# A Survey On Short Text Summarization Of Comment Streams On Social Network Sites

**Ms. Pooja S. Choudhari , Prof. S. S. Nandgaonkar**

*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.**

*Index Terms*— **Real-time short text summarization, incremental clustering, comment streams, social network services.**

## I. INTRODUCTION

In recent years, social network services are very widespread and have become important communication platforms in our daily life. The largest social networking site Facebook presented the statistics in 2012. According to it, an average of 3.2 billion interactions is generated each day which includes likes and comments. Besides this, Twitter also has millions of users and thus huge amount of messages are posted in a day. All such existing social platforms are very convenient to use and thus have gained high popularity among people. Due to this reason, the celebrities, corporations, and organizations also create their own social pages to interact with their fans and the public. For each message, users can express their opinions by forwarding, giving a like, and leaving comments on it. Due to popularity of these platforms, not only the quantity of comments is large, but also the generation rate is remarkably high. Therefore users unnecessarily have to go through the whole comment list of each message and it is almost impossible everytime. But still users desire to know what other people's are talking about and what are the opinions of these discussion participants.

Mostly, celebrities and corporations have high interest to know how their fans and customers react to certain topics and content. Thus it has created the necessity to develop an advanced summarization technique for comment streams in SNS. The popular services like YouTube [1] and Facebook [2], allow users to determine whether a comment is useful or recommendable and the comments with the top k most endorsements are displayed on the top of the list. This category depends on user contributions. Also this problem is modeled in another way as recommendation [3], [4], [5], [6] or classification [7] task. These ways employ machine learning techniques to solve it. Mostly, sentiment analysis [8] has been applied to discover hidden emotions in messages. Furthermore, providing an informative presentation interface [9] is another interesting field on the summarization of social messages. Traditional comment streams generally express more complete information, such as the discussion on products or movies. But here the main focus is on comment streams in social network sites that are in short text style with casual language. For each social message, main objective is to cluster comments with similar content together and generate a proper opinion summary for that message. For each different groups of opinions, easy and rapid overview should be generated for users and thus an efficient and effective technique should be applied to identify the clusters of all comments of a particular social message.

Grouping similar comments leads to formation of different clusters. These clusters then can be used for summarizing the comment streams from social network sites. Summarizing is defined as reducing text or any content to one-third or one-quarter its original size, clearly indicating its meaning, and retaining main thoughts expressed. The purpose of summarization is to briefly present the key points of any content in order to provide proper context for user. Summarizing is useful in many types of writing and at different points in the writing process. Summarizing is useful in many other aspects such as provide context for a paper's thesis, write literature reviews, and annotate a bibliography. The benefit of summarizing is it allows the reader to contextualize what people are saying, which is very vital in case of huge amount of social media contents generated everyday. In addition to this, summarizing helps the user to gain a better sense of what exactly the information or content is conveying.

***Manuscript received Nov, 2015***.

***Pooja S. Choudhari***, *M.E. Student, Computer Engineering, Savitribai Phule Pune University, VPCOE, Baramati,, India*

***Prof. S. S. Nandgaonkar***, *Assistant Professor Computer Engineering, Savitribai Phule Pune University, VPCOE, Baramati, India.*

  

A summary is typically generated with two main categories of techniques, called extraction and abstraction. Extractive summary involves identifying relevant sentences that belong to the summary. Abstractive summarization involves identifying or paraphrasing sections of the content to be summarized. Extractive summarization simply extracts salient information, such as sentences, from the input contents and "puts them together" to form summaries. Although summaries generated in this way may lack of coherence, but still extractive approaches are now-a-days as they are low cost and easy to be applied to general domains. To produce grammatical coherent summaries, abstractive summarization creates summaries by synthesizing and rewriting sentences based on contextual and linguistic understanding and it is heavily dependent on deep analysis and language generation techniques. Sometimes regeneration is done as a post-process for extractive summaries, i.e., make pruning or revision based on extractive summaries.

