IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

Survey on Tweet Summarization Approaches

Prashant S. Bagade[1], Prof. S. A. Shinde[2]

Dept. of Computer Engineering SKNCOE, Savitribai Phule Pune University, Pune, India[1, 2]

Abstract: Now-a-days, the large amounts of short messages are shared among multiple peoples and data. These short messages are known as tweets on social networking sites and micro blogging sites. The recent observations said that twitter receives over hundreds million tweets per day. It is a very difficult and challenging task to analyze such huge data. The querying and retrieval of data is also difficult for this situation. Such millions of tweets contain maximum amount of noise and redundancy. The searching in such raw tweets is a very complicated task. The solution is just filtering such tweets for important contents but this is also difficult for searching through such huge tweets because of noise and redundancy. So the possible solution to information overload problem is summarization. Summarization represents restating of the main ideas of the text in as few words as possible. There are various algorithms are available for tweet summarization, some of them focus on static and small-scale data set and others on dynamic, fast arriving, and large-scale tweet streams. Here, we make survey of various approaches for tweet summarization.

Keywords: Tweet stream, continuous summarization, tweet clustering, summary, timeline.

I. INTRODUCTION

The micro blogging site started in 2006 has become great popularity such as Twitter, facebook etc. This is resulted in explosion of the amount of short text messages. In February 2011, Twitter had 200 million registered users and 25 billion tweets in all of 2010. In this majority of post most of conversational or not meaningful, about 3.6% of the posts concern topics of mainstream news. Tweets, in their raw form, while being informative, can also be immense. The searching for a hot topic may yield millions of tweets, spanning weeks. The user unnecessarily goes through the millions of tweets and it is impossible every time. For this there is one solution namely filtering. Even if filtering is allowed, plowing for important contents, through such large amount of tweets is also very difficult and hard to possible task. This is happen because of enormous amount of irrelevant tweets. Another possible solution for information overload problem is summarization. The summarization is used to help what exactly the contents are conveying.

Summarization is the process of reducing a text document with a computer program for creating a summary that contains the only important points of the original document. The problem of information overload is increases, and because of the quantity of data is increasing, there is a necessity automatic summarization. This technology makes use of a coherent summary such as length, style of writing and syntax. Machine learning and data mining in which automatic data summarization is a very important area. These summarization technologies are widely used today, in a large number of micro blogging industries. Here are some examples of search engines in which summarization techniques are used such as Twitter, Facebook, and Google etc. Other category includes document summarization, image collection summarization and video summarization. The main idea behind summarization is to search a representative and common subset of the data, which represent unique information of the entire set.

Document summarization, tries to automatically create a representative summary or abstract of the entire document, by finding the most informative sentences. Similarly, in image summarization the system finds the most representative and important (or salient) images. For tweet summarization mostly document summarization technique is used.

Two types of automatic summarization approaches: extraction and abstraction. The extractive summarization identifies relevant sentences that belong to the summary. In extraction based summarization task, the automatic system extracts objects from the entire collection, without modifying the objects itself. Examples of this include key phrase extraction, where the goal is to select individual words or phrases to "tag" a document and The goal of document summarization is to select whole sentences (without modifying them) to create a short paragraph summary. Similarly, in image collection summarization, the system extracts images from the collection without modifying the images themselves. On the other hand, abstraction based summarization task, involves paraphrasing sections of the source document. In general, abstraction can condense a text more strongly than extraction, but the programs which can do this are harder to develop as they require the use of natural language generation technology, which itself is a growing field. The grouping of similar tweets forms different clusters. These clusters used for summarization of tweet streams. Summarizing is defined as reduces the size of contents and indicate which particular topic is discussed on social sites. Top tweets are found out from clusters by using ranking algorithm.

Traditional document summarization techniques are not effective for big size tweets as well as not suitably applicable for tweets which are arrived fast and continuously. To overtake this problem tweet summarization is requires which should have new functionality significantly different from traditional

Copyright to IJARCCE

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

summarization. Tweet summarization has to take into consideration the temporal feature of the arriving tweets.
Consider example of Apple tweets [10]. A tweet summarization system will monitor Apple related tweets which are produced a real-time timeline of the tweet stream. Given a timeline range, the document system may generate a series of current time summaries to highlight points where the topic or subtopics evolved in the stream. Such a system will effectively enable the user to learn major news or discussion related to Apple without having to read through the entire tweet stream.

