streams [10].
Author: C. Shen, F. Liu, F. Weng, and T. Li.
In this paper different types of events are considered. Participant take a part into event. So this paper give summarization based on participant & event. They generate a textual description of the scheduled events that are reported on Twitter.

• Title: On Summarization and Timeline Generation for Evolutionary Tweet Streams [1].
Author: Zhenhua Wang, Lidan Shou, Ke Chen, Gang Chen, and Sharad Mehrotra.

This paper is based on online summarization & historical summarization. To generate online summaries, take topic related tweet stream without any previous knowledge. Store all the tweets in each segment & select only one tweet as part of summary to reduce space as well as computation cost. To generate historical summaries maintain TCV snapshots.

3. CONCLUSION:
In this paper we have studied many different technique for clustering data stream & summarization. Existing technique provide off-line as well as online clustering. Generate cluster depends on different time durations.

For summarization most of the existing method use static & dynamic data to generate summaries.

4. REFERENCES:


