Summarizing the Evolutionary Tweet Streams with Timeline Generation

P. Siva Prasad¹, M.V. Jayasree²

¹PG Scholar, Dept of CSE, Golden Valley Integrated Campus(Affiliated to JNTU Anantapur), AP, India.
²Assistant Professor, Dept of CSE, Golden Valley Integrated Campus(Affiliated to JNTU Anantapur), AP, India.

Abstract: Short-instant messages, for example, tweets are being made and shared at an exceptional rate. Tweets, in their crude structure, while being enlightening, can likewise be overpowering. For both end-clients and information investigators, it is a bad dream to push through a large number of tweets which contain gigantic measure of commotion and repetition. In this paper, we propose a novel ceaseless synopsis structure called Sumblr to reduce the issue. As opposed to the customary record outline strategies which concentrate on static and little scale information set, Sumblr is intended to manage progressive, quick arriving, and extensive scale tweet streams. Our proposed structure comprises of three noteworthy segments. In the first place, we propose an online tweet stream grouping calculation to bunch tweets and keep up refined measurements in an information structure called tweet group vector (TCV). Second, we build up a TCV-Rank rundown procedure for creating online synopses and authentic outlines of self-assertive time lengths. Third, we plan a viable point advancement location technique, which screens rundown based/volume-based varieties to create courses of events consequently from tweet streams. Our trials on substantial scale genuine tweets show the productivity and adequacy of our system.

Keywords: Tweet Stream, Ceaseless Synopsis, Rundown, Course Of Events.

I. INTRODUCTION

Increasing notoriety of microblogging administrations, for example, Twitter, Weibo, and Tumblr has brought about the blast of the measure of short-instant messages. Twitter, for occasion, which gets more than 400 million tweets for each day1 has risen as a significant wellspring of news, web journals, sentiments, and then some. Tweets, in their crude structure, while being educational, can likewise be overpowering. For example, hunt down an interesting issue in Twitter may yield a large number of tweets, spreading over weeks. Regardless of the fact that separating is permitted, pushing through such a large number of tweets for vital substance would be a bad dream, also the tremendous measure of commotion and repetition that one may experience. To compound the situation, new tweets fulfilling the sifting criteria may arrive persistently, at an erratic rate. The abnormal state synopsis module bolsters era of two sorts of rundown: online and recorded outlines. (1) To create online synopses, we propose a TCV-Rank synopsis calculation by alluding to the present bunches kept up in memory. This calculation first processes centrality scores for tweets kept in TCVs, and chooses the top-positioned ones regarding content scope and oddity. (2) To figure a chronicled rundown where the client determines a self-assertive time term, we first recover two chronicled bunch depictions from the PTF concerning the two endpoints (the starting and consummation focuses) of the length. At that point, in view of the distinction between the two bunch previews, the TCV-Rank synopsis calculation is connected to produce outlines.

Fig1. The framework of Sumblr.

The center of the course of events era module is a subject development location calculation, which expends online/chronicled synopses to create constant/range timetables. The calculation screens evaluated variety throughout stream preparing. A vast variety at a specific minute infers a sub-theme change, prompting the expansion of another hub on the course of events. In our outline, we consider three distinct figures separately the calculation. To begin with, we consider variety in the principle substance talked about in tweets (as outline). To evaluate the rundown based variety (SUM), we utilize the Jensen-Shannon The primary commitments of this work are as per the following:

- We propose a continuous tweet stream summarization framework, namely Sumblr, to generate summaries and timelines in the context of streams.
- We design a novel data structure called TCV for stream processing, and propose the TCV-Rank algorithm for online and historical summarization.
P. Siva Prasad, M.V. Jayasree

- We propose 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.

II. RELATED WORK

In this section, we review the related work including stream data clustering, document/microblog summarization, timeline detection, and other microblog mining tasks.