A Microblog user i.e for instance, the user's of twitter and facebook usually has to browse through tens or even hundreds of posts together with their responses daily, therefore it can be beneficial if there is an intelligent tool for summarizing this information. Automatic text summarization (ATS) has been used mainly for many years, but the majority of the existing techniques might not be appropriate for Microblog sites. For instance, a popular kind of approaches for summarization tries to identify a subset of information, usually in sentence form, from longer pieces of writings as summary. Such extraction-based methods can hardly be applied to Microblog texts because many posts/responses contain only one sentence. Some special characteristics that deviates the Microblog summarization task from general text summarization are:

1. The number of sentences is limited, and sentences are usually too short and casual to contain sufficient structural information or cue phrases. Unlike normal blogs, there is a strict limitation on the number of characters for each post, For e.g. 140 characters for Twitter .

2. Microblog posts can serve several different purposes. At least three different types of posts are observed in Microblogs, expressing feeling, sharing information, and asking questions. Structured language is not the only means to achieve those goals. For example, people sometimes use attachment, as links or files, for sharing, and utilize emoticons and pre-defined qualifiers to express their feelings. The diversity of content differ Microblogs from general news articles.

3. Posts and responses in Microblogs are more similar to a multi-persons dialogue corpus.

## II. SUMMARIZATION AND CLUSTERING TECHNIQUES

1. Summarizing User-contributed Comments[3] :

It is one of the approach used for summarizing the comments contributed by user's on all types of social services which is also referred as "Social Web" plays a very important role. It is vital part of social networking. This is shared & adopted across many social media sites and also by news providers too. Although these all comments creates huge level of user engagement with online social media, but their processing and assimilation of perspectives is a burden or a problem. Thus the system in [3] overcomes this problem by doing comment summarization. Here the goal is to select the most representative comments from a large collection of user-contributed comments and these comments should contain different viewpoints and aspects about the associated resource. From the set of n-user contributed comments, best top-k comments are selected for summarization using two approaches-first, clustering based approach which identifies correlated groups of comments and second, a precedence based ranking framework which automatically selects informative user contributed comments. K-means clustering & LDA both are used for identifying group of related comments. For identifying informative and important comments from cluster, Term importance approach is used and also precedence based method is used which applies random walk over a comment graph based on page rank style. Thus the system in [3] yields overall higher performance compared to traditional document summarization methods because precedence ranking is used combinely with topic based clustering.

2. Event Summarization Using Tweet [4] **:**

This technique is applied specifically for Twitter. Now-a-days Twitter has become very popular website on which hundreds of tweets are posted everyday on variety of topics. So it has greatly helped to make real-time search applications which displays relevant tweets in response to user queries and these applications work with search engines. Most of these tweets are about "events", So the detection of such events has become vital today. But, very little applications can properly display the real-time information about events. Presently, search engines simply display all tweets matching the queries in reverse chronological order. Thus [4] gives more sophisticated techniques to summarize the relevant tweets. The solution for event-tweets summarization is given on the basis of learning underlying hidden state representation of event via "Hidden Markov Models". The method used in [4] is a 2 step process as- firstly a modified Hidden Markov Model is used that segments the event time-line, which depends on both the burstiness of the tweet-stream and the word distribution used in tweets. Each such segment represents one distinct "sub-event", which is semantically distinct portion of the full event. Then second, the key tweets are picked up to describe each segment judged to be interesting and these tweets are combined to build the summary. Here the main objective is to summarize the tweets in real time which allows for intelligent interaction of tweets into search results. For this,3 algorithms are used as – SUMALLTEXT, SUMMTIMEINT and SUMMHMM.

The SUMALLTEXT simply consider each tweet as a document, and then applies a summarization method on this corpus. With each tweet, a vector of the TF-logIDF with its constituent words is associated. Then distance between two tweets is found and it is said to be the cosine distance. Thus tweets closest to all other tweets from the event are selected.

  

In SUMMTIMEINT, so as to pick tweets from the entire duration of the event, summarization is combined with segmentation. Simply duration is split up into equal-sized time intervals and key tweets from each interval are selected. Not all intervals contain useful sub-events, because they have low tweet volume relative to the average, and so key tweets from such intervals are not selected In SUMMHMM, there are two parts to event summarization as-detecting stages or segments of an event, and summarizing the tweets in each stage by Hidden Markov Model (HMM). The HMM can learn differences in language models and its parameters are interpretable. So it can be applied to a wide variety of events. 2 modifications are done to standard HMM –output per time step and detecting bursts in tweet volume. Thus SUMMHMM algorithm performs better than other two.

