International Journal of Science and Research (IJSR)
ISSN (Online):
Index Copernicus Value (2016): 6.14 | Impact Factor (2015): 6.391

RISTS: Real-Time Incremental Short Text Summarization of Comment Streams in Social Networks

Aysha Shabin S. H.[1], Mohammed Malik C. K.[2]

[1, 2]AWH Engineering College, Calicut University, Department of Computer Science & Engineering, Kuttikkatoor, Kozhikode, India

Abstract: ***In recent years, the popularity of social networking services has increased immensely, so the number of comments can raise at a high rate immediately after a social message is published. The users of the social sites always want to get a brief understanding of a comment stream without reading the whole comment list. In order to support real-time short text summarization of comment streams in social networks, here proposed a new summarization system called RISTS . The system group comments with content similarity, semantic similarity and generate a concise opinion summary for the message. To provide immediate and instant summary of real time comment streams, the system makes use of IncreSTS algorithm which can incrementally update clustering results with latest incoming comments in real time. Moreover an at-a-glance visualization interface will be designed that enables users to get an overview understanding of a comment stream.***

Keywords: Real-time short text summarization, Incremental clustering, Comment streams, Key - term extraction, Social network services

1. Introduction

Now-a-days social network services has gained remarkable attention and have become important communication platforms in our daily life. According to the 2012 statistics by the largest social networking site Facebook, there are over 500 million daily active users and an average of 3.2 billion interactions(including Likes and Comments) is generated each day. As a result of the popularity and convenience of these platforms, celebrities and organizations also set up social pages to interact with their fans and the public. For each message, users can able to express their opinions by sharing, giving a like, and leaving comments on it. Users unnecessarily and almost impossibly go over the whole comment list of each message. However, some users may still desire to know what are other users talking about and what are the opinions of these discussion participants. With these motivations, we develop a sophisticated summarization technique called RISTS targeting at comment streams in SNS.

There are many different approaches for generating various kinds of summaries on comment streams. Most of the existing summarization approaches mainly focused on the traditional comment streams that usually express more complete information. In this paper we target at comments that are in short text style with informal language style. Moreover, there are some restrictions for the traditional summarization methods, for generating the real time summary of comment streams. Furthermore, some of the existing clustering methods doesn't consider the overall cluster quality. That is only cluster comments with similar content together [8]. Which leads to the redundant comments in clusters and decreases the overall cluster quality.

In this paper , we aim to generate the real-time summary of comment streams by considering the overall cluster quality. For each social message, the main objective is to cluster comments with content similarity, semantic similarity and generate a concise opinion summary for this 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. Finally a visualization interface is designed for presenting the summarized result. The purpose of summarization is to briefly present the key points of any content in order to provide proper context for user. There is a need to discover how many different group opinions exist and provide an overview of each group. Therefore, here the goal is developing an efficient and effective technique to identify the clusters of these comments.

The remainder of this paper is organized as follows. Section 2 surveys related works.Detailed description of real-time incremental short text summarization is given in Section 3. Finally, the conclusion and future work are presented in Section 4.

2. Related Works

Regarding the research field of text - based summarization of user generated content , in recent years, numerous works are focused on three kinds of user generated content: online reviews, blogs, and short text messages. A variety of techniques have been developed and applied to satisfy different needs of summarization. IMASS [1] is a system to summarize a microblog post and its responses with the goal to provide readers a more constructive and concise set of information for efficient digestion. The authors in [1] introduce a novel two-phase summarization scheme. In the first phase, the post plus its responses are classified into four categories based on the intention, interrogation, sharing, discussion and chat. For each type of post, in the second

phase, the system chooses different strategies, including opinion analysis, response pair identification, and response relevancy detection, to summarize and highlight critical information to display.

Before the popularity of social network services and micro-blogging websites, blog is one of the primary platforms that users publish content. As for the summarization of traditional blogs, one main research direction is to extract and discover representative sentences. The authors in [2] consider utilizing user feedback comments to identify important sentences on a blog post. The proposed sentence scoring mechanism is based on the observation that user-contributed comments can provide valuable information to better understand the blog content. To select the best top informative comments from a set of user-contributed comments for a specific object, such as a video, E. Khabiri, J. Caverlee, and C.F. Hsu [3] proposed an approach. Where initially a modified model of Latent Dirichlet Allocation (LDA) is applied to cluster comments into several groups based on the concept of topic modeling. Then, a precedence-based ranking approach is proposed to select informative comments for each cluster.

