INTERNATIONAL JOURNAL · ENGINEERING RESEARCH

International Journal for Innovative Engineering and Management Research

A Peer Revieved Open Access International Journal

www.ijiemr.org

COPY RIGHT

IJIEMR Transactions, online available on 11th Oct 2017. Link :http://www.ijiemr.org/downloads.php?vol=Volume-6&issue=ISSUE-9

Title: **TWEET STREAMS ONLINE SUMMARIZATION AND TIMELINE GENERATION**

Volume 06, Issue 09, Pages: 102– 107.

Paper Authors

PANGAL JAFFAR SADIQ, M VENKATESH NAIK , DR.G.PRAKASH BABU

St Mark Educational institution society group of institution, AP.

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

TWEET STREAMS ONLINE SUMMARIZATION AND TIMELINE GENERATION

PANGAL JAFFAR SADIQ[1], M VENKATESH NAIK[2] , DR.G.PRAKASH BABU[3]
[1]PG Scholar, CSE, St Mark Educational institution society group of institution, AP.
[2]Assistant Professor, CSE, St Mark Educational institution society group of institution, AP.
[3]Professor, CSE, St Mark Educational institution society group of institution, AP.

ABSTRACT:

Short-text messages such as tweets are being created and shared at an unprecedented rate. Tweets, in their raw form, while being informative, can also be overwhelming. For both end-users and data analysts, it is a nightmare to plow through millions of tweets which contain enormous amount of noise and redundancy. In this paper, we propose a novel continuous summarization framework called Sumblr to alleviate the problem. In contrast to the traditional document summarization methods which focus on static and small-scale data set, Sumblr is designed to deal with dynamic, fast arriving, and large-scale tweet streams. Our proposed framework consists of three major components. First, we propose an online tweet stream clustering algorithm to cluster tweets and maintain distilled statistics in a data structure called tweet cluster vector (TCV). Second, we develop a TCV-Rank summarization technique for generating online summaries and historical summaries of arbitrary time durations. Third, we design an effective topic evolution detection method, which monitors summary-based/volume-based variations to produce timelines automatically from tweet streams. Our experiments on large-scale real tweets demonstrate the efficiency and effectiveness of our framework.

INTRODUCTION

INCREASING 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. Tweets, in their raw form, while being informative, can also be overwhelming. For instance, search for a hot topic in Twitter may yield millions of tweets, spanning weeks. Even if filtering is allowed, 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. To make things worse, new tweets satisfying the filtering criteria may arrive continuously, at an unpredictable rate. One possible solution to information overload problem is summarization. Summarization represents a set of documents by a summary consisting of several sentences. Intuitively, a good summary should cover the main topics (or subtopics) and have diversity among the sentences to reduce redundancy. Summarization is extensively used in content presentation, specially when users surf the internet with their mobile devices which have much smaller screens than PCs. Traditional document summarization approaches, 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. 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. Let us illustrate the desired properties of a tweet summarization system using an illustrative example of a usage of such a system. Consider a user interested in a topic-related tweet stream, for example, tweets about “Apple”. A tweet summarization system will continuously monitor “Apple” related tweets producing a real-time timeline of the tweet stream. As illustrated in in this system, a user may explore tweets based on a timeline (e.g., “Apple” tweets posted between October 22nd, 2012 to November 11th, 2012). Given a timeline range, the summarization system may produce a sequence of times tamped summaries to highlight points where the topic/subtopics evolved in the stream. Such a system will effectively enable the user to learn major news/ discussion related to “Apple” without having to read through the entire tweet stream. Given the big picture about topic evolution about “Apple”, a user may decide to zoom in to get a more detailed report for a smaller duration (e.g., from 8 am to 11 pm on November 5th). The system may provide a drill-down summary of the duration that enables the user to get additional details for that duration. A user, perusing a drill-down summary, may alternatively zoom out to a coarser range (e.g., October 21st to October 30th) to obtain a roll-up summary of tweets. To be able to support such drill-down and roll-up operations, the summarization system must support the following two queries: summaries of arbitrary time durations and real-time/range timelines. Such application would not only facilitate easy navigation in topic-relevant tweets, but also support a range of data analysis tasks such as instant reports or historical survey.To this end, in this paper, we propose a new summarization method, continuous summarization, for tweet streams.Implementing continuous tweet stream summarization is however not an easy task, since a large number of tweets are meaningless, irrelevant and noisy in nature, due to the social nature of tweeting. Further, tweets are strongly correlated with their posted time and new tweets tend to arrive at a very fast rate. Consequently, a good solution for continuous summarization has to address the following three issues: (1) Efficiency—tweet streams are always very large in scale, hence the summarization algorithm should be highly efficient; (2) Flexibility—it should provide tweet summaries of arbitrary time durations. (3) Topic evolution—it should automatically detect sub-topic changes and the moments that they happen.Unfortunately, existing summarization methods cannot satisfy the above three requirements because: (1) They mainly focus on static and small-sized data sets, and hence are not efficient and scalable for large data sets and data streams. (2) To provide summaries of arbitrary durations, they will have to perform iterative/recursive summarization for every possible time duration, which is unacceptable. (3) Their summary results are insensitive to time. Thus it is difficult for them to detect topic evolution . In this paper, we introduce a novel summarization framework called Sumblr (continuouS sUMmarization By stream cLusteRing). To the best of our knowledge, our work is the first to study continuous tweet stream summarization.