3. Recommending Content from Information Streams [5] :

Most of the web users keep update with newest information through information streams or websites such as the Twitter. So this scheme is used for Twitter to better direct user attention and this system is termed as "Recommender". For designing recommender, 3 types of dimensions are used as-content sources, topic interest models for users and social voting. Design space for recommender consists of 3 dimensions –selection candidate URL's set, ranking topic and ranking using social voting . Thus this system shows that URL recommender on twitter is a means to better direct user attention in information.

4. Summarization of Document Streams[6] :

It is a method for summarization of short documents on microblogs such as twitter are carried out. User's posts creates short documents on twitter a certain topic such as sports matches or TV dramas, etc. These documents are often highly redundant and i.e. there can be many documents on one event in the topic. The model in [6] specifically generates a good summary on documents on sports matches. For summarizing the document stream, 2 approaches are used as-first, modification of coefficient "eij" is done means if 2 documents are temporally distant ,then they should have small "eij", even if their contents are similar. Second, Linear partition constraint is used on document assignment means 2 temporally distant documents are unlikely to be linked unless all the documents between that 2 documents are similar. Thus this system considers various characteristics of microblogs and generates the summary using fast approximate algorithm.

5. IMASS: Intelligent Microblog Analysis and Summarization System [9]:

It is used to summarize a microblog post and its responses . It helps the readers to get a more constructive set of information in a efficient manner. It works in two phases. In first phase, the post plus and its responses are classified into four categories based on the intention, interrogation, sharing, discussion and chat. Second phase uses different strategies for each type of post like opinion analysis, response pair identification and response relevancy detection, to summarize and highlight critical information to display. Microblogs has very different characteristics than other online information sources as news articles. These characteristics differ in terms of length and writing skills. Thus this scheme in [9] gives an effective strategy to summarize post and responses by first determining intention and then analyzing post types.

6. SUMCR: Subtopic Based Feature Extraction Approach[15]:

It proposed a new subtopic based extractive approach for text summarization. Relevance and coverage are two main important criteria that decide the quality of a summary. A new multi-document summarization approach SumCR via sentence extraction is used. A feature called Exemplar is introduced to help to simultaneously deal with these two concerns at the time of sentence ranking .In traditional methods, relevance value of each sentence is calculated based on the collection of sentences. Similarly the Exemplar value of each sentence in SumCR is obtained within a subset of sentences which are similar. A clustering approach based on fuzzy mediod method is used to produce clusters of sentences or subsets where each of them corresponds to a subtopic of its related topic. Such type of subtopic-based feature captures the relevance of every sentence within different subtopics and therefore helps SumCR to produce a summary with a wider coverage and less redundancy. Also one more feature is used in SumCR called as" Position", i.e., the position of each sentence occured in the corresponding document. The final score of each sentence is a combines the subtopic-level feature i.e. Exemplar and the document-level feature i.e. Position. " Exemplar" feature used here is more effective than other subtopic level features.

There are many existing clustering algorithms, some of them are specified in [10],[11],[12].

1. K-means clustering technique [10] :

It divides M points in N dimensions into K clusters, it helps to minimize the cluster sum of squares. It requires as a input a matrix of M points in N dimensions and a matrix of K initial cluster centers in N dimensions. It makes search for K-partition with locally optimal within cluster sum of squares by moving points from one cluster to other. It uses two algorithms-AS113(A transfer algorithm for non hierarchical classification) and AS58(Euclidean cluster analysis).These 2 algorithms require the time equal to number of iterations.

2. OPTICS clustering method[11] :

Ordering points are used here to identify clustering structure. OPTICS is the extension of DBSCAN algorithm, it generates cluster ordering consisting of ordering of points, reachability values and core values. The advantage of OPTICS is, it do not limit itself to global parameter setting, so it serves as versatile basis for both automatic and interactive cluster analysis.

  

3. BIRCH(Balanced Iterative Reducing and Clustering using Hierarchies) [12] :

It is an incremental data clustering method specially designed for very big databases. Recently, to find useful patterns in large datasets is creating more interest. One of the most widely studied problems in this area is to identify clusters or densely populated regions, in a multidimensional dataset. But existing methods does not adequately solve this problem of large datasets and minimization of I/0 costs. So a data clustering technique BIRCH is proposed which is especially suitable for very huge databases. BIRCH works in incremental manner and dynamically clusters incoming multi-dimensional metric data points to produce the best quality clustering with the available resources such as available memory and time constraints. It can find a better clustering with a single scan of data and improve the quality further with a few additional scans. BIRCH also handle noise effectively. Thus it efficiently works for very large databases and with any given amount of memory, but I/O complexity is little more than one scan of data.