With the flourish of the Web, online review [4][5] is becoming a more and more useful and important information resource for people. Different from traditional text summarization, review mining and summarization aims at extracting the features on which the reviewers express their opinions and determining whether the opinions are positive or negative. In [4], M.Hu and B. Liu study the problem of generating feature-based summaries of customer reviews of products sold online. They proposed different novel techniques for summarizing customers reviews. They summarizes reviews by following three ways: (1) mining product features that have been commented on by customers; (2) identifying opinion sentence in each review and deciding whether each opinion sentence is positive or negative; (3) summarizing the results. On the other hand, work in [5] is focus on a specific domain – movie review. Different from product reviews, movie reviews have some unique characteristics. The commented features in movie review are much richer than those in product review. In this paper a multi-knowledge based approach is proposed for movie review mining and summarization. Here the problem of review mining and summarization is decomposed into the following subtasks: 1) identifying feature words and opinion words in a sentence; 2) determining the class of feature word and the polarity of opinion word; 3) for each feature word, fist identifying the relevant opinion word(s), and then obtaining some valid feature-opinion pairs; 4) producing a summary using the discovered information. To perform these tasks a multi-knowledge based approach is proposed, which integrates WordNet, statistical analysis and movie knowledge. Twitter has become exceedingly popular, with hundreds of millions of tweets being posted every day on a wide variety of topics. Recent research has shown that a considerable fraction of these tweets are about "events"[6][7], and the detection of novel events in the tweet-stream has attracted a lot of research interest. For summarizing the tweets about some highly structured and recurring events, such as sports, Chakrabarti and Punera [6] proposed a solution called SUMMHMM algorithm, it consists of two steps. That is ,detecting stages or segments of an event, and summarizing the tweets in each stage. Authors in [7] explore approaches for finding representative messages among a set of Twitter messages that correspond to the same event, with the goal of identifying high quality, relevant messages that provide useful event information. Here the problem of selecting Twitter content for events can be address by two concrete steps. First, identify each event and its associated Twitter messages using an online clustering technique that groups together topically similar Twitter messages. Second, for each identified event cluster, select messages that best represent the event. To identify event content here associate twitter messages with events using an incremental online clustering algorithm.

3. Real - Time Incremental Short Text Summarization

This section gives the detailed description of real-time incremental short text summarization of comment streams in social networks. Because of the high popularity of social networking services, the quantity of comments for a social message may rise quickly and continuously. Moreover, the users of the social sites always desire to get a brief understanding of a comment stream without reading the whole comment list, but they may request summary of the comment streams at any moment. In order to generate the real-time summary of comment streams, here propose an advanced summarization technique called RISTS.

The main objective of RISTS, is to cluster comments with content similarity, semantic similarity and generate a concise opinion summary for this message. To provide immediate and instant summary of real time comment streams, an IncreSTS algorithm is used. Which incrementally update clustering results with latest incoming comments in real time. Moreover, design an at-a-glance visualization interface to help users easily and quickly get an overview summary.

3.1 System Architecture

This section describes the summarization framework of real-time incremental short text summarization system. Once a message is posted on SNS, users can leave comments immediately and the quantity of comments may increase quickly and continuously. Furthermore, readers are usually unwilling to go over the whole list of comments, but they may request to see the summary at any moment. This indicates that the RISTS approach should be able to generate the summary result at any time point of a dynamic data stream. To satisfy this requirement, here model this problem as an incremental clustering task. The system architecture of RISTS is depicted in Fig.1.

