INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 4 ISSUE 2 (2018) PAGES 2802 - 2805
RECEIVED : 05.02.2018 PUBLISHED : 18.03.2018
March 18, 2018

CLUSTERING BASED SUMMARIZATION ON TOPIC EVOLUTIONARY TWEET STREAMS

[1]Deepa G, M.Phil, Research Scholar, K.M.G College Of Arts & Science, Gudiyattam.

[2]Prof.P.Anjugam, Assistant Professor, PG & Research Department Of Computer Science & Applications, K.M.G College Of Arts & Science, Gudiyattam.

Abstract:

At an unprecedented rate, short-text messages such as tweets are being created and shared. While being informative, can also be overwhelming, tweets, in their raw form. It is a nightmare to plow through millions of tweets which contain enormous amount of noise and redundancy, for both end-users and data analysts. We propose a novel continuous summarization framework called Sumblr to alleviate the problem, in this paper. Sumblr is designed to deal with dynamic, fast arriving, and large- scale tweet streams, in contrast to the traditional document summarization methods which focus on static and small-scale data set. Our proposed framework consists of three major components.

Keywords: Sumblr, Dynamic, Nightmare.

1. INTRODUCTION

Rising popularity of microblogging services such as Twitter, Weibo, and Tumblr has resulted in the explosion of the amount of short-text messages. Twitter, for instance, which receives over 400 million tweets per day1 has emerged as an invaluable source of news, blogs, opinions, and more. While being informative, can also be overwhelming, tweets, in their raw form. Twitter may yield millions of tweets, spanning weeks, for instance, search for a hot topic. Plowing through so many tweets for important contents would be a nightmare, not to mention the enormous amount of noise and redundancy that one might encounter, even if filtering is allowed. and have diversity among the sentences to reduce redundancy, summarization represents a set of information by a summary consisting of several sentences. Specially when users surf the internet with their mobile devices which have much smaller screens than PCs, summarization is extensively used in content presentation. However, are not as effective in the context of tweets given both the large volume of tweets as well as the fast and continuous nature of their arrival, traditional document summarization approaches. Tweet summarization, therefore, requires functionalities which significantly differ from traditional summarization. In general, tweet summarization has to take into consideration the temporal feature of the arriving tweets. Since a big number of tweets are worthless, irrelevant and noisy in nature, due to the social nature of tweeting, implementing stable tweet stream summarization is however not an easy task. Using an illustrative example of a usage of such a system, Let us illustrate the desired properties of a tweet summarization system. for example, tweets about “Apple”, consider a user interested in a topic-related tweet stream. A real-time timeline of the tweet stream, a tweet summarization system will constantly monitor “Apple” related tweets producing. A user may explore tweets based on a timeline. To highlight points where the topic/subtopics evolved in the stream, Given a timeline range, the summarization system may produce a sequence of time stamped summaries. To learn major news/discussion related to “Apple” without having to read through the entire

tweet stream, such a system will effectively enable the user. a user may decide to zoom in to get a more detailed report for a smaller duration, given the big picture about topic evolution about "Apple". To get additional details for that duration, the system may provide a drill-down summary of the duration that enables the user. To obtain a roll- up summary of tweets, a user, perusing a drill-down summary, may alternatively zoom out to a coarser range. The summarization system must support the following two queries: summaries of arbitrary time durations and real-time/range timelines, to be able to support such drill-down and roll-up operations. But also support a range of data analysis tasks such as instant reports or historical survey.

2. RELATED WORK

Clusters the data based on an in-memory structure called CF-tree instead of the original large data set. Bradley et al. [3] proposed a scalable clustering framework which selectively stores important portions of the data, and compresses or discards other portions.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. 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. TCVs are considered as potential sub-topic delegates and maintained dynamically in memory during stream processing. The second structure is the pyramidal time frame (PTF), which is used to store and organize cluster snapshots at different moments, thus allowing historical tweet data to be retrieved by any arbitrary time durations. 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. 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. We propose a continuous tweet stream summarization framework, namely Sumblr, to generate summaries and timelines in the context of streams. We design a novel data structure called TCV for stream processing, and propose the TCV-Rank.algorithm for online and historical summarization. We propose a topic evolution detection algorithm which produces timelines by monitoring three kinds of variations.Extensive experiments on real Twitter data setsdemonstrate the efficiency and effectiveness of our framework.

3. PROPOSED SYSTEM

We introduced a novel summarization structure called Sumblr (continuouS sUMmarization By stream cLusteRing). To the best of our knowledge, our work is the first to study continuous tweet stream summarization. The overall framework is depicted in Fig. 2. The framework consists of three main components, namely the Tweet Stream Clustering module, the High-level Summarization module and the Timeline Generation module. In the tweet stream clustering module, we design an efficient tweet stream clustering algorithm, an online algorithm allowing for effective clustering of tweets with only one pass over the data. This algorithm employs two data structures to keep important tweet information in clusters.