EXISTING SYSTEM:

Tweets, in their raw form, while being informative, can also be overwhelming. For instance, search for a hot topic in Twitter may yield millions of tweets, spanning weeks. Even if filtering is allowed, 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. To make things worse, new tweets satisfying the filtering criteria may arrive continuously, at an unpredictable rate. Implementing continuous tweet stream summarization is however not an easy task, since a large number of tweets are meaningless, irrelevant and noisy in nature, due to the social nature of tweeting. Further, tweets are strongly correlated with their posted time and new tweets tend to arrive at a very fast rate.Unfortunately, existing summarization methods cannot satisfy the above three requirements because:

(1) They mainly focus on static and small-sized data sets, and hence are not efficient and scalable for large data sets and data streams.

(2) To provide summaries of arbitrary durations, they will have to perform iterative/recursive summarization for every possible time duration, which is unacceptable.

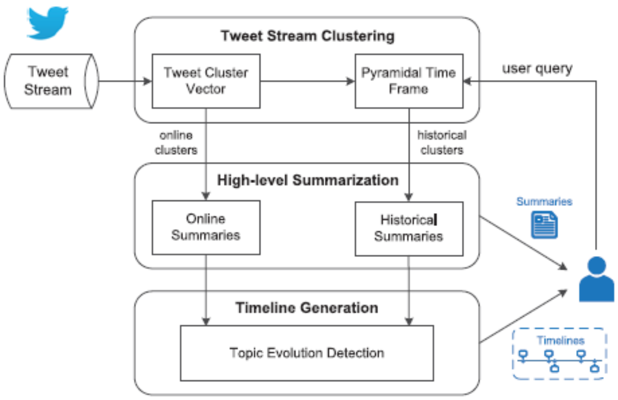
(3) Their summary results are insensitive to time. Thus it is difficult for them to detect topic evolution.

PROPOSED SYSTEM:

In this paper, we introduce a novel summarization framework called Sumblr (continuouS sUMmarization By stream cLusteRing). 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.The high-level summarization module supports generation of two kinds of summaries: online and historical 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.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 sets demonstrate the efficiency and effectiveness of our framework.

SYSTEM ARCHITECTURE:



IMPLEMENTATION

Admin

In this module, the Admin has to login by using valid user name and password. After login successful he can do some operations such as search history, view users, request & response, all topic messages and topics.

Search History

This is controlled by admin; the admin can view the search history details. If he clicks on search history button, it will show the list of searched user details with their tags such as user name, searched user, time and date.

Users

In user's module, the admin can view the list of users and list of mobile users. Mobile user means android application users.

Request & Response

In this module, the admin can view the all the friend request and response. Here all the request and response will be stored with their tags such as Id, requested user photo, requested user name, user name request to, status and time & date. If the user accepts the request then status is accepted or else the status is waiting.

Topic Tweet Messages

In this module, the admin can view the messages such as emerging topic messages and Anomaly emerging topic messages. Emerging topic messages means we can send a message to particular user. Anomaly emerging topic message means we can send message on a particular topic to all users and find the tweet stream clustering based on the topic by the end users, time line tweet streaming between two dates.

User

In this module, there are n numbers of users are present. User should register before doing some operations. And register user details are stored in user module. After registration successful he has to login by using authorized user name and password. Login successful he will do some operations like view or search users, send friend request, view messages, send messages, anomaly messages and followers.

Search Users

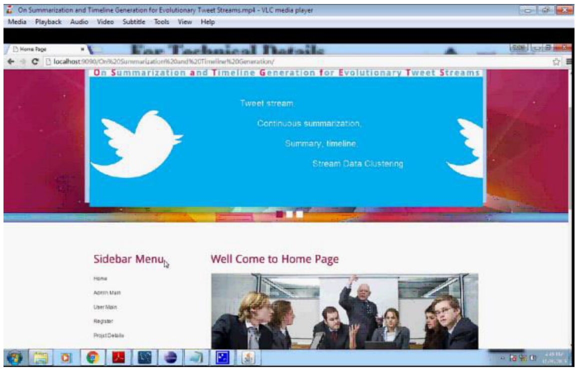
The user can search the users based on users and the server will give response to the user like User name, user image, E mail id, phone number and date of birth. If you want send friend request to particular receiver then click on follow, then request will send to the user.