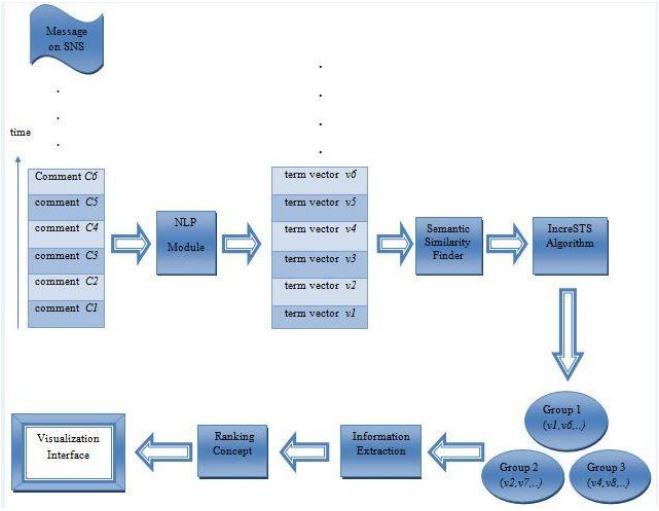

Figure 1: System architecture of RISTS

The system architecture of RISTS adopt the term vector model, and therefore each comment is transformed into a set of n-gram terms by the NLP module. Since informal and unstructured texts are widely used on SNS, and also apply some heuristics to enhance the quality of n-gram terms that can better represent each comment. Here a semantic similarity finder is used to check the semantic similarity between comments. Whenever a request for the real-time summary of comment streams is received, the IncreSTS algorithm incrementally producing latest clustering results and simultaneously outputting significant comments that are closest to the center of each cluster. Finally, for the visualization interface, representative terms will be extracted to form a key-term cloud for each group. Moreover the summaries are ranked based on their relevance. Thus, users will be provided a concise, informative, and at-a-glace presentation that can help them easily comprehend the main points of responses to one message on SNS.

3.2 Term Vector Model Representation of Comments

This section elaborate the details of NLP module that transforms each comment into a set of n-gram terms. The NLP module consist of mainly five steps. Initially for each word, the process of punctuation removal will be applied to eliminate unnecessary punctuation marks connected with words. Moreover, develop the heuristic process of redundant character removal, designed for restoring words on SNS. It can be observed that casual language style is commonly used on SNS. In particular, users often emphasize the emotion by repeating characters in a word. This phenomenon certainly causes the problem of not being able to correctly identify the original words. To cope with this problem, examine each word to find out whether there is any character consecutively appearing more than three times. If this situation is detected, appended characters will be regarded as redundant, and only one character will be retained. Meanwhile, all upper-case letters will also be changed to lower-case letters.

The following step is the stemming process. Where the inflected and derived words are reduced to their stem form . Subsequently, the process of n-gram terms extraction is carried out to extract terms that are used for representing this comment. Finally, stopwords removal process is executed to delete terms entirely composed of stopwords. Note that as long as there is at least one non- stopword appearing in a term, this term will be viewed as a valid one.

3.3 Clustering

Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. In this paper, the problem of incremental short text summarization is modeled as an incremental clustering task. Consider two comments represented in the term vector model, $\upsilon_a = (t_{1,a}, t_{2,a}, \ldots t_{N,a})$ and $\upsilon_b = (t_{1,b}, t_{2,b}, \ldots, t_{N,b})$. Each dimension corresponds to a separate term, and N is the number of dimensions. Here define that the weights of terms are equal, if the term t_i occurs in the comment υ_a, $t_{i,a}$ will be set to 1. Otherwise, $t_{i,a}$ will be set to 0. The reason for this design is that the length of each comment is usually very short compared to other text documents. Then find out the content similarity score of the comments using modified cosine similarity equation.

$$\text{sim}(\upsilon_a, \upsilon_b) = \begin{cases} \upsilon_a \cdot \upsilon_b / D & \text{if } \upsilon_a \cdot \upsilon_b \leq D \\ 1 & \text{if } \upsilon_a \cdot \upsilon_b > D \end{cases}$$

Where $\upsilon_a \cdot \upsilon_b$ is the inner product of two vectors, and D is a positive integer constant. Then check the semantic similarity between comments by a semantic similarity finder. To decide whether two words are semantically similar, it is important to know the semantic relations that hold between the words. Here WordNet is used as the semantic similarity finder. To generate the comment clusters for specific message on SNS, here first introduce the batch version of short text summarization algorithm(BatchSTS). Then propose a fully incremental algorithm to provide immediate and instant summary of real-time social comment streams(IncreSTS). Finally for representing the summarization results, here propose a key-term extraction algorithm.

BatchSTS Algorithm: BatchSTS takes the whole comment set as the input. The second input is the radius threshold theta used for determining how similar the comments are in a cluster. The aim of this algorithm is to find all connected components of the comment set . The points belonging to the same connected component will be merged as a cluster. This algorithm outputs different clusters of comments. The BatchSTS algorithm is described formally in algorithm 1.

