Indian Journal of Science and Technology, Vol 9(44), DOI: 10.17485/ijst/2016/v9i44/102441, November 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Modeling of Tweet Summarization Systems using Data Mining Techniques: A Review Report

V. Laxmi Narasamma and M. Sreedevi

Department of Computer Science and Engineering, KL University, Vaddeswaram 522502, Andhra Preadesh, India; lakshmi4540@gmail.com, msreedevi_27@kluniversity.in

Abstract

Objective: Data mining is the driving force for analysing and summarization of available data in various forms and restores it for further needs. Summarization of various literature studies have been done by the researchers based on tweets and its summarization method applied on the datasets has to be identified for analysis. **Methods/Statistical Analysis:** The analysis has been in learning the methods or techniques used from the literature of various researches in gathering knowledge of various tweets datasets used and the way in which they have analysed the datasets from small tweets of unstructured to the large blogs. **Findings:** Various pro and cons of techniques and methods used by the researchers are identified as to the knowledge for better development of new methods for fast and accurate data analysis on tweets and blog. **Application/Improvements:** The paper gives us an idea to data experts and user how to prevent issues of tweets and various methods used for tweets analysis timely for summarization and data analysis.

Keywords: Clustering, Micro-Blogging, Summarization, Timeline Generation, Tweets

1. Introduction

Twitter messages are growing exceedingly every minute day by day with combinational information of topics. The Tweets which are written by various users can be searched related to the applications for enquiry and use. This information notifies to the users and research to know ongoing day by day information and can ask query related to the information. Tweets can be accessible in irregular form which has worthless and worth full information. The worth full information is even dangerous to use it for analysis directly before summarization and refined using mining techniques.

It is very difficult to summarize the stream tweets, because tweets information will be in different forms, this information should be gathered and segregation into related groups refined using mining techniques[16]. Summarizing is a process of reducing the size of the data or datasets and refine the useful information of the essential in a quick time and continues way is shown in Figure 1.

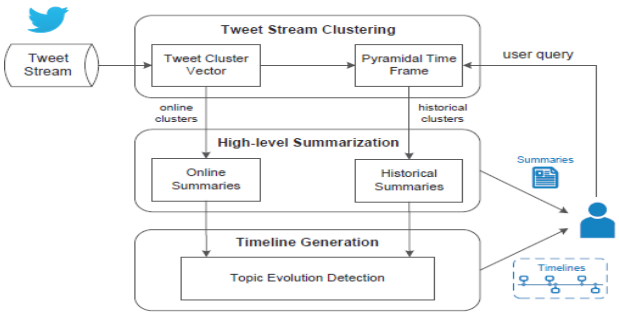

Figure 1. Architectural View

2. Related Work Done

In[1] designed a framework named TCS-Frame work which uses a paperback summarize process. The framework design has two phases. In phase I, tweets of chronicle type are clustered together using cluster tweet algorithm and the output are filtered tweets are kept in a separate structure dataset. In phase II, Rank based TC algorithm is

*Author for correspondence

proposed and evaluated on the filtered dataset, it is shown that the proposed approach requires less time compared to the existing one.

In[2] studies have been made on basic summarization concrete methods and applied the methods on ancient dataset. This concrete method summarizes only fixed and small datasets very quickly and aggressively. The authors have proposed a method which works in 3 stages. The stage algorithm clusters the stream of tweets, refine it and store in a vector database. The second rank based summarization is done on vector database, and last check for the time assessment on the tweet database.

In[3] studied the issues occurred at the user side and learnt the life cycle of tweet stream. The authors proposed unsupervised learning method, which prepares the datasets in order of each individual. Each of the individual tweets are studied based on acceptance or not acceptable, then check for the timeline of each tweets are discussed, discussion is been done based on point of specification. Further, they also extracted and sort the information based on 4 types

In[4] developed a framework of 3 components, First tweets are cluster based on clustering method using k-mean algorithm. Second clustered tweets are ranked based on the rank summarization which is done by greedy method. Third rank based tweets are monitor and analysed for better usage timely.

In[5] devised a novel model for indexing and creating multi-layer data base search to increase the speed and update the tweets collected for refining it. The search process that is used is Nearest Neighbour, which is based on pruning bound. He also added the indication of maintaining index process in effective maintenance of structure effectively in execution and monitoring of events regularly.

