# Ranking tweets using social influence and content quality

**Prof. D. M. Jadhav[1], Anjum Patel[2]**
[1, 2] Department of Computer Engineering
[1, 2] TCOER, Savitribai Phule Pune University, Kondhwa, Pune-48, Maharashtra, India.

***Abstract-*** *Social Networking site(SNS) and micro blogging site such as Twitter has attracted millions of users for sharing and keeping most up-to-date information, which has resulted in large volumes of data being produced every day. Due to which many applications of Natural Language Processing (NLP) and Information Retrieval (IR) are suffering severely from the noisy and redundant data. The short nature of these messages which are so called tweets have so much hidden and useful information in them. The traditional system which used to work on static and limited data. In this paper we will work on dynamic and fast arriving data which will give us efficient and optimized result. We will first propose an online tweet stream clustering algorithm to cluster tweets and maintain distilled statistics tweet cluster vector (TCV). Second, we develop a TCV-Rank summarization technique for generating online summaries and historical summaries of arbitrary time durations. Third, we design an effective topic evolution detection method, which monitors summary-based/volume-based variations to produce timelines automatically from tweet stream.*

***Keywords-*** Summarisation, dynamic, twitter, social networking sites

## I. INTRODUCTION

In today's world the data is growing rapidly. Twitter an online social networking service which enables users to send and read short 140-character messages called "tweets". Registered users can read and post tweets, but unregistered users can only read them. Users access Twitter through the website interface, SMS, or mobile device app. Twitter Inc. is based in San Francisco and has more than 25 offices around the world. Twitter was created in March 2006 by Jack Dorsey, Evan Williams, Biz Stone, and Noah Glass. The service rapidly gained worldwide popularity, with more than 100 million users who posted 340 million tweets per day in 2012. The service also handled 1.6 billion search queries per day. In 2013, Twitter was one of the ten most-visited websites as of May 2015, Twitter has more than 500 million users, out of which more than 302 million are active users. Tweets are for information communication and sharing ideas. These semantic phrases are well preserved in tweets. Global context derived from the Web pages (e.g., Microsoft Web N-Gram corpus) or Wikipedia therefore helps identifying the meaningful segments in tweets. The method realizing the proposed framework that solely relies on global context.

## II. EXISTING SYSTEM

In the existing system the data which was taken was static data which traditional systems used the mining systems which were used were Natural language processing (NLP) and information retrieval(IR). The problems by of ploughing through so big data was not a good option in static data .

## III. PROBLEM STATEMENT

The existing system provides a summarisation on static data .Using this system the rich information which was hidden in the tweets were having noisy data and redundant data which was not possible to process and find out meaningful rich text.( eg.:she Dacing", #happy#celebration) this example of tweets was not easy for finding rich contxt hidden inside it. So in this paper we will use dynamic fast arriving data which will eventually increase the problem of the inefficiency.

## IV. PROPOSED SYSTEM

The proposed system used filtering method and summarisation technique to where the tweets will be clustered into tweet stream clustering algorithm to cluster tweets and maintain distilled statistics in a data structure called tweet cluster vector (TCV)

### A. Data stream clustered

The Stream data clustering has been widely studied clustering is done based on an in-memory structure called CF-tree instead of the original large data set. Bradley proposed a scalable clustering framework which selectively stores important portions of the data, and compresses or discards other portions. CluStream 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 into 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.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 the authors extended CluStream to generate duration-based clustering results for text and categorical data streams. However, this algorithm relies on an online phase to generate a large number of "micro-clusters" and an offline phase to re-cluster them. In contrast, our tweet stream clustering algorithm is an online procedure without extra offline clustering. And in the context of tweet summarization, we adapt the online clustering phase by incorporating the new structure TCV, and restricting the number of clusters to guarantee efficiency and the quality of TCVs. This phase is used to information retrieval for normalization process.

**B. Document Summarization**

The document summary can be categorized as extractive and abstractive. The former selects sentences from the documents, while the latter may generate phrases and sentences that do not appear in the original documents. In this paper, we focus on extractive summarization. Extractive document summarization has received a lot of recent attention. Most of them assign salient scores to sentences of the documents, and select the top-ranked sentences set.