II. LITERATURE REVIEW

Tweet summarization includes two steps. First step requires tweet data clustering and then actually summarization is performed.
Algorithm for stream data clustering has been widely studied by various authors in literature. BIRCH is the balanced iterative reducing and clustering using hierarchies' algorithm. This algorithm is an unsupervised data mining algorithm [6]. It is used to perform hierarchical clustering over large data-sets. An advantage of BIRCH is that, it has ability to make clusters in increment and dynamic manner. This algorithm handles noise effectively and suitable for large databases. It makes cluster of incoming and multi-dimensional metric data points, to produce the best quality clustering for a given set of resources such as memory and time constraints.
Bradley proposed a scalable clustering approach, which stores only important portions of the data with compressing or discarding other portions of the data which is not useful. This framework of clustering is based on the concept that effective clustering solutions is obtained by selectively storing important portions of the data and summarizing other portions of the data. The size of prespecified memory buffer which is allowable determines the amount of summarizing and required internal book-keeping. Author assumes that an interface to the database allows the algorithm to load number of data points requested. Data compression represents group of points by sufficient statistics. The interface to database allows the algorithm to load number of data points. These are obtained from a sequential scan, a random sampling or any means provided by the database engine [7].
CluStream is one of the most typical stream clustering methods. It having online micro-clustering component and also offline macro-clustering component. Online micro-clustering component require efficient process to store summaries. Offline components use only summary statistics. The pyramidal time frame was also proposed by authors to recall historical micro clusters for different time durations [5].
Here are some document summarization approaches are explained. Random Summarizer is an approach which randomly selects k posts or each topic as summary. This method was useful in order to provide worst case performance and also set the lower bound of performance [3]. Most Recent Summarizer approach chooses the most recent k posts as a summary from the selection pool. It is able to choose the first part of a news article as summary. This approach is implemented because the intelligent summarizers cannot perform better than simple summarizer. This summarizer only uses the first part of the document as summary [3].
LexRank summarizer uses a graph based method. It detects pairwise similarity between two sentences or between two posts. It makes the similarity score that is the weight of the edge between the two sentences. The final score of a post is computed based on the weights of the edges that are connected to each other. This summarizer is helpful to provide summarization based on baseline for graph instead of direct frequency summarization. Though it does depend on frequency, this system uses the relationships among sentences to add more information. LexRank Algorithm provides better view of important sentences. This is more complex algorithm than frequency based algorithm [1]. TextRank summarizer [2] is another graph based method. This approach uses the PageRank algorithm. This provided another graph based summarizer which incorporates potentially more information than LexRank. This is happens because it recursively changes the weights of posts. The final score of each post is dependent on how it is related to immediately connected posts as well as the way in which posts are related to other posts. TextRank algorithm is graph based approach used to find top ranked sentences. TextRank includes the whole complexity of the graph rather than just pair wise similarities. ETS (Evolutionary Timeline Summarization) [9] generates timelines for large amount of data. ETS gives evolutionary trajectories on particular dates. The advantage is that it facilitates fast news browsing.
SPUR that is Summarization via Pattern Utility and Ranking is a novel algorithm used to summarize a batch of transactions with low compression ratio and high quality. It is working in a high scalable fashion. Xintian Yang et al. also develop D-SPUR which is the dynamic version of SPUR. D-SPUR is enhanced and modifies the pyramidal time window in data streams. SPUR and D-PUR algorithm compress messages with low compression ratio, high quality and fast running time [4].
Twitter streams also used for event summarization to represent information in live manner. The participant based approach is used for event summarization. The key components used for summarization are Participant Detection, Sub-event Detection and Summary Tweet Extraction. Participant detection identifies event participants then identify sub-events related to participants. The tweets are extracted from sub-events using Summary Tweet Extraction component [8].
Zhenhua Wang et al. introduce a summarization framework called Sumblr. This is the continuous summarization by stream clustering. Continuous summarization is difficult task as it contains large number of meaningless and irrelevant tweets. This is the first which studied continuous tweet stream summarization. This framework consists of three main components, namely the Tweet Stream Clustering module, the High-level Summarization module and the Timeline Generation module. Sumblr is useful to work on dynamic, fast arriving, and large-scale tweet streams [10].

DOI 10.17148/IJARCCE.2015.412122

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

III. PROPOSED WORK

Implementing continuous tweet stream summarization is not an easy task, since a large number of tweets are meaningless, irrelevant and noisy in nature, due to the social nature of tweeting. Tweets are strongly correlated with their posted time and new tweets tend to arrive at a very fast rate. Tweet streams are always very large in scale; hence the summarization algorithm should be highly efficient. It should provide tweet summaries of arbitrary time durations. It should automatically detect sub-topic changes and the moments that they happen. The previous version of sumbler was not effective in distributed area. Here we are going to develop a multi topic version of a continuous tweet stream summarization framework, namely Sumbler to generate summaries and timelines in the context of streams, which will also suitable in distributed systems and evaluate it on more complete and large scale data sets.