The first one is a novel compressed structure called the tweet cluster vector (TCV). TCVs are considered as potential sub-topic delegates and proyect dynamically in memory during stream processing. The second structure is the pyramidal time frame (PTF) [1], which is used to store and organize cluster snapshots at different moments, thus allowing historical tweet data to be retrieved by any arbitrary time durations. The high-level summarization module supports generation of two kinds of summaries: online and historical summaries. (1) To generate online summaries, we propose a TCV-Rank summarization algorithm by referring to the current clusters maintained in memory. This algorithm first computes centrality scores for tweets kept in TCVs, and selects the top-ranked ones in terms of content coverage and novelty. (2) To compute a historical summary where the user specifies an arbitrary time duration, we first retrieve two historical cluster snapshots from the PTF with respect to the two endpoints (the beginning and ending points) of the duration. Then, based on the difference between the two cluster snapshots, the TCV-Rank summarization algorithm is applied to generate summaries. The core of the timeline generation module is a topic evolution detection algorithm, which consumes online/historical summaries to produce real- time/range timelines. The algorithm monitors quantified variation during the course of stream processing. A large variation at a particular moment implies a sub-topic change, leading to the addition of a new node on the timeline. In our design, we consider three different factors respectively in the algorithm. First, we consider variation in the main contents discussed in tweets (in the form of summary). To quantify the summary based variation (SUM), we use the Jensen-Shannon divergence (JSD) to measure the distance between two word distributions in two successive summaries. Second, we monitor the volume-based variation (VOL) which reflects the significance of sub-topic changes, to discover rapid increases (or “spikes”) in the volume of tweets over time. Third, we define the sum-vol variation (SV) by combining both effects of summary content and significance, and detect topic evolution whenever there is a burst in the unified variation.

4. ANALYSIS

In the tweet summarization many tweets are repeated so using summarization we can avoid redundancy.The summarization consists of four issues efficiency, topic evolution, performance. Tweet streams or many messages of social site are very large in size so thesummarization algorithm is very efficient. Performance of summarization is very effective. We have proposed TCV rank summarization algorithm which is used for generating historical and online summaries. This algorithm selects the top rank tweets from the Tweet Cluster Vector (TCV), to generate historical and online summaries where user specifies random time duration. We retrieve cluster snapshots from the Pyramidal Time Frame (PTF) with respect to beginning and ending of time duration, based on two clusters TCV rank summarization algorithm generates summaries. Also we proposed Topic Evolution Detection algorithm which takes the input of already generated summaries to produce timeline. Also we are working on other social stream which include clustering, timeline generation, Topic evolution etc. Now a day a socially generated stream has become popular on WWW (World Wide Web). As rapid growth in an internet, use of social media also increases. There are many social sites like Twitter, Facebook, Instagram etc. in which twitter has become one of the most popular social site for users to share information like text, audio, video etc. Short messages are being created and shared at massive rate. Twitter receives thousands of tweets per hour. It is in raw form, the solution for this is summarization of tweets. Summarization represents a set of document which contain summary of related data. We have proposed Tweet Cluster Vector (TCV) algorithm which

is used formaking cluster of those retrieved tweets among which summarization will takes place. Tweet Cluster Vector (TCV) algorithm includes two data structure to keep important tweet information in cluster. These data structures are tweet cluster vector and pyramidal time frame. TCVs are considered as potential sub topic representative and maintained dynamically during stream processing in memory.
The system architecture is for historical and online summarization of social stream‖. We have proposed TCV rank summarization algorithm which is used for generating historical and online summaries. This algorithm selects the top rank tweets from the Tweet Cluster Vector (TCV), to generate historical and online summaries where user specifies random time duration. We retrieve cluster snapshots from the Pyramidal Time Frame (PTF) with respect to beginning and ending of time duration, based on two clusters TCV rank summarization algorithm generates summaries. Also we proposed Topic Evolution Detection algorithm which takes the input of already generated summaries to produce timeline.

CONCLUSION

Also we are working on other social stream which include clustering, timeline generation, Topic evolution etc.In today's world, summarization becomes necessity of social stream as millions of information posted on social sites. It is the simplest way to understand exact information using summarization by avoiding redundancy and noisy data. 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.

REFERENCES

[1] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. "A framework forclustering evolving data streams." ACMSIGMOD Conference (2003): 81-92.
[2] Radev, G. Erkan and D. R. "LexRank: Graph-based lexical centrality as salience in text summarization." J. Artif. Int. Res. 22 (2004): 457– m479.
[3] L. Gong, J. Zeng, and S. Zhang. "Text stream clustering algorithm based on adaptive feature selection." Expert Syst. Appl. 38 (2011): 1393–1399.
[4] R. Yan, X. Wan, J. Otterbacher, L. Kong, X. Li, and Y. Zhang,. "Evolutionary timeline summarization: A balanced optimization framework via iterative substitution." 34th Int. ACM SIGIR Conf.Res. 2011. 745–754.
[5] J. Nichols, J. Mahmud, and C. Drews,. "Summarizing sporting events using twitter." ACM Int. Conf. Intell (2012): 189–198.
[6] Zhenhua Wang, Lidan Shou, Ke Chen, Gang Chen and Sharad Mehrotra. "on summarization and timeline generation forevolutionary tweet stream." IEEE 27 (2015): 1301-1315.
[7] N. A. Diakopoulos and D. A. Shamma, "Characterizing debate performance via aggregated twitter sentiment," in Proc. SIGCHI Conf. Human Factors Comput. Syst., 2010, pp. 1195–1198.
[8] M. Dork, D. Gruen, C. Williamson, and S. Carpendale, "A visual backchannel for large-scale events," IEEE Trans. Vis. Comput. Graph., vol. 16, no. 6, pp. 1129–1138, Nov. 2010.