ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 5, May 2018

Friend's Page

Bhagyashree Tatiya, Revati Deshmukh, Bhagyashree Tupe, Komal Pokharkar, Vishnu Kamble

Dept. of IT, P.E.S Modern College Of Engineering (Shivajinagar), Pune, Maharashtra, India

ABSTRACT: Now-a-Days, the popularity of social networking services has increased tremendously, so the quantity of comments can increase at a high rate immediately after a social message is published. Thus the system focuses on the problem of short text summarization on the comment stream of a particular message from social network services. The users of the social sites always desire to get a brief understanding of a comment stream without reading the whole comment list. So this system attempts to group comments with similar content together and generate a concise opinion summary for the message. Since different users can request the summary at any moment, existing clustering methods cannot be directly applied because they cannot meet the real-time need of such application. So this comment stream summarization problem is modeled as incremental clustering problem. This approach can incrementally update clustering results with latest incoming comments in real time. As a result visualization interface is generated that help users to rapidly get an overview summary.

KEYWORDS: Real-time short text summarization, incremental clustering, comment streams, social network services

I. INTRODUCTION

A social network is a social structure made up of a set of social actors (such as individuals or organizations) and a set of the dyadic ties between these actors. The social network perspective provides a set of methods for analyzing the structure of whole social entities as well as a variety of theories explaining the patterns observed from the structures. Due to the popularity and convenience of these platforms, celebrities, corporations, and organizations also set up social page: to interact with their fans and the public. Note that for each message, users are able to express their opinions by forwarding, giving a like, and leaving comments on it. Not only the quantity of comments is huge, but also the generation rate is remarkably high.

Users unnecessarily and almost impossibly go over the whole comment list of each message. However, we may still desire to know what are they talking about and what are the opinions of these discussion participants. Moreover, celebrities and corporations will have high interest to understand how their fans and customers reacting to certain topics and content.

With these motivations, inspired to develop an advanced summarization technique targeting at comment streams in social network services. Numerous studies and systems have proposed techniques and mechanisms to generate various types of summaries on comment streams. One major category aims to extract representative and significant comments from messy discussion.

Our proposed system does not focus on traditional comment streams that usually express more complete information. such as the discussion on products or movies. We target at comment streams in SNS that are it short text style with casual language usage.

For each social message, our main objective is to cluster comments with similar content together and generate a concise opinion summary for this message. We want to discover how many different group opinions exist and provide an overview of each group to make users easily and rapidly understand.

For instance, when lady Gaga uploads a photo to SNS, there are hundreds and thousands of comments given by her fans during a short period of time. Some of them may say that she is very beautiful and another group of fans may think that the outfit is too weird. Even more, some may particularly discuss the hair style of this photo.

Our goal is developing an efficient and effective technique to identify the clusters of these comments. Note that this problem is clearly different from existing research and possesses numerous unique characteristics and challenges.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 5, May 2018

Explore the problem of incremental short text summarization on comment streams from social network. We model this problem as an incremental clustering task and propose the Incremental Short Text Summarization algorithm to discover the top-k clusters including different groups of opinions towards one social message.

For each comment cluster, important and common terms will be extracted to construct a key-term cloud. This key-term cloud provides an at-a-glance presentation that users can easily and rapidly understand the main points of similar comments in a cluster. Moreover, representative comments in each group will also be identified.

II. LITERATURE SURVEY

Learning similarity metrics for event identification in social media [1] paper represents using this rich context, which includes both textual and non-textual features, we can define appropriate document similarity metrics to enable online clustering of media to events. Ensemble clustering is an approach that combines multiple clustering solutions for a document set. The classification based approach involves the selection of training examples from which to learn the similarity classifiers. The advantages of this modification to the ensemble similarity learning technique include improved efficiency via the use of centroids, providing for a more direct similarity metric computation. To produce high quality clustering results. The similarity metric learning techniques yield better performance than the baselines. Disadvantages are: The work focuses on event documents only, not non documents. It does not summarize event contents.

