International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2016): 79.57 | Impact Factor (2015): 6.391

Continuous Multitopic Tweet Summarization and Timeline Generation using Clustering

Seema Shivajirao Malkar[1], Dr. D. V. Kodavade [2]

[1] PG Student, Department of Computer Science and Engineering, D.K.T.E.Society's Textile and Engineering Institute, Ichalkaranji, India

[2] Professor & Dean, Department of Computer Science and Engineering, D.K.T.E. Society's Textile and Engineering Institute, Ichalkaranji, India

Abstract: ***Every day twitter receives 500 million tweets with emerged as an invaluable source of news, blogs, unwanted information and more. Continuous tweet cannot show information correctly. Our proposed work consist summarization and opinion mining technique for data analysis. First collect the tweet online and historical from internet, in first technique opinion mining can show fast result and show emotion with score about online tweet by using sentiment analysis. Second technique summarization first cluster the tweet using K-means clustering algorithm ,tweet data structure represent statically known as tweet cluster vector and then formulation of incremental cluster is done. In summarization incremental tweet match with present tweet then add into the specific cluster; if not then declare it is in new cluster. By using summarization evolutes most trending topic very fast. The paper discussed study report of new approach for tweet summarization.***

1. Introduction

Large number of Tweeted data is being created and shared daily. The rate in 500 million tweets posted per day. In data analysis it is required to extract the tweets, cluster them and summarize them for proper viewing of data. We propose to develop a framework will cluster the tweets according to topic and then according to subtopic. The tweets are then summarized and graph will be generated depicting the current trend of the tweeted messages. In continuous tweets summarization is most important part by using the summarization we can easily find out our main topic of Tweet.

Using summarization method analysis the data from continuous tweet stream which are historical and online tweets and provide trending topic on basis of tf idf calculation and cosine similarity. Another one method for analysis data is opinion mining; in opinion mining fast analysis data provide sentiment about the tweet with score.

Continuous tweet are clustered first; for clustering required some space. Merging is best method for save space and reduce complex data. If most similar functionality cluster known as composite cluster are merged then some space of our memory saved as like these merging done.

Another one method is delete outdated tweet. For example consider sport tweets some oldest tweets related to sport are not useful yet. If as unwanted tweets from our clustered tweet are deleted then it was easy for summarization .

2. Literature Review

T. Zhang et.al. [1] discussed previous clustering algorithm which are less effective for large data set and problem for fitting it into large data set in main memory. To overcome these problem BIRCH cluster algorithm used, BIRCH incrementally and dynamically incoming multidimensional data points to try to produce best quality cluster. It can be first check memory limit by removing noise can adjust data in disk. BIRCH is single scan good clustering algorithm.

P. S. Bradley et.al. [3] saved important portion of data and compress or delete other part of data by using traditional clustering method time required for cluster are more for large scale data set. Scaling technique can not required more time for clustering because it is done in single scan.

C. C. Aggarwal et.al [4]discoursed about clustering problem if large volume of data come then clustering are difficult for single scan of data set. This problem can be solving using CluStream method. It can be divide the clustering method into two parts one is online component and other is offline component, in online component store periodic summary and offline component stores statistic summary.

Table I

Paper Name	Technique	Advantages	Disadvantages	Result
BIRCH: An efficient data clustering method for very large databases	BIRCH are single scan good clustering algorithm	BIRCH incrementally and dynamically incoming multidimensional data points to try to produce best quality cluster. It can be first check memory limit by removing noise can adjust data in disk.	Required more time for clustering of large scale data set.	Provide best quality cluster.
Scaling clustering algorithms to large databases	saved important portion of data and compress or delete other part of data	Using traditional clustering method time required for cluster are more for large scale data set. Scaling technique can not required more time for clustering because it is done in single scan.	Clustering problem if large volume of data come then clustering is difficult for single scan of data set.	Provide result in single scan of dataset

A framework for clustering evolving data streams	CluStream method.	It can be divide the clustering method into two part one is online component and other is offline component ,in online component store periodic summary and offline component stores statistic summary	Scan extra offline cluster	Relies large number of micro cluster on online phase and at offline phase re cluster again.
LexRank: Graph-based lexical central-ity as salience in text summarization	Degree centrality, Lex rank	Graph-based lexical centrality as salience in text summarization, introduce a stochastic graph-based method for computing relative importance of textual units for Natural Language Processing.	Path of graph are not in sequence.	Provide graph-based lexical centrality as salience in text summarization
On Summarization and Timeline Generation for Evolutionary Tweet Streams	summarization and timeline generation	Summarized tweet can be detect current trending topic, online tweet can be analysed and place into particular topic cluster, outdated cluster can be deleted and composite cluster can be merged.	summarization and timeline generation done for only single topic	Produce Summary about continuous tweet and trending topic