IV. ARCHITECTURAL VIEW

As shown in Fig. 1, the sumbler [10] framework consists of three main modules: the tweet stream clustering module, the high-level summarization module and the timeline generation module. The tweet stream clustering module maintains the online statistical data. The incremental clustering is used to maintain tweets in online fashion. The topic-based tweet stream is given, it is able to efficiently cluster the tweets and maintain compact cluster information. The high-level summarization module provides two types of summaries: online and historical summaries. An online summary describes what is currently discussed among the public. Thus, the input for generating online summaries is retrieved directly from the current clusters maintained in memory. On the other hand, a historical summary helps people understand the main happenings during a specific period, which means we need to eliminate the influence of tweet contents from the outside of that period. As a result, retrieval of the required information for generating historical summaries is more complicated. The core of the timeline generation module is a topic evolution detection algorithm which produces real time and range timelines in a similar way.

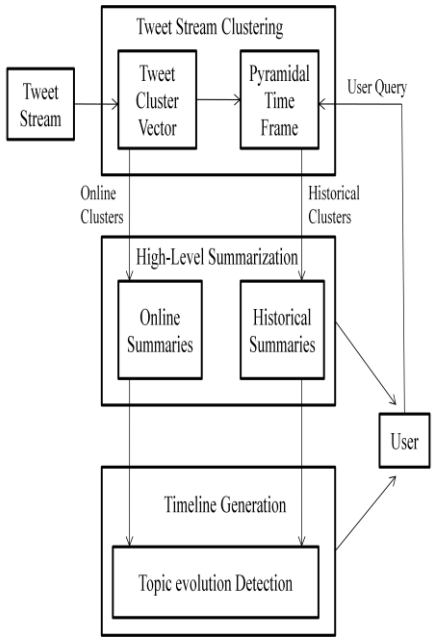

Fig1. Architectural View [1]

V. CONCLUSION

Here we studied various approaches for document summarization such as filtering and tweet summarization. These approaches are used for managing huge amount of tweets. Filtering is not an efficient approach because of tweet data is noisy and redundant. Because of the summarization is used to summarize the tweet data. Traditional document summarization techniques are not effective for big size tweets as well as not suitably applicable for tweets which are arrived fast and continuously, also they are not focus on static and small-scale data set. To overcome this problem, develop a multi topic version of a continuous tweet stream summarization framework, namely Sumbler to generate summaries and timelines in the context of streams, which will also suitable in distributed systems and evaluate it on more complete and large scale data sets, which deals with dynamic, fast arriving, and large- scale tweet streams. This will discovers the changing dates and timelines dynamically during the process of continuous summarization. Moreover ETS (Evolutionary Timeline Summarization) does not focus on efficiency and scalability issues which are very important in our streaming context.

REFERENCES

[1] G. Erkan and D. Radev, "Lexrank: graph-based centrality as salience in text summarization," Journal of Artificial Intelligence Research, vol. 22, pp. 457–480, 2004.
[2] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," in EMNLP. Barcelona: ACL, 2004, pp. 404–411.
[3] David Inouye, Jugal K. Kalita, "Comparing Twitter Summarization Algorithms for Multiple Post Summaries", IEEE Trans. Knowl. Data Eng., 23(8):1200–1214, 2011
[4] Xintian Yang, Amol Ghoting,, "A Framework for Summarizing and Analyzing Twitter Feeds", In KDD'12, August 12–16, 2012, Beijing, China.
[5] 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.
[6] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," in Proc. ACM SIGMOD Int. sConf. Manage. Data, 1996, pp. 103–114.
[7] 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.
[8] C. Shen, F. Liu, F. Weng, and T. Li, "A participant-based approach for event summarization using twitter streams," in Proc. Human Lang. Technol. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics, 2013, pp. 1152–1162.
[9] R. Yan, X. Wan, J. Otterbacher, L. Kong, X. Li, and Y. Zhang, "Evolutionary timeline summarization: A balanced optimization framework via iterative substitution," in Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2011, pp. 745–754.
[10] Zhenhua Wang, Lidan Shou, Ke Chen, "On Summarization and Timeline Generation for Evolutionary Tweet Streams", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 5, MAY 2015.