In the [2] paper represents to identify each event and its associated Twitter messages using an online clustering technique that groups together topically similar Twitter messages. Each identified event cluster, we select messages that best represent the event. We use centrality-based techniques to select messages that have high textual quality, strong relevance to the event, and, importantly, are useful to people looking for information about the event. Advantages are: To communicate Twitter content effectively. The re-ranking techniques that boost the centrality score of tweets with potentially useful features (e.g., URLs, tags). Automatically select event messages that are both relevant and useful. Disadvantages are: The sub-event content and topic analysis, such that multiple views or temporal variations not represent in the data.

The paper [3] represents Eddi is a novel interface for browsing Twitter streams that clusters tweets by topics trending within the user's own feed. An algorithm for topic detection and a topic-oriented user interface for social information streams such as Twitter feeds. (i) Benchmark TweeTopic against other topic detection approaches, and (ii) compare Eddi to a typical chronological interface for consuming Twitter feeds. Advantages are: A simple, novel topic detection algorithm that uses noun-phrase detection and a search engine as an external knowledge base. Eddi is more enjoyable and more efficient to browse than the traditional chronological Twitter interface. Disadvantages are: Users had access to our clients for a limited time, making it difficult to extrapolate conclusions on how the tool might be used longitudinally. Users were viewing the history of their feed rather than tweets they had never seen before, making our task slightly less realistic.

In [4] paper, measure the sentiment of each tweet using an extended version of the Profile of Mood States (POMS) and compare our results to a timeline of notable events that took place in that time period. We argue that sentiment analysis of minute text corpora (such as tweets) is efficiently obtained via a syntactic, term-based approach that requires no training or machine learning. Advantages are: The events in the social, political, cultural and economical sphere do have a significant, immediate and highly specific effect on the various dimensions of public mood. Long-term changes seem to have a more gradual and cumulative effect. To speculate that the social network of Twitter may highly affect the dynamics of public sentiment. Sentiment analysis techniques rooted in machine learning yield accurate classification results.

In [5] paper, we argue that for some highly structured and recurring events, such as sports, it is better to use more sophisticated techniques to summarize the relevant tweets. The problem of summarizing event-tweets and give a solution based on learning the underlying hidden state representation of the event via Hidden Markov Models. The advantage of leveraging existing query matching technologies and for simple one-shot events such as earthquakes it works well. The HMM is able to learn differences in language models of sub-events completely automatically. The disadvantage that SUMMHMM has to account for tweet words that only occur in some of the events, but not in others.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 5, May 2018

In social network consist of problems during when a user like Lady Gaga uploads a photo to social network services; there are hundreds and thousands of comments given by her fans during a short period of time. Some of them may say that she is very beautiful, and another group of fans may think that the outfit is too weird. Even more, some may particularly discuss the hair style of this photo. Therefore, our goal is developing an efficient and effective technique to identify the clusters of these comments.

The quantity of comments may increase at a high rate right after a social message is published. Moreover, distinct users will request the summary result at any moment. For these reasons, in order to immediately generate a summary based on the current comment stream, an incremental approach is preferable to meet the real-time needs of this application.

The comments in SNS are usually short, and users widely make use of informal and unstructured texts that contain acronyms, shortening words, etc. This phenomenon increases the difficulty of determining the similarity between comments. On the other hand, it is worth mentioning that instead of emphasizing on the quality of clustering, the most crucial point of this task is to produce a general summary promptly so that users can easily get the overview of a comment stream. It can be perceived that this problem is able to be modeled as a clustering task. However, traditional clustering methods have several inherent restrictions that cannot be directly applied here.

First, the computational complexity of existing methods is high, and they cannot straightforwardly adapt to satisfy the incremental need. Moreover, in this problem, defining the number of desired clusters in advance is unreasonable, which is required in many clustering algorithms. In addition, there will be a lot of outliers in a comment stream, meaning that without employing a good strategy for selecting initial cluster centers, existing method may be prone to poor results.