In[6] Derived a model which streams tweets using clustering methods. The developed model is based on statistical tweet stream online algorithm using vectors. Summarization based on Rank TCV procedure is developed to summarize the records of tweets and they are self-assisted timely based on required or not required. The algorithm developed is added with detection method, which detects the tweets online and summarizes them chronically and timeliness with accurate and efficiently. He also conducted experiments in comparing he algorithms efficiency with other using various methodology.

In[7] derived a model which streams tweets using clustering methods, clustered tweets are refined using statistical methods named as vector cluster.

In[8] have developed extended Hidden Markov Model for solving the issues raised in summarization of tweets and tweets events, the questions are solved using learning hidden underlying method for event representation. **Extensive analysis was also made related to real world data and practically implemented the significance of the model comparing with the outcomes of the existing one.**

In[9] devised a system INTWEEMS, which is used to tweet clusters progressively into existing clustering of the new ones. The new tweets are converted into existing one using incrimination methods of perception one, these clusters of the existing are differenced with latency points in the tweets. The paper also proposes the framework of incremental clustering tweets and its advantages.

In[10] summarized a framework for managing the data issues. He builds data stream tweets of larger scale which analysis data at a faster rate. The developed system works on 3 steps, in the first step, clustering is done to tweet data which is accepted dynamically online and stored in a data structure called vector cluster (TCV). Next, proposed a novel TCV rank summarization method to summarize online data chronically at any duration of time. At last and final, the developed method identifies various type of data continuously from tweets online timely of large volumes, delivers and summarizes it and stored it in a database.

In[11] has proposed an event which continuously evaluate the stream of tweet online and produces the require information timely. To generate timely information, clustering method is used on stream tweets on related information and summarizes them continuously and stores each of the clustered stream tweets appropriately. Conventional method on date is used for summarization of tweets using timeline. The paper proposed by the author also specifies the dynamic procedure for continuous tweet data summarization.

In[12] studied various literatures on tweets and their methodologies on summarization. They stated that, the principle idea behind summarization is to principles the content based on expected results. Comparison is also done by the authors on the previous existing summarization algorithms, based on the comparison data tweets are categorised and algorithms mainly focus on passive, tweet extensive streams and scalar data stream.

In[13] developed an innovative summarization model, which deals mainly on small dataset of passive nature and concretes the dataset at a rapid speed on arrival dynamically. At first the arrived tweets are clustered using vector clustering method, next TCV summarization is done based on rank for creating summaries online at a rapid rate and stored them timely. Lastly evolution method is used to screen the tweets based on time and type of stream tweet.

In[14] proposed an issue on summarizing data stream in various stage for pre-processing and clustering of text. He has shown the improvement in quality of summarization by grouping related into cluster posts.

In[15] developed a graph word algorithm and optimization technique for pruning and windows decay. This method gives quality output comparatively based on memory efficiency and time.

In[16] proposed a novel method which deals with summarization on topic specific and customer specific datasets adequately for analysis. This method is compared with the available approaches based on query management, cost of capacity, flexibility of extraction on query and also compared with the pattern mining summarization. The developed method initiate scalable and performance was better.

3. Other Micro Blog Mining Tasks

Numerous Mining techniques has been used in modelling unstructured and structured datasets such as generation of storyline, exploration of events, recommendation made on blog sites of the product have been implemented using summarization of tweets. Most of the works done by the researchers are based on unreceptive data sets rather than data streams. Compression and Pattern Frequency mining are used for stream analysis on twitter. Past studies shown that twitter summarization has done only to tweets of static not to dynamic, but the changes in recent requires summarization to be done for large-scale and tweet evolutionary streams, which has to created timely and has to be summarized online in substantial manner.

4. Conclusion

The summarization of various studies made by different people is based on document summarization technique such as tweet and filtering. The techniques developed are well utilized for overseeing tweets counts. Filtering methodology does not give effective results on tweet datasets because of its repetitive and noisy information. So Summarization is a process used to review the datasets of tweet. The work which was done earlier and the techniques used to summarization is for dynamic and static data only.

