ISSN: 2349 - 4891

International
Journal of Recent Research and Applied Studies
(Multidisciplinary Open Access Refereed e-Journal)

Clustering Based Summarization on Topic Evolutionary Tweet Streams

M.Arunkumar[1], K.Arulanandam[2] & N.Suresh[2]

[1]*Research Scholar, Department of Computer Science, Government Thirumagal Mills College, Gudiyattam, Tamilnadu, India.*
[2]*Assistant Professor, Department of Computer Science, Government Thirumagal Mills College, Gudiyattam, Tamilnadu, India.*

Received 20th August 2016, Accepted 10th September 2016

Abstract

Twitter is a massively popular social network website that allows users to send short messages to the general public or a set of acquaintances. The topic of these messages is range from news items to notes of a more personal nature. General, the tweet to make a summarise of and third to detecting and the monitors of the summary – based to the volume based variation to produce timeline automatically from tweet stream. Implementing continuous tweet stream reducing a text document is however not a simple task, (ie) A huge number of the tweets are worthless, nonrelated and the raucous in nature of thedue to the social nature of tweeting. Collecting tweets and extracting information from them could be very valuable in many areas including market analysis and political research. In some cases, tweets have even been used to detect where earthquakes have recently occurred. Extracting useful information from Twitter is a very challenging endeavor. This research compares using traditional clustering techniques to a simpler statistical analysis in order to group common tweets for further analysis. The research shows that the statistical approach finds a solution much quicker than a traditional clustering approach, and has similar cluster quality. At a minimum, the statistical based methods used in this research could be used to determine the number of clusters used in a traditional clustering solution and summarization of tweet streams.

Keywords: Clustering, Tweet Streams, Algorithm.

Introduction

Internet enabled social networks have been around for nearly two decades. However, the landscape of social networking has changed dramatically over the years. The Internet as grown from the simple static information sharing network to a complex high speed network is designed to deliver data in real-time and on demand. The real-time nature of the Internet has given social networking websites the ability to provide information exchange among users as events are unfolding. Exchange the information as events are unfolding is not a new concept on the Internet. Much of the information exchanged on social networking sites is conversational in nature. Conversational capabilities have been part of the Internet since its inception. The Simple Mail Transfer Protocol (SMTP) was created to allow electronic mail to be exchanged over interconnected networks. The early Internet had other tools to exchange messages, such as the UNIX utilities "talk" and "write". Gradually, new forms of information sharing began to emerge. Near real-time instant messaging services, such as AOL Instant Messenger and ICQ2 was started gaining popularity in the mid to late 1990's. As the ability to have conversations and share information with people began to mature, social networking websites started to include these capabilities.

As the Internet began to boom, so did social networking. Early sites, like classmates. Com was built to simply allow people to connect with one another. A social networking became most popular, and more interactive sites like myspace.com and the facebook.com have been introduced. These websites include the ability to send instant messages to friends and post statuses about one. Some sites allow posting journals or logs. These web-logs, or blogs, allow followers to receive updates when new entries have been created by the author. People blog opinions, like news and any other even tutorials on a variety of topics. Sometime in the mid 2000's a new form of web-logging, known as micro-blogging, began to emerge. The term "microblog" was coined3 to refer to short sentences, links or individual images that could be exchanged with followers of a blog.

Correspondence
M.Arunkumar
E-mail:arunkumarasho01@gmail.com, Ph. +9191597 78122

Overall System Architecture

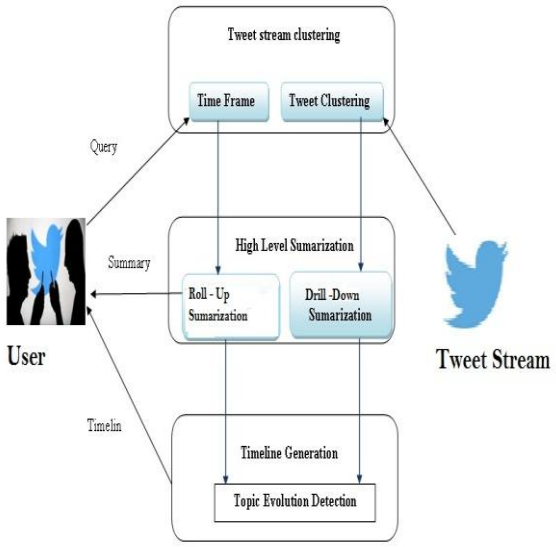

Figure I. Proposed System Architecture

The high-level summarization module provides two types of summaries: Drill -Down and Roll -Up summaries. The Drill -Down summary describes what is currently discussed among the public. Thus, the input for generating Roll -Up 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 need to eliminate the influence of tweet contents from the outside of that period.[1]. The result, get back of the required message for generating and historical summarizing is more complicated, and this shall be focus in the following discussion. Suppose the length of a user - defined time duration is H, and the ending timestamp of the duration is use.