A. Stream Data Clustering

Stream data clustering has been widely studied in the literature. BIRCH [2] clusters the data based on an in-memory structure called CF-tree instead of the original large dataset. Bradley et al. [3] proposed a scalable clustering framework which selectively stores important portions of the data, and compresses or discards other portions. CluStream [1] is one of the most classic stream clustering methods. It consists of an online micro-clustering component and an offline macro-clustering component. The pyramidal time frame was also proposed in [1] to recall historical microclusters for different time durations. A variety of services on the Web such as news filtering, text crawling, and topic detecting etc. have posed requirements for text stream clustering. A few algorithms have been proposed to tackle the problem [4], [5], [6], [7]. Most of these techniques adopt partition-based approaches to enable online clustering of stream data. As a consequence, these techniques fail to provide effective analysis on clusters formed over different time durations. In [8], the creators stretched out CluStream to produce term based bunching results for content and clear cut information streams. In any case, this calculation depends on an online stage to produce countless "bunches" and a disconnected stage to re-group them. Conversely, our tweet stream grouping calculation is an online method without additional disconnected bunching. What's more, with regards to tweet synopsis, we adjust the internet bunching stage by joining the new structure TCV, and confining the quantity of groups to ensure productivity and the nature of TCVs.

B. Document/Microblog Summarization

Record synopsis can be arranged as extractive and abstractive. The previous chooses sentences from the records, while the last may create expressions and sentences that don't show up in the first reports. In this paper, we concentrate on extractive synopsis. Extractive report outline has gotten a great deal of late consideration. A large portion of them allotted striking scores to sentences of the records, and select the top-ranked sentences [9], [10], [11]. Some works try to extract summaries without such salient scores. Wang et al. [12] used the symmetric non-negative matrix factorization to cluster sentences and choose sentences in each cluster for summarization. He et al. [13] proposed to summarize documents from the perspective of data reconstruction, and select sentences that can best reconstruct the original documents. In [14], Xu et al. modeled documents (hotel reviews) as multi-attribute uncertain data and optimized a probabilistic coverage problem of the summary. While document summarization has been studied for years, microblog summarization is still in its infancy. Sharifi et al. proposed the Phrase Reinforcement algorithm to summarize tweet posts using a single tweet [15]. Later, Inouye and Kalita proposed a Hybrid TF-IDF algorithm and a Cluster- based algorithm to generate multiple post summaries [6]. In [7], 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 pmedian problem, which takes posted time of microblogs into consideration. Unfortunately, almost all existing document/microblog summarization methods mainly deal with small and static data sets, and rarely pay attention to efficiency and evolution issues. There have also been studies on summarizing microblogs for some specific types of events, e.g., sports events. Shen et al. [9] proposed to identify the participants of events, and generate summaries based on sub-events detected from each participant. Chakrabarti and Punera [2] introduced a solution by learning the underlying hidden state representation of the event, which needs to learn from previous events (football games) with similar structure. In [1], Kubo et al. summarized events by exploiting “good reporters”, depending on event-specific keywords which need to be given in advance. Conversely, we mean to manage general point pertinent tweet streams without such earlier learning. In addition, their technique stores every one of the tweets in every section and chooses a solitary tweet as the rundown, while our strategy keeps up refined data in TCVs to diminish stockpiling/calculation cost, and creates different tweet synopses as far as substance scope and curiosity. Notwithstanding online synopsis, our strategy likewise bolsters verifiable rundown by keeping up TCV previews.

C. Timeline Detection

The interest for dissecting huge substance in social medias fills the improvements in representation methods. Timeline is one of these techniques which can make analysis tasks easier and faster. Diakopoulos and Shamma [2] made early efforts in this area, using timelines to explore the 2008 Presidential Debates by Twitter sentiment. Dork et al. [3] presented a timeline-based backchannel for conversations around events. In [4], Yan et al. proposed the transformative timetable rundown (ETS) to process development courses of events like our own, which comprises of a progression of time-stamped synopses. Nonetheless, in [4], the dates of outlines are dictated by a pre-characterized timestamp set. Interestingly, our strategy finds the changing dates and creates courses of events powerfully amid the procedure of nonstop rundown. In addition, ETS does not concentrate on productivity and versatility issues, which are essential in our spilling setting. A few frameworks identify critical minutes when fast increments or "spikes" in announcement volume happen. TwitInfo [5] built up a calculation in light of TCP blockage discovery, while Nichols et al.[6] utilized a slant based...
strategy to discover spikes. After that, tweets from every minute are distinguished, and word mists or outlines are chosen. Unique in relation to this two-stage approach, our strategy distinguishes theme development and produces synopses/timetables in an online manner.