4. Topical Clustering of Tweets [7] :

A technique is for automatically clustering and classifying twitter messages i.e. "tweets", into different categories for eg. GoogleNews Service. Due to micro-blogging and social communication services, users post thousands of short messages every day. But keeping track of all the messages posted by your friends or other people is impossible & tedious. Unsupervised and Supervised Clustering both methods are used here. Unsupervised clustering uses LDA & K-means algorithms. Supervised clustering uses Rocchio classifier. Thus from each cluster, top few tweets are found to summarize a cluster.

Different incremental clustering algorithms are proposed in [13],[14].

1. Incremental K-means algorithm[13] :

It is simple and computationally efficient. But the main problem with this method is its tendency to converge at a local minimum. Thus [13] explains the cause of this problem and an existing solution involving a cluster centre jumping operation is examined. The jumping technique eliminates the problem with local minima by enabling cluster centres to move in such a radical way as to reduce the overall cluster distortion. But, the method is very sensitive to errors in estimating distortion. The other clustering method is also presented, that is also based on distortion reduction through cluster centre movement but is not so sensitive to inaccuracies in distortion estimation . The scheme is an incremental version of the K-means algorithm, it involves addition of cluster centres one by one as clusters are being formed. Compared to K-means, incremental K-means algorithm requires K-times more iterations because when there are K clusters, this new algorithm need to run K times and each iteration is equivalent to one execution of K-means algorithm.

2. Incremental hierarchical text document clustering algorithms[14] :

Incremental hierarchical text document clustering algorithms are important in organizing documents generated from streaming on-line sources, such as, Newswire and Blogs.

Popular incremental hierarchical clustering algorithms are Cobweb and Classit, but not suitable for text clustering. So an alternative method is given which includes changes to the assumption of the algorithm in order to conform with the empirical data. For incremental hierarchical text document clustering, a Cobweb based algorithm is used where word occurance attributes follows Katz's distribution instead of normal distributions. It also gives a way to evaluate quality of hierarchy generated by hierarchical clustering algorithms.

Therefore when compared to the existing methods, the scheme proposed in [16] gives the IncreSTS algorithm which is first fully incremental algorithm that provide immediate and instant summary of real-time social comment streams. IncreSTS and BatchSTS clustering algorithms are used to generate comment clusters. These both algorithms slightly sacrifices cluster quality slightly but can achieve the real-time processing need of the comment stream summarization problem. The main difference from existing clustering methods is that the IncreSTS maintains the radius of every cluster should be smaller than a predefined threshold.

## III. TABLE IN BRIEF FOR SUMMARIZATION TECHNIQUES

| Method | Advantage | Disadvantage |
|---|---|---|
| Clustering approach and precedence ranking framework for selection of best top-k comments. [3] | Combination of precedence based ranking & term importance approach with topic clustering, so it ranks the clusters according significance measure. | This approach is not suitable for microblog social sites. [3] |
| Event summarization using tweets[4] | Three algorithms - SummallText, SummTimeint and Summhmm, and these are combined with Hidden markov model. So summary of relevant tweets is obtained. | Cannot provide summaries for unpredictable events such as revolutions & natural disasters. Unable to give real time summary. [4] |
| Recommender: Technique for Recommen-ding | Due to use of 3 dimensions –social voting, content sources & topic interest models, gives proper content | Provides only URL recommendation .[5] |

  

| | | |
|---|---|---|
| content from information streams [5] | recommendation. | |
| Document Stream Summarization [6] | Linear partition constraint on document assignment links only relevant documents. | Not suitable for diverse datasets. []6 |
| IMASS: Intelligent Microblog Summarization technique. [7] | Determines the user intention based on analysis of opinions and response-pair identification | Cannot be applied on web blogs and news articles. [7] |
| SUMCR: Extractive approach for text summarization [8] | Exemplar feature is used, it is more effective than other subtopic-level features because it gives value of sentence. | Finding Proper combination of weights for every feature of sentences creates overhead. [8] |
| IncreSTS:Real time short text summarization method. [16] | Provides updated real time comment stream summary, by maintaining radius restrictions for each cluster. | Cluster quality is slightly sacrificed. [16] |

## IV. CONCLUSION

Thus, it seems that, for real time comment stream summarization problem on social networking sites, incremental clustering technique will prove useful. This technique makes use of IncreSTS algorithm which can incrementally update clustering results with latest incoming comments in real time. These clusters will be then summarized so that users can get an overview understanding of a comment stream easily and rapidly without going through the whole comment list of each social message. Existing clustering methods which are presented here mainly focus on maintaining cluster quality & hence they cannot provide real time updated summary of comments.