Disadvantages:

1. The difficulty of determining the similarity between comments is more.
2. Cluster identification is difficult one.
3. Computational complexity is high.
4. Due to sparse information contained in each comment, these approaches are not suitable for short text summarization especially when the number of comments is not large.

III. PROPOSED SYSTEM

In this paper, explore the problem of incremental short text summarization on comment streams from social network services. To model this problem as an incremental clustering task and propose the IncreSTS (standing for Incremental Short Text Summarization) algorithm to discover the top-k clusters including different groups of opinions towards one social message. For each comment cluster, important and common terms will be extracted to construct a key-term cloud. This key-term cloud provides an at-a-glance presentation that users can easily and rapidly understand the main points of similar comments in a cluster.

Representative comments in each group will also be identified. Our objective is to generate an informative, concise, and impressive interface that can help users get an overview understanding without reading all comments. Our proposed system depends on a fully incremental algorithm that is almost parameter-free and can handle the outlier problem. The most significant advantage of our algorithm is its high efficiency, indicating that it can generate clustering results with latest incoming comments in real time. These capabilities certainly meet the need of comment stream summarization on SNS.

To verify the effectiveness of IncreSTS algorithm, to collect real comment streams from Facebook and conduct extensive experiments with comparative methods to show the strength and superiority of our approach. Overall, the contributions of this paper can be summarized as follows.

- To model a novel incremental clustering problem based on the requirements of comment stream summarization on SNS.
- To propose IncreSTS algorithm that can incrementally update clustering results with latest incoming comments in real time.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 5, May 2018

- To design an at-a-glance presentation, which is concise, informative, and impressive, to help users easily and rapidly get an overview understanding of a comment stream.

Advantages:

I. High efficiency indicating that it can generate clustering results with latest incoming comments in real time.
II. High scalability and better handling outliers.
III. Justifies the practicability of Incremental Short Text Summarization on the target problem.

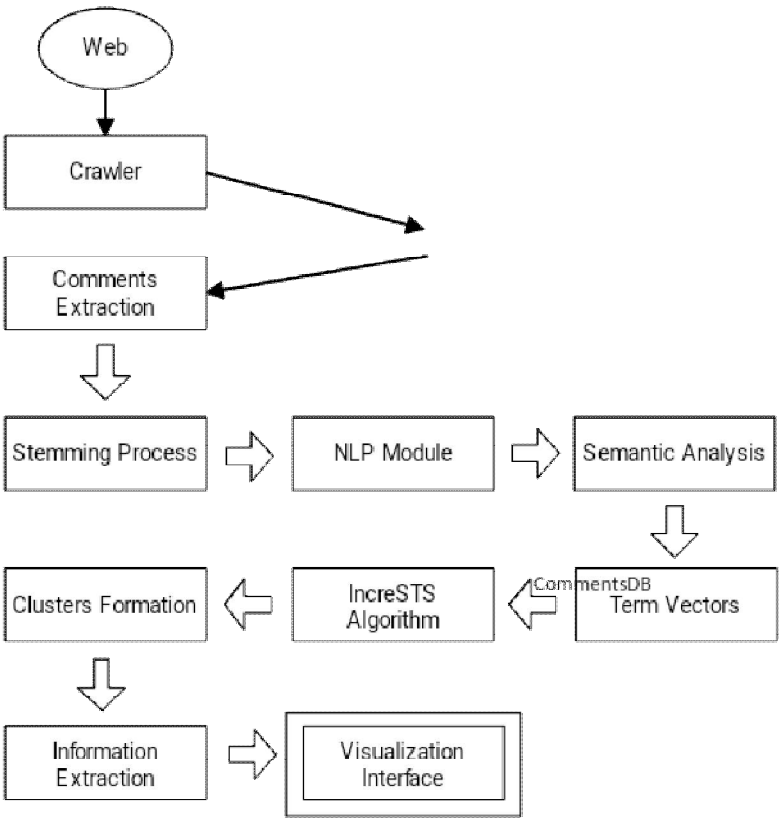

Fig.2 System model of the proposed framework.