D. Other Microblog Mining Tasks
The emergence of microblogs has engendered researches on many other mining tasks, including topic modeling [7], storyline generation [8] and event exploration [5]. Most of these researches focus on static data sets instead of data streams. For twitter stream analysis, Yang et al. [9] studied frequent pattern mining and compression. In [3], Van Durme aimed at discourse participants classification and used gender prediction as the example task, which is also a different problem from ours. To sum up, in this work, we propose a new problem called continuous tweet summarization. Different from previous studies, we aim to summarize large-scale and evolutionary tweet streams, producing summaries and timelines in an online fashion.

III. THE SUMBLR FRAMEWORK
As shown in Fig.1, our framework consists of three main modules: the tweet stream clustering module, the high-level summarization module and the timeline generation module. In this section, we shall present each of them in detail.

A. Tweet Stream Clustering
The tweet stream clustering module maintains the online statistical data. Given a topic-based tweet stream, it is able to efficiently cluster the tweets and maintain compact cluster information.

1. Initialization: Toward the begin of the stream, we gather a little number of tweets and utilize a k-implies bunching calculation to make the underlying groups. The relating TCVs are introduced by 1. Next, the stream bunching process begins to incrementally overhaul the TCVs at whatever point another tweet arrives.

2. Incremental Clustering: Assume a tweet t touches base at time ts, and there are N dynamic bunches around then. The key issue is to choose whether to assimilate t into one of the present groups or overhaul t as another bunch. We first discover the bunch whose centroid is the nearest to t. In particular, we get the centroid of every group taking into account Equation (1), figure its cosine similitude to t, and discover the bunch Cp with the biggest likeness (meant as MaxSim(t)). Note that despite the fact that Cp is the nearest to t, it doesn't implied actually has a place with Cp. The reason is that t may in any case be extremely removed from Cp. In such case, a new cluster should be created. The decision of whether to create a new cluster can be made with the following heuristic.

The above updating process is executed upon the arrival of each new tweet. Meanwhile, when the current timestamp is divisible by ai for any integer i, we store the snapshot of the current TCVs into disk and index it by PTF. Algorithm 1 describes the overview of our incremental clustering procedure, operations: deleting outdated clusters and merging similar clusters. Since the computational complexity of deletion is O(N) and that of merging is O(N^3), we use the former method for periodical examination and use the latter method only when memory limit is reached.

IV. CONCLUSION
We proposed a model called Sumblr which bolstered persistent tweet stream rundown. Sumblr utilizes a tweet stream grouping calculation to pack tweets into TCVs and keeps them up in an online manner. At that point, it utilizes a TCV-Rank rundown calculation for creating online outlines and chronicled synopses with self-assertive time spans. The subject advancement can be recognized consequently, permitting Sumblr to create dynamic courses of events for tweet streams. The test results exhibit the proficiency and viability of our strategy. For future work, we plan to build up a multi-theme adaptation of Sumblr in an appropriated framework, and assess it on more finish and expansive scale information sets.

V. REFERENCES


Author’s Profile:

**P. Siva Prasad** is currently PG scholar of CSE in Golden Valley Integrated Campus, Madanapalli, Chittoor (Dist), Affiliated to JNTU Anantapur.

**M.V. Jayasree** working as Assistant Professor in the department of CSE at Golden Valley Integrated Campus, Madanapalli, Chittoor (Dist), Affiliated to JNTU Anantapur.