Visualization Interface of Comment Streams from Social Network Services

Ms. Athira Sajeevan[1] Mrs. Jisha Mariyam John[2]
[1]M.Tech Student [2]Assistant Professor
[1,2]Department of Computer Science Engineering
[1,2]MBCCET Peermade, Idukki, Kerala, India

Abstract— This paper focus on generating a clear visualization interface of comment streams from Social Network Services. For each message on social network services, clients can express their conclusions by sending, giving a like, and leaving comments on it. The amount of comment is expansive, as well as the generation rate is also high. Clients may want to get a concise comprehension of a comment stream without analyzing the entire comment list, so we attempt to cluster comments with comparable substance together to produce a brief summary of message. A bunch form of short content rundown calculation (Batch STS) is first presented. Since particular clients will ask for the summary at any minute, a novel incremental grouping calculation called IncreSTS calculation can incrementally refresh comments with most recent comments continuously. IncreSTS has the benefits of high productivity, high versatility, and better dealing with anomalies. This paper uses two algorithms Batch STS and IncreSTS for efficient clustering of comments and focus on language processing by implementing a 4 Gram generation of comments in order to take semantic details of comments into account. Various NLP techniques are implemented such as stemming, redundant character removal and n gram generation for efficient short text summarization. Finally a clear visualization of the summary is generated.

Key words: Batch Clustering, Incremental Clustering, Natural Language Processing, Social Network Services

I. INTRODUCTION

Social Network Services (SNS) are common and have turned out to be imperative correspondence stage in our day by day life. VIPs, enterprises, and associations will make their own social pages to communicate with their fans and general society. Clients can respond to the social messages by giving likes, share and leaving comments. Comments for particular social messages are generated at a higher rate due to the popularity of social network services. It is time consuming and impossible to go over the whole comment list of each social message to understand what is being discussed on social medias. Summarization of comments can provide a brief understanding of social message. Summarization of the comments allows interaction at a higher level and can lead to an understanding of the overall discussion. Abstractive summarization is a very complex task. Several mechanisms and techniques were proposed to generate various types of summaries on comment streams. Clustering of comments can generate an efficient summary which involves the grouping of comments with similar contents together.

Moreover, the comments in SNS are generally short, and clients broadly make utilization of casual and unstructured writings that contain acronyms, shortening words, and so forth. This wonder builds the trouble of deciding the similarity between comments. The purpose of this assignment is to deliver a general summary speedily with the goal that clients can without much of a stretch get the diagram of a comment stream .Clustering involves the use of two algorithms, Batch STS and IncreSTS. The Batch STS algorithm is designed to cluster large data sets in batches but maintain the efficiency and quality.

Several experiments confirm that batch clustering algorithm for big data sets is more efficient in using computational power, data storage and results in better clustering. Since the users/clients may request the summary at any time real time generation of comments are needed. IncreSTS algorithm contributes for this. Besides this clustering algorithms the implementation of NLP technique can provide an efficient summary generation and a visualization interface can be provided for an at- a -glance representation of the generated summary.

II. RELATED WORK

Several researches and studies were done earlier to overcome the information overload problem and the problem of extracting useful knowledge from the large quantity of client generated data on SNS. This can be summarized into the following categories.

A. Human-Assisted Mechanisms

The principle idea of this mechanism is to feature significant comments by the help of client criticism and judgment. For example, clients can "like" a comment on well-known social sites, for example, YouTube, Facebook, and Amazon. The comments with more "likes" will be shown on the highest priority on the top of the list. Without a doubt, human judgment can create most right outcomes. However, since a client is probably not going to skim all comments and assess the goodness of everyone some comments which are more informative may be disregarded.

B. Sentiment Analysis

This category first applied a neural network model, namely the self- organizing maps, to cluster similar messages and sentiment keywords, respectively. Then it developed an association discovery process to find the associations between a message and some sentiment keywords. The sentiment of a message is then determined according to such associations. Sentiment analysis or opinion mining is a popular topic for social network applications since it is crucial for knowing the emotions of people for commercial and political purposes. For example, if the opinions of one person on some political issues is discovered, proper promotional materials can be send to the person to enhance or change his political preference.

