Clustering Article Comments Based On Text Usage

Anusree K C
Dept. of ISE
RV College of Engineering
Bangalore, India
anusreeanil759@gmail.com

Prof. Srinivas B K
Assistant Professor
Dept. of ISE
RV College of Engineering
Bangalore, India
bksrinivas@rvce.edu.in

Abstract:-Nowadays, everything is available in the web, like articles, books etc. We need to find whether its good sorting out the best is a time consuming task. This paper helps to sort out comments on articles, with same meaning. Batch Short Text Summarization (BSTS), is a technique which is done as a addition to the existing system i.e., short text summarization in social networks like Facebook. The output shows that the system provides an easier way for users to get a brief idea of articles comments without actually reading comments. The system can be run on any online websites which makes it more reliable than the exisisting system.

Keywords – Short Text Summarization,Batch-STS, N-Gram, Comment Vector, Key-Term Extraction.

I.INTRODUCTION

Batch Short Text Summarization (BSTS), a summarization system helps to sort out comments in an articles with same meaning. The users always to get a brief idea of articles comments without actually reading comments.The proposed approach groups comments with content similarity, semantic similarity and make ashort text summary of article comments.

This paper focuses on short text comments and informal language style to provide immediate summary of article comments in real time. The proposed system makes use of BSTS algorithm which can update clustering results with latest incoming comments in real time. Grouping similar comments leads to formation of different clusters. Key-term extraction algorithm helps in identifying the count of keys from the cluster centre.Finally a visualization interface is designed for presenting the summarized result.For each comment, the main objective is to discover top cluster comment groups with content similarity, semantic similarity and generate a summary for the comments. The purpose of summarization is to determine how many different group opinions exist and also gives an overview of each group.

II. LITERATURE SURVEY

In recent years, numerous works are focused.For summarization purpose different kinds of techniquesis used. A summarization system implemented on social media[1] networking website (Facebook) summarizes the comments received in a web page into various groups. Cheng-Ying Liu, Ming-SyanChen,Chi-Yao Tseng describes IncreSTS algorithm that can incrementally update clustering results with latest incomingcomments in real time. From the experimental results and demonstration shows the advangtages of IncreSTS algorithm. By using the IncreSTS algorithm provides high efficiency, high scalability etc[1].

IMASS[2] system for summarizing micro blog post and responses with goal to provide readers a more constructive set of information for efficient digestion. The authors of [2] introduce a two-phase summarization scheme. Based on the intention, sharing, discussion and chat, the post plus its responses are classified as four categories in first phase. For each type of post, in the second phase, the system chooses different strategies response pair identification, including opinion analysis, and response relevancy detection, to summarize and highlight critical information to display.

The popularity of social network services and micro-blogging websites, blog is the primary platforms that publish content. For the summarization of blogs, main research direction is to extract and discover representative sentences. Authors [3] consider utilizing feedback comments foridentifying important sentences from a blog post. The proposed sentence scoring mechanism was based on the observation that user-contributed comments will provide valuable information for better understandingof the blog content. For selecting best top informative comments from set of user-contributed comments for a specific object, such as video, E. Khabiri, J. Caverlee, and C.F. Hsu [3] proposed a approach. Wherea modified model of Latent Dirichlet Allocation (LDA) was applied in cluster comments to several groups based on the concept of topic modeling. Then, a precedence-based ranking approach is proposed for selecting informative comments for each cluster.

With flourish of Web, online review [4][5] has become more useful and important information resource for people. Different from traditional text summarization, review mining and summarization aims at extracting the features from which the reviewers express their opinions and determining, the opinions are positive or negative.

Authors in [4] focuses on a specific domain – movie review. Different from product reviews, movie reviews have some unique characteristics. The commented features in movie review are richer than those as product review. This paper shows a multi-knowledge based approach is proposed for movie review mining and summarization.

III. METHODOLOGY OF THE PROPOSED SYSTEM

In this proposed system various steps are used to generate a summary ofan article comments.

A. Architecture Diagram of the Proposed System

Once anarticle comments posted on webpage, users can view the comments immediately.To provide immediate summary of article comments in real time, the proposed system makes use of BSTS algorithm which can updateclustering results with latest incoming comments in real time. Grouping similar comments leads to formation of different clusters. Key-term extraction algorithm helps in identifying the count of keys from the cluster centre. Finally a visualization interface is designed for presenting the summarized result.The system architecture of clustering article comments is depicted in Fig.1.

