Topic Detection and Summarization of Events on Social Media Data

Rajani D. Gavali
Department of Computer Science and Engineering
Walchand Institute of Technology, Solapur

A.R. Kulkarni
Department of Computer Science and Engineering
Walchand Institute of Technology, Solapur

ABSTRACT
Millions of Internet users use social sites for sharing and storing data, blogs for connecting with people and sharing information. Twitter is one of the fastest growing social sites. Short-text messages are being posted and shared at a unique rate. Twitter collects millions of tweets, which contain lots of noise and redundancy, un-structured tweets.

As redundancy leads to inconsistency and less accurate result, users are unable to understand current topics of discussion. Due to time constraints readers are unable to read each and every tweet, so user requires summary to understand important information on social media. To generate summary for large amount of data, a new summarization method is proposed, namely sequential summarization, which provides a topic detection of social media and generate ordered short sub-summaries for a trending topic in order to convey the important information in few sentences. The system will implement two approaches, stream-based and semantic-based, for detecting the necessary and non-redundant subtopics within trending information.

Keywords
Topic detection, Social media data, Summarization, event detection

1. INTRODUCTION

1.1 Motivation
Social media is a good way for millions of people to express their ideas, thoughts and share information. Many people constantly post on social media regarding many topics. Twitter is fastest growing social networking service that allows users to send and read the huge amount of short-text messages, which contains news, blogs, opinions, messages and they are continuously increasing.

It is not possible for human to read each and every Tweet to get correct and accurate information related to specified topic. To understand all information of social media in less time, user needs summary which is a shortened version of a text that contains only the key points of the original content. The traditional summarization methods focus only on static and small-scale data set. To give correct and accurate information related to social media, topic detection is helpful.

1.2 Problems and Solutions
Some problems faced by people when reading tweets of any topic such as, tweets related to specific topic are not in serial order, noisy and redundant tweets. To avoid such problems one possible solution is to give summary of social sites. A good summary should cover the key points. Users can easily read or process this text or tweet in less time. There are so many text summarization algorithms are available, from that system will use correct, accurate and efficient algorithm for social media topic summarization. Using algorithm, Sequential summarization method will produce summary.

1.3 Significance
- Summarization method provide summary in short, meaningful and user accepted manner.
- Summarization method conveys the important information in short sentence.
- Summarizing important points quickly which is time saving work.

2. RELATED WORK
In the previous research, different techniques were presented for generating summary Related to Twitter

In this paper, a novel continuous summarization framework Sumblr is used. It is designed to deal with dynamic, fastest arriving, and huge scale tweet streams. It consists of different major components. They proposed an algorithm for clustering called online tweet stream clustering for the tweets and maintain statistics in a data structure as tweet cluster vector (TCV). Second, they proposed new summarization technique, a TCV-Rank for online summaries and historical summaries of arbitrary time durations. Third, they implemented an effective topic detection method, which is used to monitor variations in summary-based/volume-based to produce timelines easily and automatically from streams [2].

In this paper, they presented the real-time interaction of events like earthquakes in Twitter and designed an algorithm to monitor tweets to detect a target event. To detect a target event, they made use of devise classifier of tweets which is based on features like keywords in a tweet, the number of words, and their context. They also considered each user of twitter as a sensor and apply filtering such as Kalman filtering and particle filtering, which are used for detection of location estimation in ubiquitous/pervasive computing [3].

In this work presented the Phrase Reinforcement algorithm for Automatic Twitter topic summarization. Phrase Reinforcement algorithm is used to produce summary of social site Twitter topic which is accurate and meaningful. In this method it calculates occurrences of each word in every sentence and according high word occurrence rank or count, it builds a graph and then it produces summary by combining or merging nodes in the graph with the help of count of highest word occurrence. They used a front-end classifier for trending topics within different categories of sports, politics, and world events, then summarizing these topics in order to generate an automated real-time newspaper [4].

This paper presents algorithms for summarizing micro blog posts. Algorithm is used for collections of short posts or messages on particular topics on the social site Twitter and

presented short summaries from these collections of messages on that specific topic. Here, goal is to produce summaries which are similar to human produced summary for the same collection of messages on a particular topic. They also compare the summaries which are produced using summarizing algorithms with human created summaries and get the excellent results [5].

In this paper, they presented summarization of Topics on Twitter of Tweet Ranking by using the method Social Influence and Content Quality. In this work they used and performed sub-topic segmentation and summarization. Also they performed content quality estimation, generation of summary by removing redundancy of tweets related to topics [6]

In this paper, they stated and formalized the problem which occurred during summarizing of event and presented a solution which was based on learning the underlying hidden state representation of the event by using Hidden Markov Models. They also showed through extensive experiments on real-world data and showed that their model significantly outperforms some intuitive and competitive baselines. They showed frequently occurred events like sports. For that it is good to use more sophisticated methods for summarizing the relevant tweets [7].

3. PROPOSED METHODOLOGY

3.1 System Design

The information of proposed system is given below:

Twitter has provided open source Twitter Stream API for downloading tweets. Tweet collector is used to collect various tweets. Tweet filter is used to remove links, non-English, symbols, redundant tweets.

Topic is composed of subtopics and approaches the summarization method by means of two models, i.e., detection of subtopics and sequential summary generation. The goal of subtopic detection is to identify a serial of time-ordered tweet sets such that each tweet set represents a subtopic of the topic. The final Summary is actual output of the proposed system (refer Fig 1).