C. Automatic Summarization

This category focus on a trend called micro-blogging, and in particular a site called Twitter that allows a huge number of users to release frequent short messages. Such sites contain a large number of small textual messages, posted by millions of clients, at random or in response to some events or situations. However, out of such random disorganization of messages usually large number of users post similar messages on similar topics. This can be discovered using statistical analysis of huge amount of posts. This consists of an algorithm that takes a trending phrase or any phrase as input specified by a user, collects a large number of posts containing the phrase, and provides an automatically created summary of the posts related to the term. It is possible to get a global view of the content of the text message repository in terms of a set of short summaries of trending terms during the course of a period of time such as an hour or a day.

III. EXISTING SYSTEM

To generate summary of comment stream from social network services such as Facebook, Twitter … etc several clustering strategies are existing. One among the recent mechanisms include Real time incremental short text summarization on comment streams using IncreSTSalgorithm[1]. Since new messages are appearing in social networks and different clients may request the summary in real time, IncreSTS a real time summarization algorithm can generate the summary incrementally. IncreSTS algorithm has the capability of generating incremental update and can produce the latest top-k clusters. Existing system works as follows. For a newly incoming comment the algorithm checks for an already existing cluster to which the comment can be added. If a match is found when mapping with the existing cluster contents the incoming comments are added. Otherwise a new cluster is created. The efficiency of IncreSTS is increased by adding a specific data structure for each cluster. From the comments important and identical terms are extracted to generate a key-term-cloud. This mechanism of generating summary possesses high efficiency performance. IncreSTS is fully parameter free and can handle outlier effectively.

IV. DISADVANTAGES

- Does not focus on traditional comment streams.
- Does not focus on natural language processing.
- Semantic constraints of comments are ignored.
- Problem of redundancy.
- Poor cluster quality.

V. PROPOSED SYSTEM

Earlier works were not centered on the traditional comment streams which are able to express complete information. Natural Language Processing techniques were ignored. The researches fully aims at generating summary of comment streams obtained at social network services. The quantity of comments may increase at higher rate after a social message is published. In order to generate an efficient short text summarization of comments the proposed system applies various NLP techniques to the whole set of comments. Thus the comments are preprocessed before generating clusters. Natural language processing systems take strings of words (sentences) as their input and produce structured representations capturing the meaning of those strings as their output. The nature of this output depends heavily on the task at hand. The application of NLP techniques led to the conversion each comment into a set of n-gram terms. The basic NLP techniques implemented includes the following.

A. Removing Punctuations:

Here the punctuations in the comments such as dot, exclamation, commas, star etc are removed. For eg. The comment “hai!!!!!!!!!!!!!!!!!!” will be transformed into “hai”.

B. Removing Redundant Characters:

In this the repeated characters from the comments are ignored. For eg. The comments like “helloiiiiiiiiiiiiiiiiiiiiiiiiii” will be termed as “hello”.

C. Stemming:

Stemming is the process of reducing inflected words to their word stem, base or root form generally a written word form. For example, should identify the string "cats" (and possibly "catlike", "catty" etc.) as based on the root"cat",and "stems","stemmer", "stemming", "stemmed" as based on "stem". A stemming algorithm reduces the words "fishing", "fished", and "fisher" to the root "fish. Instead of storing all form of a word a search engine can store only the stems, greatly reducing the size of index while increasing retrieval accuracy.

D. Vectorization:

Used for a vector model representation of comment streams.

E. N-gram generation:

An n-gram is a contiguous sequence of n items from a given sample of text or speech. These items, syllables, letters, words or base pairs according to the application. The n-grams typically are collected from a text or speech corpus. An n-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram" (or, less commonly, a "digram"); size 3 is a "trigram". English cardinal numbers are sometimes used, e.g. "four-gram", "five-gram", and so on. This paper implements a 4- gram model of generation.

