ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Clustering Based Concise Opinion Summary

R.Rajani, G.Deepthi , P.Sujatha

Head & Associate Professor, Dept. of MCA, Narayana Engineering College, Nellore, AP, India

Student, Dept. of MCA, Narayana Engineering College, Nellore, AP, India

Student, Dept. of MCA, Narayana Engineering College, Nellore, AP, India

ABSTRACT: This paper focuses on the problem of short text summarization on the comment stream of a specific message from social network services (SNS). Due to the high popularity of SNS, the quantity of comments may increase at a high rate right after a social message is published. Motivated by the fact that users may desire to get a brief understanding of a comment stream without reading the whole comment list, we attempt to group comments with similar content together and generate a concise opinion summary for this message. Since distinct users will request the summary at any moment, existing clustering methods cannot be directly applied and cannot meet the real-time need of this application. In this paper, we model a novel incremental clustering problem for comment stream summarization on SNS. Moreover, we propose IncreSTS algorithm that can incrementally update clustering results with latest incoming comments in real time. Furthermore, we design an at-a-glance visualization interface to help users easily and rapidly get an overview summary. 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.

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

I. INTRODUCTION

In recent years, social network services (SNS) are prevalent and have become important communication plat-forms in our daily life. According to the 2012 statistics by the largest social net working site Face book. 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. In this paper, 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. 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 You Tube and Face book, 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. On the other hand, some researchers model this problem as recommendation [35],[17], [54], [16], [12] or classification [53], [48], [45] tasks and employ machine learning techniques to solve it. Moreover, sentiment analysis [6], [21], [36], [15], [55] has been applied as well to discover hidden emotions in messages. Further-more, providing an informative presentation inter face [40], [50], [58] is another active research field on the summarization of social messages. As will be thoroughly surveyed in Section 2.2, 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.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

II. RELATED WORK

Owing to the large quantity of user-generated data on SNS, the research topics on alleviating the information overload problem and discovering useful knowledge have attracted much attention recently. In addition to the comment stream data discussed in this paper, previous works also target at different types of social data and explore various research topics related to solving the information overload problem on SNS. In this section, we can broadly classify these works into five categories: 1) Human-assisted mechanisms, 2) Summarization, 3) Rating and filtering, 4) Topic and event detection, and 5) Sentiment analysis. Note social network services are not restricted to well-known social web sites, such as Face book, Twitter, etc. The web services providing interaction functionality for users can be generally included**.**

Human Assisted Mechanisms. The main notion of this category is to highlight significant comments or provide summary by the assistance of user Feedback and judgment. For instance, users can "like" a comment on popular social websites, such as YouTube [7], Face book [3], and Amazon [1]. The comments with more "likes" will be displayed on the top of the list. Undoubtedly, human judgment is able to produce most correct results. However, since a user is unlikely to skim all comments and evaluate the goodness of each one, good comments may still be ignored.

Summarization. Regarding the research field of short text summarization, in recent years, numerous works [40], [50], [58], [52], [27] are focused on micro-blogging messages. A variety of techniques have been developed and applied to satisfy different needs of summarization. In [40], a visualization system Twit Info is presented to enable the convenient browsing of a large collection of Twitter messages (also known as tweets)by detecting and highlighting peaks of highly-discussed activity. Another map-like presentation system Twitter Stand [50] further considers incorporating geographic location of tweets to automatically obtain late breaking news. In addition, with collections of short posts on a specific topic, the authors in [52] aim to create short summary sentences that best describe the primary gist of what users are saying about. With similar intention of [52], work of [27] employs both generative model and user behavior model to synthesize content from micro-blogging messages on the same topic into a prose description of fixed length.

Rating and Filtering. Some researchers attempt to relieve the information over-load problem by selecting representative messages that better express group opinions or contain significant information. The rating mechanism [35], [17], [54], [16], [12]is widely developed to determine the importance of messages. In addition, several types of filtering approaches [53],[45], [29], [19] have also been devised to keep important messages and exclude redundant ones. The work of [35] aims to select the best top-k informative comments from a set of user-contributed comments for a specific object, such as a video. Initially, a modified model of Latent Dirich let Allocation is applied to cluster comments into several groups based on the concept of topic modeling. Then, the authors propose a precedence-based ranking approach to select informative comments for each cluster. Note that the intention of this work [35] is most relevant to this paper.

Topic and Event Detection. The key motive of the topic detection on SNS is to help users facilitate the social stream understanding [38], [13], [43],[42]. In [42], the authors propose to generate an entity-based topic profile for each user by examining the entities this user mentions in his/her previous tweets. Works of [13]and [43] aim to develop the topic-based browsing interfaces. Trending topics will be clustered for more directed exploration [13], and a set of themes will be extracted and grouped for faceted navigation [43]. Moreover, the authors in [38]further incorporate multiple text sources to enhance the effectiveness of topic summarization on Twitter.

Sentiment Analysis. Many researchers explore the sentiment analysis [44] to discover public mood and emotion hidden in social messages. In general, sentiment classification model is employed to classify messages into pre-defined sentiment labels, such as positive and negative. The system Tweet Feel [6] is a real-time Twitter search system that can estimate the ratio of positive to negative tweets mentioning a specific search keyword. In [21], the authors propose to appraise the presidential debate performance via Twitter plat from according to the numbers of positive and negative tweets. In addition to the binary classification, the authors in [36] and [15] further consider evaluating each message by aggregating the sentiment distribution over six sentiment labels. Besides, predicting election results based on the detection of political sentiment on Twitter has also been investigated in [55].

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

III. PROPOSED ALGORITHM

A. *Design Considerations:*

In this section, we aim to develop efficient approaches in discovering top-k groups of opinions towards a specific message on SNS. A batch version of short text summarization algorithm is first introduced.