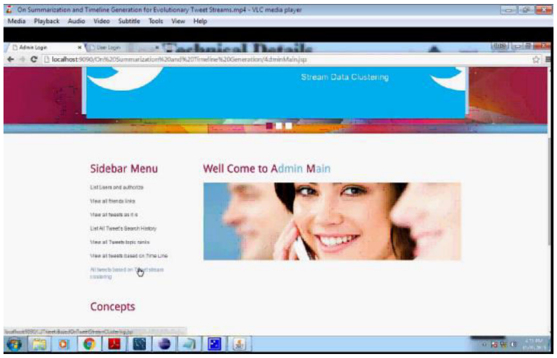
Messages

User can view the messages, send messages and send anomaly messages to users. User can send messages based on topic to the particular user, after sending a message that topic rank will be increased. Then again another user will also re-tweet the particular topic then that topic rank will increases. The anomaly message means user wants send a message to all users.

Followers

In this module, we can view the followers' details with their tags such as user name, user image, date of birth, E mail ID, phone number and ranks.


On Summarization and Timeline Generation for Evolutionary Tweet Streams.mp4 - VLC media player
Media Playback Audio Video Subtitle Tools View Help
Home Page
For Technical Details
localhost:9090/On%20Summarization%20and%20Timeline%20Generation/
On Summarization and Timeline Generation for Evolutionary Tweet Streams
Tweet stream
Continuous summarization
Summary, timeline
Stream Data Clustering
Sidebar Menu
Home
Admin Main
User Main
Register
Projct Details
Well Come to Home Page


On Summarization and Timeline Generation for Evolutionary Tweet Streams.mp4 - VLC media player
Media Playback Audio Video Subtitle Tools View Help
Admin Login
User Login
Technical Details
localhost:9090/On%20Summarization%20and%20Timeline%20Generation/AdminMain.jsp
Stream Data Clustering
Sidebar Menu
List Users and authorize
View all friends links
View all tweets as it is
List All Tweet's Search History
View all Tweets topic ranks
View all tweets based on Time Line
All tweets based on Tweet stream clustering
Concepts
Well Come to Admin Main

CONCLUSIONS

We proposed a prototype called Sumblr which supported continuous tweet stream summarization. Sumblr employs a tweet stream clustering algorithm to compress tweets into TCVs and maintains them in an online fashion. Then, it uses a TCV-Rank summarization algorithm for generating online summaries and historical summaries with arbitrary time durations. The topic evolution can be detected automatically, allowing Sumblr to produce dynamic timelines for tweet streams. The experimental results demonstrate the efficiency and effectiveness of our method. For future work, we aim to develop a multi-topic version of Sumblr in a distributed system, and evaluate it on more complete and large-scale data sets.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] Q. He, K. Chang, E.-P. Lim, and J. Zhang, "Bursty feature representation for clustering text streams," in Proc. SIAM Int. Conf. Data Mining, 2007, pp. 491–496.

[6] J. Zhang, Z. Ghahramani, and Y. Yang, "A probabilistic model for online document

clustering with application to novelty detection," in Proc. Adv. Neural Inf. Process. Syst., 2004, pp. 1617–1624.

[7] S. Zhong, "Efficient streaming text clustering," Neural Netw., vol. 18, nos. 5/6, pp. 790–798, 2005.

[8] C. C. Aggarwal and P. S. Yu, "On clustering massive text and categorical data streams," Knowl. Inf. Syst., vol. 24, no. 2, pp. 171–196, 2010.

[9] R. Barzilay and M. Elhadad, "Using lexical chains for text summarization," in Proc. ACL Workshop Intell. Scalable Text Summarization, 1997, pp. 10–17.

[10] W.-T. Yih, J. Goodman, L. Vanderwende, and H. Suzuki, "Multidocument summarization by maximizing informative contentwords," in Proc. 20th Int. Joint Conf. Artif. Intell., 2007, pp. 1776–1782.

[11] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," J. Artif. Int. Res., vol. 22, no. 1, pp. 457–479, 2004.

[12] D. Wang, T. Li, S. Zhu, and C. Ding, "Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization," in Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2008, pp. 307–314.

[13] Z. He, C. Chen, J. Bu, C. Wang, L. Zhang, D. Cai, and X. He, "Document summarization based on data reconstruction," in Proc. 26th AAAI Conf. Artif. Intell., 2012, pp. 620–626.

[14] J. Xu, D. V. Kalashnikov, and S. Mehrotra, "Efficient summarization framework for multi-attribute uncertain data," in Proc. ACM SIGMOD Int. Conf. Manage., 2014, pp. 421–432.