After applying the NLP techniques to the comment streams the implementation of the two main algorithms BatchSTS and IncreSTS is done. Finally informative comments are extracted to generate the summary. Finally through a visualization interface an informative and impressive brief understanding of the comments are delivered to the clients.

VI. ADVANTAGES

- High quality and efficiency of clustering results.
- Remove comments with similar information.
- At –a-glance visualization of summary.
- Avoid information overload problem.
- Efficient handling of outliers.

VII. DESIGN & IMPLEMENTATION

The overall system is designed[Fig.2] as an effective solution for the real time short text summarization of comments from social network services .This paper center around the methods of natural language processing (contracted as NLP).[Fig.2]. With utilizing some fundamental NLP techniques, each comment is changed to an arrangement of n-gram terms. Then again, we characterize new closeness measures to enough decide the separation amongst comments and groups.

Both IncreSTS and BatchSTS[6]are applied for efficient clustering. Moreover, the most noteworthy preferred standpoint of our calculation is its high productivity, showing that it can create clusters comes about with most recent approaching comments continuously. These capacities unquestionably address the issue of comment stream rundown on SNS.

The main focus is to incorporate NLP [9] implementation for generating a summary of the comments that are obtained for a particular message on social network sites. As the generation rate of comments may increase at a higher rate real time generation of summary is necessary. The main execution of NLP module is demonstrated below.

After the preprocessing of comments the algorithms are implemented. The two main algorithms are IncreSTS & BatchSTS. Both algorithm aims at discovering the top-k clusters of opinions regarding a message on SNS. First the implementation of BatchSTS takes place.

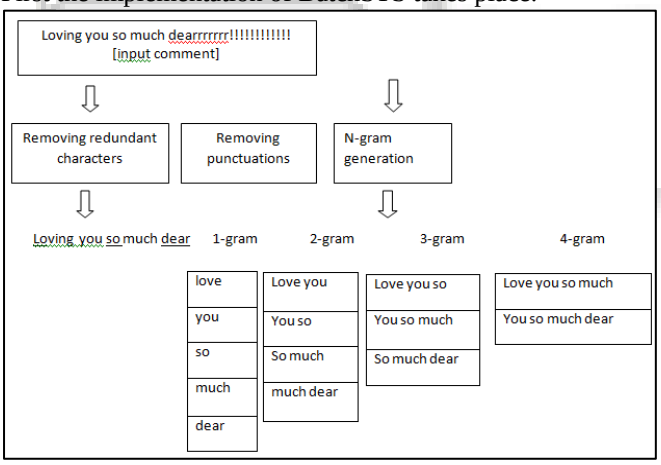

Fig. 1: NLP Module

A. *BatchSTS Algorithm*

This algorithm is the batch version of short text summarization. Batch STS takes the whole set of comments as the input. A radius threshold is used to determine the range of similarity of comments in a cluster. For each incoming comments we check whether there exists a cluster to which the incoming comments can be added. If there exist then the comment is added else a new cluster is created with a single comment. Again it is checked to find the existence of a cluster with similar comments. If found the clusters are merged together. Radius of each cluster should be less than the threshold value. If the criteria are met BatchSTS outputs the top-k clusters with top-k most comments.

Algorithm BatchSTS
Input: S: the comment set
θ_r: the radius threshold
Output: top-k clusters which have top-k most comments
1. Initialize $C = \emptyset$;
2. **for** each element v_i of S
3. **if** there exists any cluster C_j where $dis(v_i, C_j)$ is not infinite
4. Add v_i into anyone of these clusters;
5. **else**
6. Form a new cluster C_{new} with the comment v_i;
7. $C = C \cup C_{new}$;
8. **for** each non-single-point element C_i of C
9. **for** each non-single-point element C_j of C where $i \neq j$
10. **if** $dis(C_i, C_j)$ is not infinite
11. Merge C_i and C_j;
12. **for** each non-single-point element C_i of C
13. **while** the radius of C_i is larger than or equal to θ_r
14. **for** each comment v_j in C_i
15. **if** $dis(v_j, C_i) \geq \theta_r$
16. Exclude v_j from C_i;
17. Check whether v_j can be merged with other excluded comments;
18. Output top-k clusters in C which have top-k most comments;
End