Proposed Approach

Algorithm

The basic function of algorithm 1 is to take a set of tweets and divide them into two new distinct sets. Line 2 defines a function that takes a set of tweets to be divided, a set of terms in the tweets (ordered by descending number of occurrences) and a stopping threshold that defines the minimum number of terms allowed to continue processing

```
clusters ()
        function process(tweets terms minTweets)
 oldTerms ()
         while |terms| > 0 and terms 6= oldTerms do
         retainedTweets tweets having terms[0]
        discardedTweets tweets not having terms[0]
 if |retainedTweets| > minTweets then
        if |discardedTweets| = 0 then
                terms ShiftLeft(terms, 1)
        end if
process(retainedTweets terms)
        else
        clusters[k] discardedTweets
                k k + 1
end if
if discardedTweets > minTweets and retainedTweets > minTweets then
        terms discarded Terms
else
        clusters[k] discardedTweets
        k k + 1
        terms ()
end if
end while
end function
```

The Zipfian Clustering Algorithm

Lines 5 through 22 will divide the tweets into two sets: the first set having the most common term and the second set not having the most common term. If the set having the most common term is larger than the stopping point, then the list of terms is shifted (popping of the most common term) and the retained tweets are processed into two new groups using the second most common term as the discriminator. The same process is used to divide the "discarded" (those tweets not having the most common term).[2] This process of dividing groups of tweets by most common term continues until the stopping point is achieved. The stopping point is the only tunable parameter in the process. For very large dataset a good value for this parameter can be rather high. For this research a value of 100 was used for the large datasets. For smaller datasets, a value near 40 or 50 appears to work best.

The Zipfian clustering algorithm produced some number of grouped tweets for each dataset. This number was used to select K for a hierarchal clustering. [3]The hierarchal clustering results completion of the clustering, and the Zipfian clustering results. The comparing the results from the deterministic Zipfian clustering algorithm as well as the results from hierarchal clustering are the tweets into the same number of clusters as the Zipfian clustering.

Classification of Tweet Streams

The main algorithms SAR (Speech Act Recognition) used in Twitter Summarization. SAR for Twitter texts helps in keyword or phrase extraction and summarization. Cue Words and Phrases Non-cue Words (Abbreviations, opinion words, vulgar words, emoticons, etc.). The timeline generation module of the particular topic and the evolution detection of algorithm, which consumes online/historical summaries to produce real-time/range timelines. A most variation of particular moment that implies a to the sub topic change and leading the addition of a new node on the timeline.

Tweet Summary Approaches

Drill –Down Approach

The drill-down approach should provide summaries of the duration that enables to the user get additional details for that particular day. A user, using a drill-down summaries, alternatively zoom out to the coarser range of the obtain a roll-up summary of tweets.

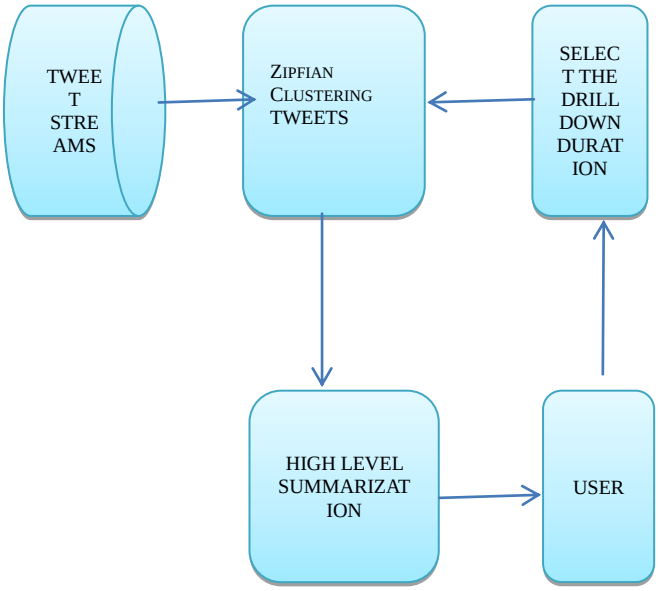

Figure II. Drill – Down Sumarization

Roll – Up Summarization

A roll-up summary field calculates tweets from related records, such as those in a related list. Create a roll-up summary field to display a tweet in a tweet cluster based on the tweet dataset detail record. (example) October 21st to October 30th), want to display particular days between the month or dates topic summarization amounts for all related tweets on that particular day.