IV. SYSTEM DESIGN

Consider two comments represented in the term vector model, $v_a = (t_{1,a}, t_{2,a}, \dots t_{N,a})$ and $v_b = (t_{1,b}, t_{2,b}, \dots t_{N,b})$. Each dimension corresponds to a separate term, and N is the number of dimensions. Since we define that the weights of terms are equal, if the term t_i occurs in the comment v_a , $t_{i,a}$ will be set to 1, otherwise, $t_{i,a}$ will be set to 0.
Thus, we define the modified cosine similarity of two comments as follows:

$$sim(v_a, v_b) = \{\frac{v_a * v_b}{D}, \quad if\ v_a.v_b \leq D, \quad 1, \qquad if\ v_a.v_b > D \quad (1)$$

Where, $v_a.v_b$ is the inner product of two vectors,
D is a positive integer constant.
The denominator of original cosine similarity is the product of the lengths of two vectors.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 5, May 2018

Comment-comment distance: The distance between two comments v_a and v_b is defined as:

$$dis(v_a, v_b) = \{\frac{1}{sim(v_a, v_b)} - 1, \ if\ sim(v_a, v_b) \neq 0, \ \infty, \quad if\ sim(v_a, v_b) = 0 \qquad (2)$$

Center of cluster: Let $(v_1, v_2, \dots v_i, \dots, v_n)$ be the set of comments belonging to cluster C_e, where $v_i = (t_{1,i}, t_{2,i}, \dots, t_{j,i}, \dots, t_{N,i})$. The center vc_e of cluster C_e is defined as:

$$vc_e = (tc_{1,e}, tc_{2,e}, \dots, tc_{j,e}, \dots, tc_{N,e}) \qquad (3)$$

$$tc_{j,e} = \sum_{p=1}^{n} t_{j,p} \qquad (4)$$

Similarly, the vector of cluster center is not normalized. In addition, each element in vc_e is not averaged by the number of total comments in the cluster since such an operation is not necessary for distance calculation.

Comment-cluster distance: The similarity between comment v_i and cluster C_e (whose center is vc_e) is defined as:

$$sim(v_i, C_e) = \{\frac{f(v_i, C_e)}{T}, \ if\ f(v_i, C_e) \leq T, \ 1, \quad if\ f(v_i, C_e) > T \qquad (5)$$

T is a positive integer constant.

$$f(v_i, C_e) = \sum_{p=1}^{n} \{ \ 2, \quad if\ t_{p,i} \times tc_{p,e} > 2, \ t_{p,i} \times tc_{p,e}, \quad Otherwise \qquad (6)$$

Accordingly, the distance between comment v_i and cluster C_e is defined as:

$$dis(v_i, C_e) = \{\frac{1}{sim(v_i, C_e)} - 1, \ if\ sim(v_i, C_e) \neq 0, \ \infty, \quad if\ sim(v_i, C_e) = 0 \qquad (7)$$

Short text summarization on comment streams: Given a set of comments S, and a desired number of cluster k, find top-k clusters $\{C_1, C_2, \dots, C_j, \dots, C_k\}$ which have top-k most comments, and the number of comments in C_j is larger than or equal to that in C_{j+1} (i.e., $|C_j| \geq |C_{j+1}|$). Not all comments in S should be included in top-k clusters, and $C_j \subset S$. In addition, the radius of each cluster should be smaller than the radius threshold θ_r.