Algorithm 1: Algorithmic form of BatchSTS

B. *IncreSTS algorithm*

Since different users request the summary during different time and the comments may increase at a higher rate BatchSTS will not be sufficient to generate a real time summary. Here comes the application of IncreSTS algorithm. In addition, since new messages rise up consistently, clients by and large just view the outline of a particular message once and won't backpedal to peruse the refreshed synopsis later on. In such a specific situation, to quickly deliver the most recent best k groups, we propose the IncreSTS calculation that has the capacity of incremental refresh. The essential idea of IncreSTS is to keep up the grouping aftereffect of the past stage, and to incrementally refresh the bunching result with the recently approaching comments.

Algorithm IncreSTS
Input: C: the set of previous clustering result
v_{new}: the newly-incoming comment
θ_r: the radius threshold
Output: top-k clusters which have top-k most comments
1. $C^a = \{C_i \mid C_i$ is an element of $C \cap dis(v_{new}, C_i)$ is not infinite$\}$;
2. $C^b = \{C_j \mid C_j$ is an element of $C^a \cap dis(v_{new}, C_j) < \theta_r\}$;
3. **if** C^b is not empty
4. Add v_{new} into C_{added} which have most comments in C^b;
5. Initialize $C_{changed} = \emptyset$;
6. **for** each element C_i of C^a where $C_i \neq C_{added}$
7. **for** each comment v_j in C_i
8. **if** $dis(v_j, C_{added}) < \theta_r$
9. Add v_j into C_{added};
10. Exclude v_j from C_i;
11. $C_{changed} = C_{changed} \cup C_i$;
12. **for** each element C_i of $C_{changed}$
13. **while** $V = \{v_j \mid dis(v_j, C_i) \geq \theta_r\}$ is not empty
14. Exclude all elements in V from C_i;
15. Try to add each comment in V into other clusters from large to small sizes;
16. **else**
17. Form a new cluster C_{new} with the comment v_{new};
18. $C = C \cup C_{new}$;
19. Output top-k clusters in C which have top-k most comments;
End

Algorithm 2: Algorithmic form of IncreSTS

At first we find a cluster to which the new incoming can be added. In line 1 of Algorithm 2, the separations amongst V new and every single existing group are figured. Among the clusters (in the set Cb) whose separations are littler than threshold radius, we pick the group Cadded that has generally comments. It is significant that V new isn't included into the group where dis(V new ; Cj) is the smallest. This is on the grounds that our foremost objective is finding top-k clusters where each gathering communicates comparative assessments (by applying the range limitation), however not limiting the inside group whole of squares. In light of this goal, the outline of the proposed new comment task is more reasonable for our objective application. In the next step again we check to determine whether there exist other clusters which can be merged.

With the plan of incremental clustering and information structure, IncreSTS has noteworthy proficiency execution. The extra cost of IncreSTS is the space utilized for saving to-date clustering comes about. Quickening the procedure by utilizing extra space is the quintessence of the proposed technique. The essential objective of this paper isn't to create groups with best quality, however to locate a neighborhood ideal in an extremely proficient manner. Such the plan idea significantly satisfies the prerequisites on the quick pace of SNS.

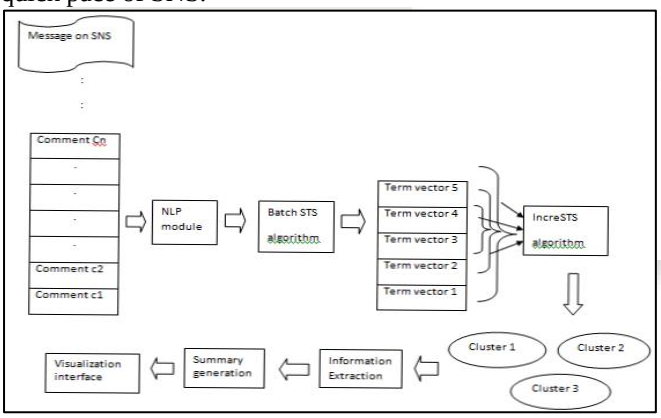

Fig. 2: System model of the proposed Frame work.