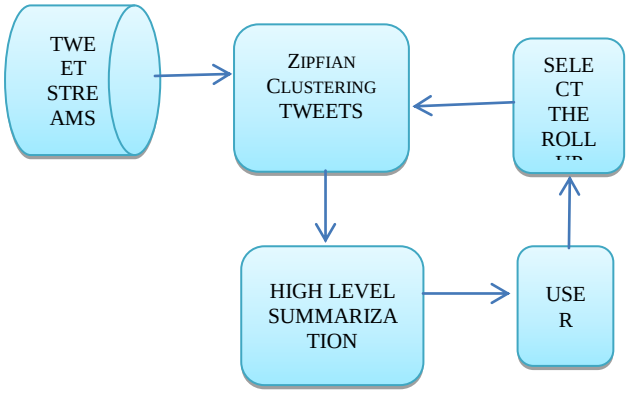

Figure III. Roll – Up Summarization

Timeline Generation

The demand for analyzing massive contents in social Medias fuels the developments in visualization techniques. Timeline techniques which can make analysis and the tasks easy and faster. A timeline based backchannel using conversations around the events. This timeline summarization computes evolution timelines similar to this method, which consists of a series of time - stamped summaries. In contrast, this method discovers the changing dates and generates timelines dynamically during the process of continuous summarization. Moreover, existing does not focus on efficiency and scalability issues, which are very important in this streaming context.[4]The several systems detected and the important moments of rapid increasesing of "spikes" to the status update the volume happen. The timeline generation developing an algorithm that based on the TCP and the detection, employed that slope based method to find spikes. After that, tweets like each moment are identified, and the word clouds or summariesing are selected. Different from this two – step approach, this method detects topic evolution and produces summaries/timelines in an online fashion.

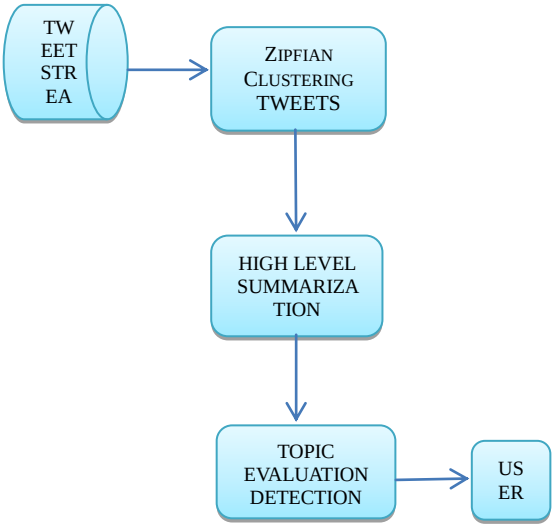

Figure IV. Timeline Generation

Experimental Results

The Zipfian clustering algorithm produced some number of grouped tweets for each dataset. This number was used to select K for a K-Means clustering and Tweet Cluster Vector (TCV). The K-Means results as well as the Zipfian clustering results.Comparing the results from the deterministic Zipfian clustering algorithm as well as the results to the kmeans to the TCV tweets to the clusters the Zipfian clustering.

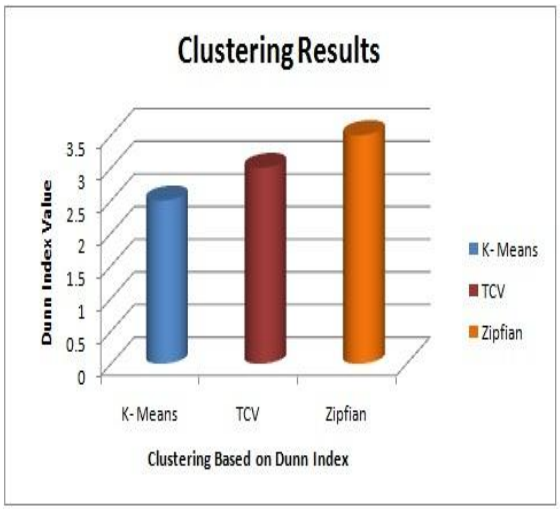

Figure V. Zipfian Clustering Results (Dunn Index).

The Timeline Focused on the User Activity. When providing the high-level summaries of the interface between the semi structured interview, that most users cited the timeline and event labels as the most memorable and helpful elements[5].The timeline also the most actively used for the component and it interface. Performance of the system is analyzed by comparing the solution generated by manually annotated event evolution graph and system generated event evolution graph results. Performance of system is measured using standard Investor *Relations* (IR) that divided into two types such as precision, recall Assuming the manually annotated of the set of event evolution of the relationships among the truth set O and the system generated by certain algorithms as A.

The precision is the ratio of the number of true and valid event evolution relationships retrieved by the system to the total number of tweet event evolution relationships retrieved by the system.