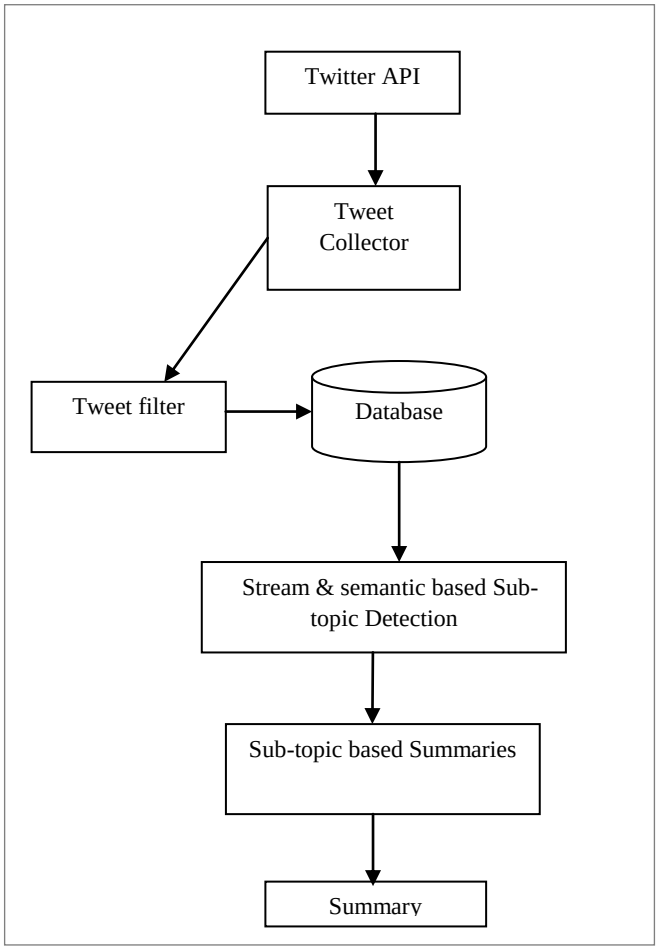

Fig 1: System Design

3.2 Sequential Summarization

Sequential summarization proposed to generate a series of chronologically ordered sub summaries of subtopics for a given Twitter topic. Each summary is supposed to represent one main subtopic or one main aspect of the topic, By using this way, the sequential summary will provide a general plan of the entire topic development.

3.3 Subtopic Segmentation

One of the keys to sequential summarization is subtopic segmentation. Proposed system will develop the subtopic segmentation. For that it is important to find some answers for questions like which subtopics have attracted the public attention, and how are they developed? It is important to provide the valuable and organized materials for more fine-grained summarization approaches. The following two approaches will automatically detect and chronologically order the subtopics.

3.3.1 Stream based Subtopic Detection and Ordering

When a subtopic is most popular enough, it will create a certain level of forward or upward movement in the tweet stream. In other words, every movement in the tweet stream can use as an indicator of the appearance of a subtopic that is of being summarized. The stream-based subtopic detection approach employs the offline peak area detection (OPAD) algorithm to locate such movements by tracing tweet volume changes. It related to the collection of tweets at change of volume time range can consider as a new subtopic.

The subtopics detected by the OPAD algorithm are naturally ordered in the timeline. Proposed system will use the Offline Peak Area Detection (OPAD) algorithm to locate volume changes and movements of tweet stream. It will define the concept, Peak Area (PA) which represents the tweets within the time spans.

3.3.2 Semantic based Subtopic Detection and Ordering

The stream-based approach focuses on the changes of the level of user attention. It is easy to implement, but it fails to handle the cases where the same subtopic are received at different time ranges. In order to segment the subtopics from the semantic aspect, the semantic-based subtopic detection approach divides the time of tweet stream, and regards each tweet as an individual short document

4. CONCLUSION

The proposed work categorizes the tweets into different topics and finds the trending topics (topics that have large number of tweets). It then provides the summary of the tweets on trending topics. It gives summary as per user requirement which is meaningful and conveys important information in few sentences saving the time of the user.

Proposed system can be extended in future to provide multimedia feature like pictures and its related tweets in the summary to make it more meaningful to the readers of the summary.

5. REFERENCES

[1] Dehong Gao, Wenjie Li, Xiaoyan Cai, Renxian Zhang and You Ouyang. " Sequential Summarization of Twitter Trending Topics" *" IEEE/ACM Trans on Audio ,speech and language processing,vol.22 No 2,Feb 2014*

[2] Zhenhua Wang, LidanShou Ke Chen, Gang Chen and Sharad Mehrotra,. "On summarization And Timeline Generation for Evolutionary Tweet Streams*". "IEEE Trans Knowl. Data Eng.*,vol 27,No 5,May 2015

[3] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: Real-time event detection by social sensors," in *Proc WWW-10*,2010, pp. 851–860

[4] B. Sharifi, M.-A. Hutton, and J. K. Kalita, "Automatic summarization of Twitter topics," in Proc. National Workshop Design Anal.Algorithms,2010.

[5] B. Sharifi, M.-A. Hutton, and J. K. Kalita, "Experiments in microblog summarization,"in Proc. SOCIALCOM-10, 2010, pp. 49–56.

[6] Y. Duan, Z. Chen, F. Wei, M. Zhou, and H.-Y. Shum, "Twitter topic summarization by ranking tweets using social influence and content quality,"*Proc. Coling-12*, pp. 763–780, 2012.

[7] D. Chakrabarti and K. Punera, "Event summarization using tweets," in *Proc. AAAI-11*, 2011.

[8] S. Petrovic, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to Twitter," in *Proc. ACL-10*, 2010

[9] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller, "Twitinfo: Aggregating and visualizing microblogs for event exploration," in *Proc. CHI-11*, 2011, pp. 227–236.

[10] Gulab R. Shaikh, Digambar M. Padulkar "Template Based Abstractive Summarization of Twitter Topic with Speech Act" by Asst. Prof., Department of CSE, VPCOE Baramati, Pune, India, India in June 2014.