## ACKNOWLEDGMENT

I wish to thank Prof S.S. Nandgaonkar for giving me proper guidance and suggestions regarding this work.

## REFERENCES

[1] YouTube [Online]. Available: http://www.youtube.com/, 2014.
[2] Facebook [Online]. Available: http://www.facebook.com/, 2014.
[3] E. Khabiri, J. Caverlee, and C.-F. Hsu, "Summarizing User- Contributed Comments," in Proc. 5th Int. AAAI Conf. Weblogs Soc.
[4] D. Chakrabarti and K. Punera, "Event summarization using tweets," in Proc. 5th Int. AAAI Conf. Weblogs Social Media, 2011, pp. 66–73.
[5] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. H. Chi, "Short and tweet: Experiments on recommending content from information streams," in Proc. ACM SIGCHI Conf. Human Factors Comput. S.P. Bingulac, "On the Compatibility of Adaptive Controllers," Proc. Fourth Ann. Allerton Conf. Circuits and Systems Theory, pp. 8-16, 1994.
[6] H. Takamura, H. Yokono, and M. Okumura, "Summarizing a document stream," in Proc. 33rd Eur. Conf. IR Res. Adv. Inf. Retrieval, 2011, pp. 177 188.
[7] K. D. Rosa, R. Shah, B. Lin, A. Gershman, and R. Frederking, "Topical clustering of tweets," in Proc. ACM SIGIR's 3rd Workshop Social Web Search Mining, 2011, http://www.cs.cmu.edu/ kdelaros/ sigir¡Vswsm¡V2011.pdf
[8] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in Proc. 4th Int. AAAI Conf. Weblogs Social Media, 2010, pp. 178–185.
[9] J.-Y. Weng, C.-L. Yang, B.-N. Chen, Y.-K. Wang, and S.-D. Lin, "IMASS: An intelligent microblog analysis and summarization system," in Proc. ACL/HLT Syst. Demonstrations, 2011, pp. 133–138.
[10] J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm," J. Roy. Statist. Soc.. Series C (Appl. Statist.), vol. 28, no. 1,pp. 100–108, 1987.
[11] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: Ordering points to identify the clustering structure," in Proc. ACM SIGMOD Int. Conf. Manag. Data, 1999, vol. 28, no. 2, pp. 49–60.
[12] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," vol. 25, no. 2, pp. 103–114, 1996.
[13] D. T. Pham, S. S. Dimov, and C. D. Nguyen, "An incremental k-means algorithm," Proc. Institution Mech. Eng., C, J. Mech. Eng.Sci., vol. 218, no. 7, pp. 783–795, 2004.
[14] N. Sahoo, J. Callan, G. Duncan R. Krishnan, and R. Padman, "Incremental hierarchical clustering of text documents," in Proc. 15th ACM Int. Conf. Inf. Knowl. Manag., 2006, pp. 357–366.
[15] J.-P. Mei and L. Chen, "SumCR: A new subtopic-based extractive approach for text summarization," Knowl. Inf. Syst., vol. 31, no. 3, pp. 527–545, 2012.
[16] Cheng-Ying Liu, Ming-Syan Chen,Chi-Yao Tseng," IncreSTS: Towards Real-Time Incremental Short Text Summarization on Comment Streams from Social Network Services",IEEE Transactions on Knowledge and Data Engineering, vol.27,No.11,Nov 2015.

**Pooja S. Choudhari** recieved B.E degree in Computer Engineering from MMCOE, Savitribai Phule Pune University, Pune. Currently pursuing Masters degree in Computer Engineering from Savitribai Phule Pune University, VPCOE, Baramati, Pune.

**Prof S. S. Nandgaonkar** recieved B.E degree in Computer Science and Engineering from WIT, Solapur, Shivaji University and M.Tech in Computer Engineering from COEP, Pune University, Pune. Currently working as Assistant Professor in VPCOE, Baramati, Pune .

  