B. *Description of the Proposed Algorithm:*

Batch STS Algorithm: Batch Version According to the problem definition of we propose the algorithm Batch STS that is the batch version for solving this problem. The algorithmic form of Batch STS in outlined in Algorithm 1.

Batch STS takes the whole comment set S as the input. The second input is the radius threshold u_r used for deter-mining how similar the comments are in a cluster.

There are two main steps in Batch STS. The aim of the first step, shown in lines 2-11 of Algorithm 1, is to find all connected components of the comment set S. The points belonging to the same connected component will be merged as a cluster. It can be imagined that there will be a link between two comments as their distance is not infinite. In lines 2-7, for each comment v_i of S, we examine whether there is any existing cluster C_j where disðv_i; C_jÞ is not infinite. If there is, this comment will be added into anyone of these clusters. Otherwise, a new single-point cluster is formed with the comment v_i. In lines 8-11, we calculate the distances between two centers of any pair of non-single-point clusters, and if the distance between two clusters is not infinite, they will be merged together. Subsequently, in lines 12-17, the objective of the second main step is to guarantee the radius of each cluster is smaller than the threshold u_r. To meet this requirement, for each non-single-point cluster C_i, we exclude the comment v_j where disðv_j; C_i Þ is larger than or equal to u_r. Meanwhile, in line 17, it will be checked whether v_j can be merged with other excluded comments. After this step, all clusters will satisfy the radius restriction, and finally, Batch STS outputs the top-k clusters with top-k most comments.

IV. PSEUDO CODE

Step 1: Finds all connected components of the comment set S..
Step 2: The second input is the radius threshold u_r used for deter-mining how similar the comments are in a cluster.
Step 3: If There exists any cluster C_j *Where* $dis(v_i,C_j)$ *is not infinite.*
Step 4: Add v_i into anyone of these clusters;
else
Step 5 : Form a new cluster C_{new} with the comment v_i;
$C = C \sqcup C_{new}$;
Step 6: **for** each non-single-point element C_i of C
Step 7: **while** the radius of C_i is larger than or equal to θ_r **for** each comment v_j in C_i
Step 8: **if** $dis(v_j,C_i) \geqq \theta_r$
Step9: Check whether v_j can be merged with other excluded comments;
Step10: Output top-k clusters in C which have top-k most comments;

End

V. SIMULATION RESULTS

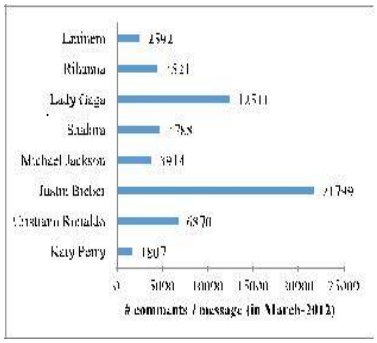
We target at comment streams in SNS that are in short text style with casual language usage. For each social message, our main objective is to cluster comments with similar content together and generate a concise opinion summary for this message. We want to discover how many different group opinions 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.

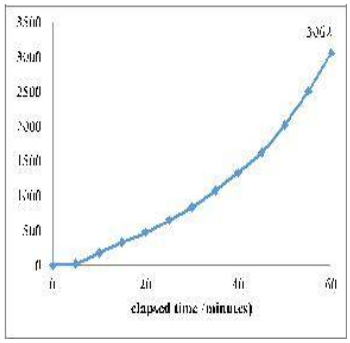
ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016





VI. CONCLUSION AND FUTURE WORK

In this paper, to enable the capability of comment stream summarization on SNS, we model a novel incremental clustering problem and propose the algorithm Concise Opinion Summary, which can incrementally update clustering results with latest incoming comments in real time. With the output of Concise Opinion Summary, 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 Concise Opinion Summary possesses the advantages of high efficiency, high scalability, and better handling outliers, which justifies the practicability of Concise Opinion Summary on the target problem. , this data structure design will incur additional storage space. In this section, we study the effects of efficiency and storage space caused by this design for the execution time.

REFERENCES

[1] Amazon [Online]. Available: http://www.amazon.com/, 2014.
[2] Experimental Demo Page [Online]. Available: http:// 140.109.21.214/public/IncreSTS/index.htm, 2014.
[3] Facebook [Online]. Available: http://www.facebook.com/, 2014.
[4] HotelClub [Online]. Available: http://www.hotelclub.com/, 2014.
[5] TripAdvisor [Online]. Available: http://www.tripadvisor.com/, 2014.
[6] TweetFeel [Online]. Available: http://www.tweetfeel.com/, 2014.
[7] YouTube [Online]. Available: http://www.youtube.com/, 2014.
[8] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "On-line new event detection and tracking," in Proc. 21th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 1998, pp. 37–45.
[9] 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.
[10] S. Baccianella, A. Esuli, and F. Sebastiani, "Multi-facet Rating of Product Reviews," in Proc. 31st Eur. Conf. IR Res. Adv. Inf. Retrieval, 2009, pp. 461–472.

BIOGRAPHY

Mrs.R.RAJANI is an Associate Professor and head of the department of MCA, Narayana Engineering College, Nellore, AP, India. She is pursuing her Ph.D from Sri Padmavathi Mahila University. She guided many projects for B.Tech and PG students. Her research interests include Datamining, Query Optimization, Concise Opinion Summary,Computer Networks and Software Engineering etc.,

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Ms. Deepthi G, completed master of computer applications in Narayana engineering college, Nellore. Her research interest is Concise Opinion Summary.

Ms. Sujatha P, completed master of computer applications in Narayana engineering colleg, Nellore. Her research interest is Concise Opinion Summary.

.