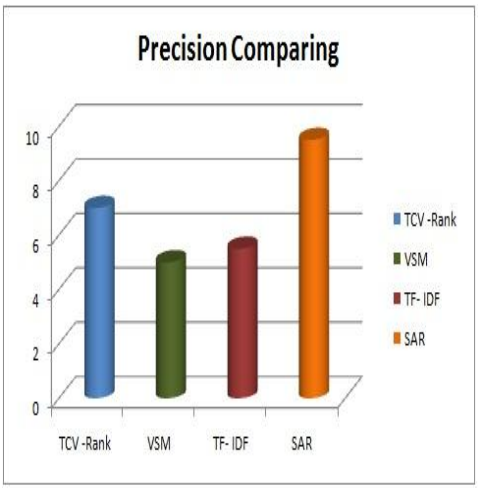

Figure VI. Precision Comparing between Methods

The recall is the ratio of the number of true and valid event evolution relationships retrieved the automatic system the total number of true and the valid and event evolution the relationships annotated manually.

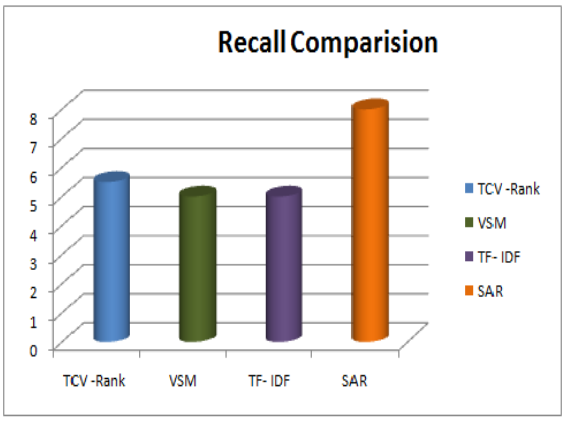

Figure VII. Recall Comparing between Methods

The Zipfian clustering algorithm (a statistical based approach) requires significantly less processing time and appears to produce equal or higher quality clusters and the results seem to more accurately represent the number of topics in the datasets. [6]From the values of recall and precision obtained for sample scenarios that conclude this approach gives better result than the existing methods is promising in producing an event evolution graph satisfactory precision and the recall about to the support user navigation and understanding of the development of events in a given tweet summarization .

Conclusions and Future Work

Conclusion

Clustering tweets and summarization has proven to be a major challenge. Both the dimension of the term vectors as well as their sparsity lead to high processing time as well as poor cluster performance. These problems are improved only marginally when reducing datasets by removing duplicate data as well as isolating by language. The clustering is used to the K-Means requires to the knowledge of the structure and the data a priori this itself presents a challenge with very large datasets. While some methods exist to automatically determine the best value for "K", these methods appear to fall short when applied to the short tweet texts. The statistical based Zipfian clustering algorithm appears to quickly identify appropriate values for "K", The grouped tweet are better performance between compared to the "K-Means" using the same value for "K". The SAR approach summarizing tweet streams with regard to topics along time line to produce an overview of topic evolution, which is expressed by sub-topics. A new initiative for Twitter topic summarization — speech act-guided to tweet post summarization. The problem of the propose is a set of

words based and the symbol-based on the features that can be easy harvested from raw data or free resources. Then results demonstrate how the algorithm predict and work efficiently.

Future Work

In future improve clustering performance is one that adds additional features to the tweets. Any tweets containing links to other pages could include the content of those pages as part of the document vectors. Tweet streams are having lot of temporal data like images and sounds that may be issue to summarize and clustering the tweet stream.

Intuition also suggests that clusters of tweets occur in a common timer period as well as a common geographic location. Including the date and time as well as originating location in the document vectors could improve clustering performance. It seems that these features would be more discriminating those simple terms, so a weighting mechanism should be chosen in order to appropriately represent this importance.

References

1. Zhenhua Wang, Lidan Shou, Ke Chen and Sharad Mehrotra," *On Summarization and Timeline Generation for Evolutionary Tweet Streams",*IEEE Transactions on Knowledge and Data Engineering (Volume:27, Issue: 5),2015.
2. Elena Baralis, Tania Cerquitelli, Silvia Chiusano, Luigi Grimaudo, Xin Xiao," Analysis of Twitter Data Using a Multiple-Level Clustering Strategy", Springer Heidelberg NewYork Dordrecht London, DOI:10.1007/978-3-642-41366-7-2013.
3. J. Nichols, J. Mahmud, and C. Drews," Summarizing Sporting Events using Twitter", Atlanta, Georgia, USA - 2012.
4. A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller, "*Twitinfo: Aggregating and visualizing micro-blogs for event exploration*," in Proc. SIGCHI Conf. Human Factors Comput. Syst. , 2011, pp. 227–236.
5. 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.
6. Beaux Sharifi, Mark-Anthony Hutton, and Jugal K. Kalita," Automatic Summarization of Twitter Topics", Proceeding of International Conference on Multimedia and Expo. ICME 2010.