With the plan of incremental clustering and information structure, IncreSTS has noteworthy proficiency execution. The extra cost of IncreSTS is the space utilized for saving to-date clustering comes about. Quickening the procedure by utilizing extra space is the quintessence of the proposed technique. The essential objective of this paper isn't to create groups with best quality, however to locate a neighborhood ideal in an extremely proficient manner. Such the plan idea significantly satisfies the prerequisites on the quick pace of SNS.

VIII. VISUALIZATION INTERFACE

After the top-k clusters have been generated the summary should be presented to the clients. Residing within the vast amount of information our goal is to provide an at-a-glance visualization of comment streams from social network services. For each cluster a representative key-term cloud is constructed with the key terms. That is the final visualization or the generated summary consists of the cluster heads with comments added to the cluster heads and indicates the relevant term about the message which is more discussed. At –a-glance visualization interface enables to get a brief understanding of the summary which is more informative and impressive.

IX. CONCLUSION

In this paper, to empower the capacity of comment stream outline on SNS, we demonstrate the implementation of Natural Language Processing(NLP)that preprocess the comments and a novel incremental grouping issue and propose the calculation IncreSTS, which can incrementally refresh bunching comes about with most recent approaching remarks progressively. Initially the execution of BatchSTS algorithm takes place. With the yield of IncreSTS, we plan a perception interface comprising of fundamental data, key-term mists, and delegate comments .This initially introduction empowers clients to effortlessly and quickly get a review comprehension of a comment stream. From broad trial comes about and a genuine case exhibition, we check that IncreSTS has the benefits of high effectiveness, high adaptability, and better taking care of anomalies, which legitimizes the practicability of IncreSTS on the objective issue.

REFERENCES

[1] IncreSTS: Towards Real-Time Incremental Short Text Summarization on Comment Streams from Social Network Services-Cheng-Ying Liu, Ming-Syan Chen, Fellow IEEE, Chi-Yao Tseng.2016.

[2] H. Becker, M. Naaman, and L. Gravano, "Selecting quality Twitter content for events," in Proc. 5th Int. AAAI Conf. Weblogs Social Media, 2011, pp. 442–445.

[3] J. Bollen, H. Mao, and A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," in Proc.5th Int. AAAI Conf. Weblogs Social Media, 2011, pp. 450–453.

[4] D. Chakrabarti and K. Punera, "Event summarization using Tweets," in Proc. 5th Int. AAAI Conf. Weblogs Social Media, 2011 pp. 66–73

[5] S. Baccianella, A. Esuli, and F. Sebastiani, Batch Clustering Algorithm for Big Data Set in Proc. 31st Eur. Conf. IR Res. Adv. Inf. Retrieval 2014, pp.461–472.

[6] E. Shi, J. Bethencourt, H. Chan, D. Song, and A. Perrig, "Multi-Dimensional Range Query over Encrypted Data," Proc. IEEE Symp. Security and Privacy (S&P'07), pp. 350-364, May2007.

[7] J.-P. Mei and L. Chen, "Sum CR: A new subtopic-based extractive approach for text summarization," Knowledge. Inf. Syst., vol. 31, no. 3,pp. 527–545, 2012.

[8] M. Michelson and S. A. Maccskassy, "Discovering users' topics of interest on twitter: A first look," in Proc. 4th Workshop Analytics Noisy Unstructured Text Data, 2010, pp. 73–79.

[9] Baoshan Sun,Peng Zaho"Feature Extension for Chinese Short Text Classification Based on TopicaN-grams978-50904/17/ IEEEICIS 2017, May 24-26, 2017,