G. Erkan and D. R. Radev [2]discoursed LexRank: graph-based lexical centrality as salience in text summarization We introduce a stochastic graph-based method for computing relative importance of textual units for Natural Language Processing. We test the technique on the problem of Text Summarization (TS). Extractive TS relies on the concept of sentence salience to identify the most important sentences in a document or set of documents. Salience is typically defined in terms of the presence of particular important words or in terms of similarity to a centroid pseudo-sentence. We consider a new approach, LexRank, for computing sentence importance based on the concept of eigenvector centrality in a graph representation of sentences. In this model, a connectivity matrix based on intra-sentence cosine similarity is used as the adjacency matrix of the graph representation of sentences. Our system, based on LexRank ranked in first place in more than one task in the recent DUC 2004 evaluation. In this paper we present a detailed analysis of our approach and apply it to a larger data set including data from earlier DUC evaluations. We discuss several methods to compute centrality using the similarity graph. The results show that degree-based methods (including LexRank) outperform both centroid-based methods and other systems participating in DUC in most of the cases. Furthermore, the LexRank with threshold method outperforms the other degree-based techniques including continuous LexRank.

Zhenhua Wang et.al. [12] discoursed the summarization and timeline generation for single topic. Tweet stream clustering algorithm create number of clusters, if any new tweet come compare with available cluster tweet data set if it matches then add into it, if not then declare it is new cluster. These method face problem for multitopic tweet clustering if apply on that then provide wrong result. The detailed literature survey is discussed in Table I.

3. System Architecture

System architecture shown in Fig.1 Proposed approach work like following steps:

3.1 Opinion Mining

By using opinion mining, the tweet data streams are filtered first then analyzed and provide fast feedback. The twitter data is obtained and filtered, so that it is in text format. This data is then stored in an external database. Wordnet is lexical database that group's English word into set of synonyms called as synset. The SentiWordNet is an extension of wordnet that adds for each synset into 3 measures that are positive, negative and object score measure.

Opinion mining can be done in five steps first Tokenization in that split the tweet into very simple token such as punctuation, number and words of different types, speech tagging tag the notation as like noun ,verb ,adverb based on role of each word in the tweet, second part word sense disambiguation(WSD) is used to determine meaning of every word after WSD, third part is Sentiwordnet interpretation here find sentiment positive and negative score associate to the synset, fourth step is sentiment orientation here summed separately positive and negative score of each term found in tweet. In last step tweet classification sentiment of tweet is determined based on higher value of positive and negative score.

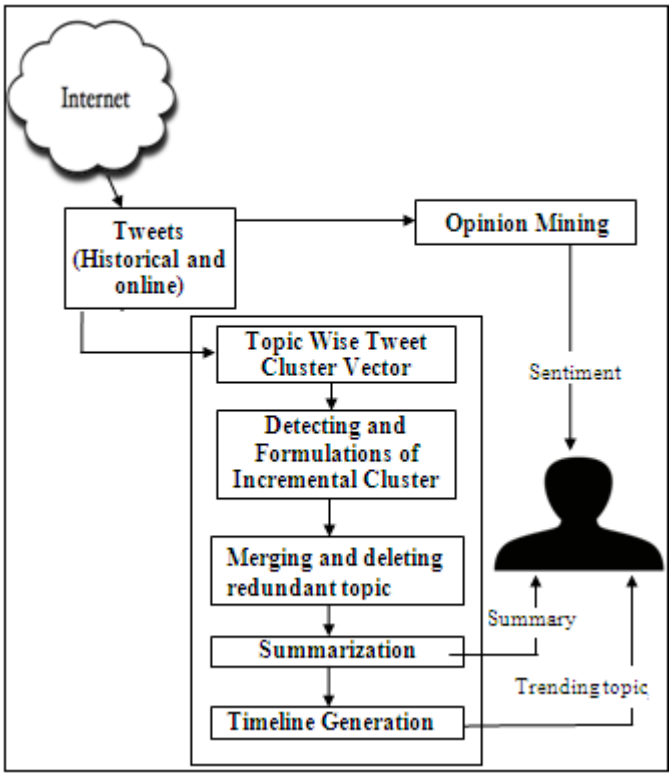

Figure 1: System architecture

3.2 Create Tweet Cluster Vector

The k means algorithm is used to create the initial cluster but tweet stream clustering algorithm required statistics for tweet to generate summary. For that purpose tweet cluster represent in data structure is known as tweet cluster vector. First get a sample collection of tweets, along with the time stamp of the tweet. Taking the sum of weighted textual vector, sum of normalized textual vectors, sum of posted timestamp of tweet arrive by using this tweet cluster vector is created.

