INTRODUCTION

IN recent years, social network services (SNS) are prevalent and have become important communication platforms in our daily life. According to the 2012 statistics by the largest social networking site Facebook, 1. There are over 500 million daily active users and an average of 3.2 billion interactions (including Likes and Comments) is generated each day. Besides, Micro-blogging giant Twitter has over 400 million user base and there are close to 200 million messages posted in a day.2. Due to the popularity and convenience of these platforms, celebrities, corporations, and organizations also set up social pages 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. Compiling the data from 8 celebrities on Facebook, As can be observed, not only the quantity of comments is large, 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, we are inspired to develop an advanced summarization technique targeting at comment streams in SNS. 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. Like YouTube and Facebook, these popular services 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 relies on user contributions and intends to leverage the wisdom of crowds.
On the other hand, some researchers model this problem as recommendation or classification tasks and employ machine learning techniques to solve it. Moreover, sentiment analysis has been applied as well to discover hidden emotions in messages. Furthermore, providing an informative presentation interface is another active research field on the summarization of social messages. Despite some effort has been spent on solving this information overload problem, a generalized approach for summarizing rapid-increasing comment streams in SNS, based on text content, is yet to be fully explored.

We do 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 in 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. Therefore, 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. First, 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. Furthermore, 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 methods may be prone to poor results.

On the other hand, the techniques of document clustering based on topic modeling concepts, such as latent dirichlet allocation (LDA) and latent semantic analysis (LSA) are another possibility to cope with this problem. 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. Besides, the process of parameter estimation is time-consuming, and thus they are not applicable to real-time tasks.

We explore the problem of incremental short text summarization on comment streams from social network services. We 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 resonation 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. Our objective is to generate an informative, concise, and impressive interface that can help users get an overview understanding without reading all comments.

Note that this one does not focus on the techniques of natural language processing (abbreviated as NLP). With employing some basic NLP procedures, each comment is transformed to a set of n-gram terms. On the other hand, we define new similarity measures to adequately determine the distance between comments and clusters. According to the new definitions, we propose a fully incremental algorithm that is almost parameter-free and can handle the outlier problem. Furthermore, 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, we 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 can be summarized as follows.

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