**C. Detection (Evaluation of List)**

In the similarity computation the keywords of the active user and the previous user is matching with similar taste of the neighbourhood of the user. If the keyword is not matching then there is no similarity between the active user and the previous user. So the comment which are not similar are removed from tha set.

**D. Others**

The emergence of microblogs has engendered researches on many other mining tasks, including topic modeling , storyline generation and event exploration . Most of these researches focus on static data sets instead of data streams. For twitter stream analysis, Yang et al. studied frequent pattern mining and compression. In, 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..

Each review will be converted into a keyword according to the keyword candidate list and Domain thesaurus.

**E. Analysis (Evaluation Process)**

After performing the Map Reduce task of the process, the result will aggregate to generate the recommendation list to the user. Handling noises. The effect of clusters of noises can be diminished by two means in Sumblr. First, in tweet stream clustering, noise clusters which are not updated frequently will be deleted as outdated clusters. Second, in the summarization step, tweets from noise clusters are far less likely to be selected into summary, due to their small LexRank scores and cluster sizes. The process which is used is collaborative filtering algorithm to generate the result of the recommendation list.

## V. ANALYSIS

The summarisation process takes place to predict the accuracy of the recommendation list. The Evaluation will be in the form clusters. Evaluate the performance we will compared this with the other method and item based algorithm.

## VI. CONCLUSION

We proposed a sytem which support continuous tweet stream summarization. Sumblr employs a tweet stream clustering algorithm to compress tweets in an online fashion. It, then uses a TCV-Rank summarization algorithm for generating online summaries and historical summaries with arbitrary time durations. This evolution can be detected automatically, allowing Sumblr to produce dynamic timelines for tweet streams. The experimental results demonstrate the efficiency and effectiveness of our method. For future work, we aim to develop a multi-topic version of Sumblr in a distributed system, and evaluate it on more complete and large-scale data sets..

## ACKNOWLEDGMENT

I consider myself most fortunate and to have worked under guidance of Prof. D. M. Jadhav Faculty and Computing. I would like to thank to all other teachers and friends who are really helping me to make project successfully.

## REFERENCES

[1] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in Proc. 29th Int. Conf. Very Large Data Bases, 2003, pp. 81–92.

[2] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 1996, pp. 103–114.

[3] P. S. Bradley, U. M. Fayyad, and C. Reina, "Scaling clustering algorithms to large databases," in Proc. Knowl. Discovery Data Mining, 1998, pp. 9–15.

[4] L. Gong, J. Zeng, and S. Zhang, "Text stream clustering algorithm based on adaptive feature selection," Expert Syst. Appl., vol. 38, no. 3, pp. 1393–1399, 2011.

[5] Q. He, K. Chang, E.-P. Lim, and J. Zhang, "Bursty feature representation for clustering text streams," in Proc. SIAM Int. Conf. Data Mining, 2007, pp. 491–496.

[6] J. Zhang, Z. Ghahramani, and Y. Yang, "A probabilistic model for online document clustering with application to novelty detection," inProc.Adv.NeuralInf.Process. Syst.,2004,pp.1617–1624.

[7] S. Zhong, "Efficient streaming text clustering," Neural Netw., vol. 18, nos. 5/6, pp. 790–798, 2005.

[8] C. C. Aggarwal and P. S. Yu, "On clustering massive text and categorical data streams," Knowl. Inf. Syst., vol. 24, no. 2, pp. 171–196, 2010.

[9] R. Barzilay and M. Elhadad, "Using lexical chains for text summarization," in Proc. ACL Workshop Intell. Scalable Text Summarization, 1997, pp. 10–17.

[10] W.-T. Yih, J. Goodman, L. Vanderwende, and H. Suzuki, "Multidocument summarization by maximizing informative contentwords,"inProc.20thInt.JointConf.Artif. Intell.,2007,pp.1776–1782.

[11] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," J. Artif. Int. Res., vol. 22, no. 1, pp. 457–479, 2004.

[12] D. Wang, T. Li, S. Zhu, and C. Ding, "Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization," in Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2008, pp. 307–314.