3.3 Incremental Clustering

When a tweet arrives online, initially find the cluster whose centroid is the closest to the arrived tweet. Specifically, then get the centroid of each cluster compute its cosine similarity to that tweet then find the cluster Cp with the largest similarity. If largest similarity found then add tweet into particular cluster. If the distance between the tweet and cluster is very large then a new cluster is created.

3.4 TimeLine

The algorithm discovers topic changes by monitoring quantified variations during the course of stream processing by using summary based variation evaluate the timeline. A large Variation at a particular moment suggests a sub-topic change, which is a new node on the timeline.

3.5 Merging

Same functionality cluster are merge together for simplicity

3.6 Deleting

Timeliness is more important in tweet because tweets are not final for long time. Find out cluster having subtopic are rarely discussed, delete these kind of cluster and create space in memory.

3.7 Summaries

In order to derive summaries of the current trending topic the entire set of current clusters are subjected to the summarization algorithm.

4. Conclusion

We propose framework support continuous tweets which are collected from internet in from historical and online tweets. Framework shows sentiment analysis about tweet and tweet stream clustering algorithm apply on collected tweet. Tweet cluster vector rank summarization algorithm generated with arbitrary time duration. Incremental clustering which apply on online new tweet. If new tweet match with available tweet the add into particular cluster otherwise declare it new cluster. Our framework deleted outdated clusters and merge similar clusters.

5. Future Scope

For future work our aim to develop multi topic version frameworks in distributed system and evaluate continuous large scale datasets. The algorithm that has been used here for the summarization and clustering of tweets in cosine similarity algorithm and frequency of keywords. The different techniques like hierarchical clustering can be used for same.

References

[1] A. Bonnaccorsi, “On the Relationship between Firm Size and Export Intensity,” Journal of International Business Studies, XXIII (4), pp. 605-635, 1992. (journal style)
[2] G. Erkan and D. R. Radev, “LexRank: Graph- based lexical central-ity as salience in text summarization,” J. Artif. Int. Res., vol. 22, no. 1, pp. 457–479, 2004.
[3] M. Clerc, “The Swarm and the Queen: Towards a Deterministic and Adaptive Particle Swarm Optimization,” In Proceedings of the IEEE Congress on Evolutionary Computation (CEC), pp. 1951-1957, 1999. (conference style)
[4] H.H. Crokell, “Specialization and International Competitiveness,” in Managing the Multinational Subsidiary, H. Etemad and L. S, Sulude (eds.), Croom-Helm, London, 1986. (book chapter style)
[5] K. Deb, S. Agrawal, A. Pratab, T. Meyarivan, “A Fast Elitist Non-dominated Sorting Genetic Algorithms for Multiobjective Optimization: NSGA II,” KanGAL report 200001, Indian Institute of Technology, Kanpur, India, 2000. (technical report style)
[6] J. Geralds, "Sega Ends Production of Dreamcast," vnunet.com, para. 2, Jan. 31, 2001. [Online]. Available: http://nl1.vnunet.com/news/1116995. [Accessed: Sept. 12, 2004]. (General Internet site)

Author Profile

Seema Shivajirao Malkar has completed B.E. Computer Science and Engineering from Shivaji University, Kolhapur. She is currently pursuing M.E. in Computer Science and Engineering at D. K. T. E.’s Textile and Engineering Institute, Ichalkaranji, India. Her areas of interest include Information Security and Data Mining.

Prof. Dattatraya.V.Kodavade, working as Professor in Computer Sc. & Engg,, he is member of Board Of Studies Computer Sc. & Engg. Shivaji University , Kolhapur , he has completed M.E. and PhD and having 25 years teaching experience in teaching He has presented and published more than 25 research papers in International Conferences and Journals. His areas of research includes Artificial Intelligence ,IoT, Data structures, Algorithms, Big Data etc.