V. EXPERIMENTAL RESULTS

We collect the real comment streams from Facebook through Facebook Graph API. After transforming each comment into the term vector model representation, if the longest n-gram (denoted as N-GRAM) is set to 3. It can thus be perceived that the lengths of most comments are short. Table 1 gives the quality comparison between different clustering approaches. Based on these results, we focus on evaluating the quality (by F-measure) of 1st cluster and assess the precision of the extracted key terms. IncreSTS is capable of efficiently processing the incremental update with newly-incoming comments and always providing the up-to-date summary. The performance of BatchSTS and IncreSTS clustering algorithm graphically show below:

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 5, May 2018

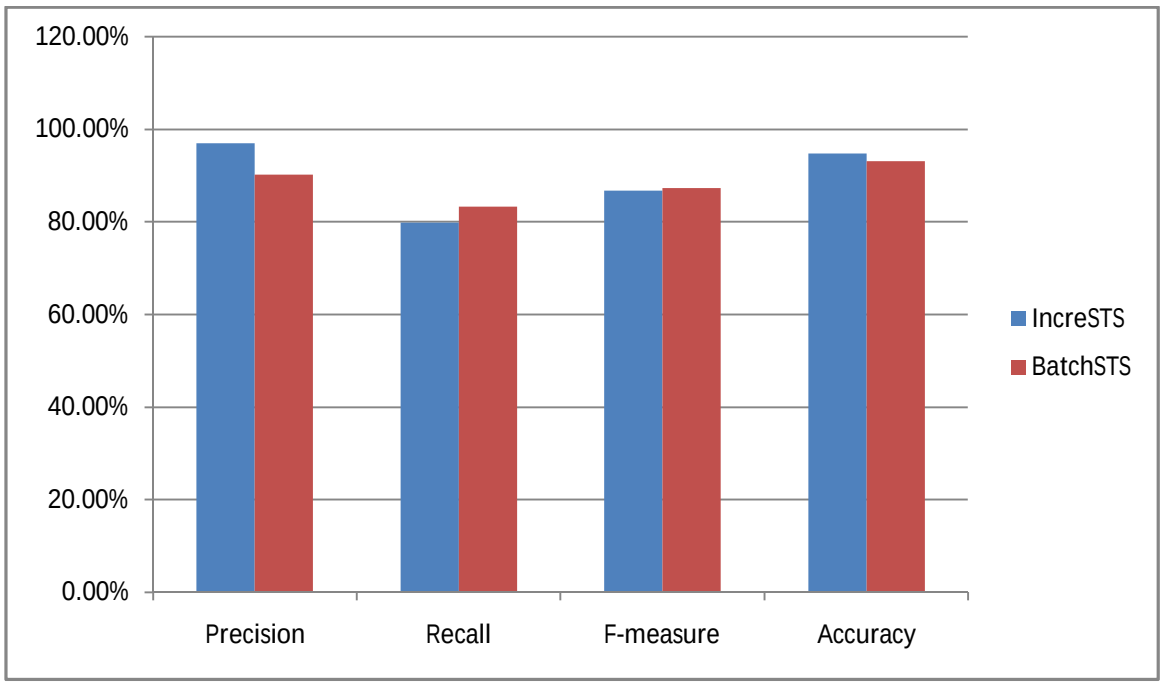

Fig. 2 Performance analysis between BatchSTS and IncreSTS clustering

VI. CONCLUSION

To enable the capability of comment stream summarization on SNS, we model a novel incremental clustering problem and propose the algorithm IncreSTS, which can incrementally update clustering results with latest incoming comments in real time. With the output of IncreSTS, we design a visualization interface consisting of basic information, key-term clouds, and representative comments. This at-a-glance presentation enables users to easily and rapidly get an overview understanding of a comment stream. From extensive experimental results and a real case demonstration, we verify that IncreSTS possesses the advantages of high efficiency, high scalability, and better handling outliers, which justifies the practicability of IncreSTS on the target problem.

REFERENCES

[1] H. Becker, M. Naaman, and L. Gravano, "Learning similarity metrics for event identification in social media," in Proc. 3rd ACM Int. Conf. Web Search Data Mining, 2010, pp. 291–300.
[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] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi, "Eddi: Interactive topic-based browsing of social status streams," in Proc. 23nd Annu. ACM Symp. User Interface Softw. Technol., 2010, pp. 303–312.
[4] 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.
[5] D. Chakrabarti and K. Punera, "Event summarization using tweets," in Proc. 5th Int. AAAI Conf. Weblogs Social Media, 2011, pp. 66–73.