Procedure (BatchSTS)

2. For each comnt $_i$ in comment set S ,create term vector TV_i
3. Initialize cluster C=ϕ
4. Create first cluster with comnt $_0$
5. Cluster C= comnt $_0$
6. For each comment comnt $_i$ in comment set S ,for all i!=0
7. Initialize simlist=ϕ
8. For each cluster C_i in C
9. Add sim(C_i , comnt $_i$) to simlist
10. End of loop 6
11. If max(simlist) ≥ Θ
12. Add comnt $_i$ to C_i
13. Update the cluster center C_c
14. Else
15. Create new cluster with comment comnt $_i$
16. End of loop
17. For any two clusters C_i , C_j in C
18. If sim (C_i , C_j) ≥ Θ
19. Merge (C_i , C_j)

End

Algorithm 1: BatchSTS algorithm

IncreSTS Algorithm: It is an iterative version of BatchSTS algorithm. Which is aiming to provide immediate and instant summary of real-time social comment streams. The primary notion of this algorithm is to maintain the clustering result of the previous phase, and to incrementally update the clustering result with the newly-incoming comment. Here first check whether the last comment that is considered in the BatchSTS algorithm is equal or not to the newly-incoming comment comnt $_{new}$. If it is not equal then clear the previous term vectors ,clusters, cluster elements. Then call the BatchSTS algorithm for clustering Comment streams. The IncreSTS algorithm is described formally in algorithm 2.

Procedure (IncreSTS)

1. If comnt $_{old}$!= comnt $_{new}$
2. Clear term vectors
3. Clear clusters
4. Clear cluster elements
5. Initiaize wordlist
6. Call BatchSTS
7. Save comnt $_{old}$ = comnt $_{new}$

Algorithm 2: IncreSTS algorithm

Key-Term Extraction Algorithm: Finally for the design of visualization interface , representative terms will be extracted to form a key-term cloud for each group. For this here propose a key-term extraction algorithm. Which intent to provide a concise at-a-glance visualization interface that enables users to quickly get an a overview understanding of comment stream.

Procedure (Key-Term Extraction)

1. For each cluster C_i in cluster C, initialize k_i= ϕ
2. For each comment comnt $_i$ € C_i
3. Create 1-gram,2-gram,3-gram
4. End for each loop
5. For each word W_i in wordlist W
6. If W_i in 3-gram $_i$
7. If k_i contains W_i == false
8. Add W_i to k_i
9. Else if W_i in 2-gram $_i$
10. Add W_i to k_i
11. Else if W_i in 1-gram $_i$
12. Add W_i to k_i

End

Algorithm 3: Key-Term Extraction algorithm

4. Conclusion and Future Work

In this paper, a new summarization system is proposed for generating the real time summary of comment streams in social network services. For enabling the capability of comment stream summarization, it 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. Moreover the system provides a visualization interface that consists of basic information and key-terms present in the comments.

In the future work, we will further improve our approach from two aspects. Firstly, a filter module will be added for removing unwanted comments of a particular message, which may increase the quality of comments. Secondly, we will consider the representative comments from each cluster for one type of summary presentation.

References

[1] J.-Y. Weng, C.-L. Yang, B.-N. Chen, Y.-K. Wang, and S.-D. Lin. "IMASS: An Intelligent Microblog Analysis and Summarization System". Proc. of the ACL/HLT Systems Demonstrations (ACLHLT 11), pages 133 138, 2011.

[2] M. Hu, A. Sun, and E.-P. Lim. "Comments-Oriented Blog Summarization by Sentence Extraction". Proc. of the 16th ACM International Conference on Information and Knowledge Management (CIKM07), pages 901904, 2007.

[3] E. Khabiri, J. Caverlee, and C.-F. Hsu. "Summarizing User-Contributed Comments". Proc. of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM11), pages 534537, 2011.

[4] M. Hu and B. Liu. "Mining and Summarizing Customer Reviews". Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD04), pages 168177, 2004.

[5] L. Zhuang, F. Jing, and X.-Y. Zhu." Movie Review Mining and Summarization" Proc. of the 15th ACM International Conference on Information and Knowledge Management (CIKM'06), pages 43–50, 2006.
[6] D.Chakrabarti and K. Punera. Event Summarization Using Tweets". Proc. of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM11), pages 6673, 2011.
[7] H. Becker, M. Naaman, and L. Gravano."Selecting Quality Twitter Content for Events". Proc. of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11), pages 442–445, 2011.
[8] 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.

Author Profile

Aysha Shabin S.H received the B Tech degree in Computer Science and Engineering from university of Calicut in 2014 . Currently pursuing M Tech in Computer Science and Engineering from university of Calicut , respectively.

Mohammed Malik C.K received the Diploma, B Tech and M Tech degrees in Computer Science and Engineering from Technical Board and Anna University in 2002, 2006, and 2013 respectively. During April 2006 to November 2007, he was working in HCL Technologies, Chennai. He had been worked in Al Bassami International Business Group, Riyadh as software engineer in 2009. Now he is currently working as assistant professor in AWH Engineering College.