5. References

1. Rutuja K. Ingavale, Saniya D. Latkar, et.al. Tweet Summarisation and Timeline Generation using Clustering, International Journal of Innovative Research in Computer and Communication Engineering. 2016 Feb; 4(2).
2. Tejaswini K, Madhavi H, Sreenivasa BR. Automatic Range Timeline Generation for Volume Based Tweet Stream, International Journal of Advanced Networking and Applications (IJANA), 1st International Conference on Innovations in Computing and Networking (ICICN16). p. 431–35
3. Jiwei Li, Claire Cardie. Timeline Generation: Tracking Individuals on Twitter, International World Wide Web Conference Committee (IW3C2), WWW'14, April 7–11, 2014.
4. Sonali Karle, Nawathe AN. Survey on Summarization of Tweet Data, IJARIIE, 2(2), 2016.
5. Hongyun Cai, Zi Huang, Divesh Srivastava, Qing Zhang. Indexing Evolving Events from Tweet Streams, IEEE Transactions on Knowledge and Data Engineering, 27(11), Nov. 2015.
6. Lidan Shou, Zhenhua Wang, et.al. Sumblr: Continuous Summarization of Evolving Tweet Streams, Proceedings of the 36th International ACM SIGIR'13 Conference on Research and development in Information Retrieval, p. 533–42.
7. Deepayan Chakrabarti, Kunal Punera. Event Summarization Using Tweets, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011.
8. Muhammad Farid Khan, Rabeeh Ayaz Abbasi, Naif Radi Aljohani, Aiiad A. Albeshri, Mubashar Mushtaq. INTWEEMS: a Framework for Incremental Clustering of Tweet Streams. In: Proceedings of the 17th International Conference on Information Integration and Web-Based Applications and Services (iiWAS'15).
9. Miss. Geeta G. Dayalani. Tweet Streams Online Summarization and Timeline Generation, International Journal of Advance Scientific Research and Engineering Trends, 2016 Jun; 1(3).
10. Dhanshri A. Nevase, Amrit Priydarshi. Timeline Generation for Progressive Tweet Stream, International Journal of

Advanced Research in Computer and Communication Engineering. 2016 Jun; 5(6).
11. Prashant S. Bagade, Shinde SA. Survey on Tweet Summarization Approaches, International Journal of Advanced Research in Computer and Communication Engineering. 2015 Dec; 4(12).
12. Sree Harika G, Sreenivasulu M. A Novel Approach for Tweet Summarization through Timeline Generation, International Journal of Computer Science Engineering and Scientific Technology, Aug 2016.
13. Andrei Olariu. Clustering to Improve Microblog Stream Summarization, 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2012. DOI: 10.1109/SYNASC.2012.10.
14. Andrei Olariu. Efficient Online Summarization of Microblogging Streams, Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics.
15. Xintian Yang, Amol Ghoting. A Framework for Summarizing and Analyzing Twitter Feeds, In: KDD'12, Beijing, China, 12–16 August, 2012.
16. Dhanshri A. Nevase, Amrit Priyadarshi. Survey on Data Clustering and Summarization, AR Digitech International Journal of Engineering Education and Technology (ARDIJEET), 2016, 4(1).
17. Thiyagarajan VS. Platfora Method for High Data Delivery in Large Datasets, Indian Journal of Science and Technology. 2015 Dec. DOI: 10.17485/ijst/2015/v8i33/76517.
18. Ehsan Sargolzaei, Mohammad Javad Khazali and Fateme Keikha. Privacy Preserving Approach of Published Social Networks Data with Vertex and Edge Modification AlgorithmIndian, Journal of Science and Technology, 2016 Mar. DOI: 10.17485/ijst/2016/v9i12/81982
19. Iqbaldeep Kaur, Namita Arora. Comparative Analysis of Information Extraction Techniques for Data Mining Amit Verma, Indian Journal of Science and Technology. 2016 Mar. DOI: 10.17485/ijst/2016/v9i11/80464.
20. Karthick S. Analysis of Data Mining Techniques for Weather Prediction, Indian Journal of Science and Technology, 2016 , Oct. DOI: 10.17485/ijst/2016/v9i39/93184,
21. Nayan Mattani J. Sharath Kumar A, Prabakaran, Maheswari N. Privacy Preservation in Social Network Analysis using Edge Weight Perturbation, Indian Journal of Science and Technology, 2016 Oct. DOI: 10.17485/ijst/2016/v9i37/93810.