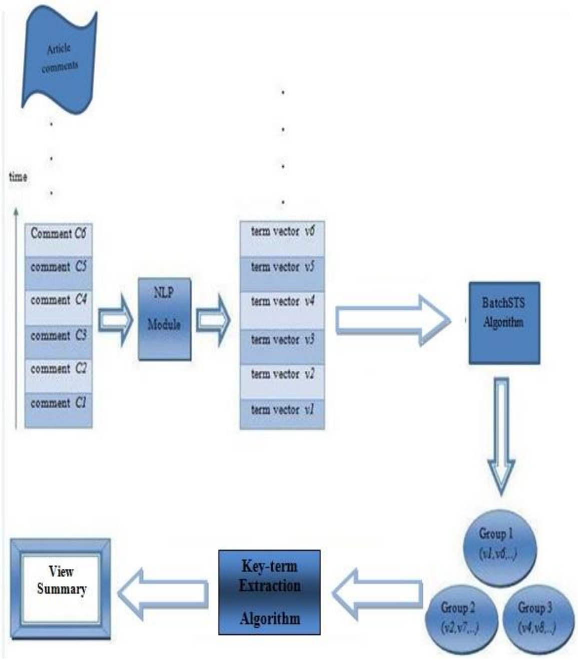


Fig.1 Architecture Diagram of Clustering article comments

B. N-Gram

N-Gram is a contiguous arrangement of n items from a given array or sequence of text or speech. Anuni-gram is an n-gram where the value of n is 1, bi-gram is an n-gram of size 2, tri-gram is an n-gram of size 3. When the size of n-gram is large then it is referred to by the value of n such as ‘four-gram’, ‘five-gram’ and so on.

C. Term Vector Representation of Comments

Each comment converted into a set of n-gram terms. Term vector can be represented as 0’s and 1’s.The first step is the stemming process. Porter stemming algorithm [8]is to reduce inflected and derived words to their stem form (e.g., loving is turned into love). Subsequently, the process of n-gram terms extraction is carried out to extract terms that are used for representing this comment.

D. Clustering

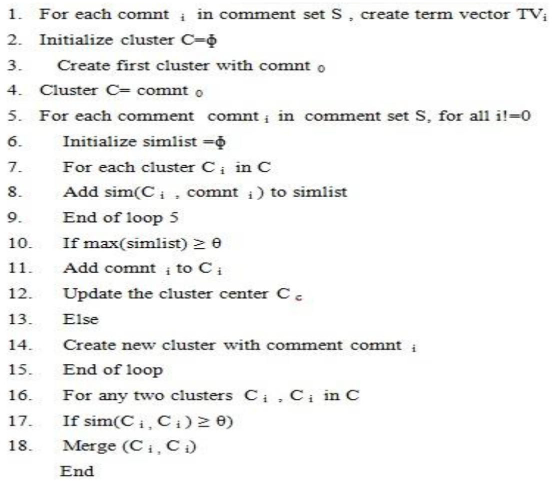
The task of grouping a set of objects in such a way that objects in the same group are more similarto each other than to those in other groups is known as clustering. Consider two comments represented in the term vector model, υa = (t1,a, t2,a, ... tN,a) and υb =(t1,b, t2,b,...., t N, b). N is the number of dimensions. Then calculate the content similarity value of the comments using cosine similarity equation.

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

Where υa .υb is the inner product of two vectors, and D is a positive integer constant.

***Batch-STS algorithm**:-*

Theproposed system makes use of BSTS algorithm which can update clustering results with latest incoming comments in real time. Grouping similar comments leads to formation of different clusters.The whole comment set S and the radius threshold θrcan be considered as the input Batch-STS algorithm.Threshold θr used for determining how similar the comments are in a cluster. Batch-STS consist of two steps. The first step focuses to find all connected components of the comment set S. The points belonging to the same connected component will be merged asa clusters. The second step focuses to guarantee the radius of each cluster is smaller than the threshold θr.

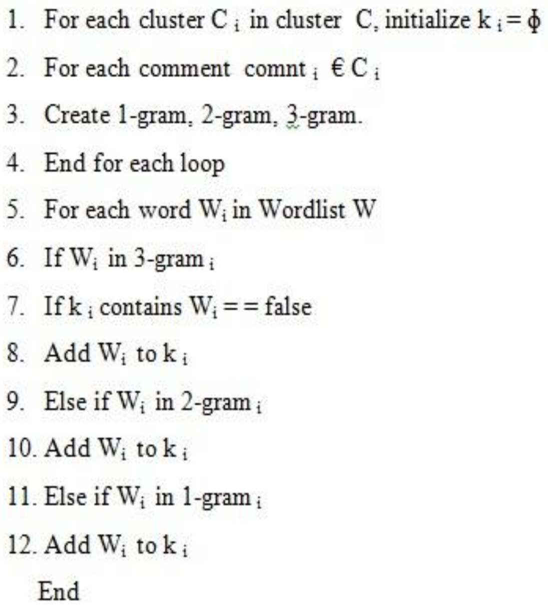


```
1.  For each comnt i in comment set S , create term vector TVi
2.  Initialize cluster C=Φ
3.    Create first cluster with comnt 0
4.  Cluster C= comnt 0
5.  For each comment comnt i in comment set S, for all i!=0
6.    Initialize simlist =Φ
7.    For each cluster C i in C
8.    Add sim(C i , comnt i ) to simlist
9.    End of loop 5
10.   If max(simlist) ≥ θ
11.   Add comnt i to C i
12.   Update the cluster center C c
13.   Else
14.   Create new cluster with comment comnt i
15.   End of loop
16.   For any two clusters C i , C i in C
17.   If sim(C i , C i ) ≥ θ)
18.   Merge (C i , C i)
      End
```

Figure2Batch-STS algorithm

Key-Term Extraction:-

Key-term extraction algorithm helps in identifying the count of keys from the cluster centre. N – gram and clusters are the input of this algorithm.



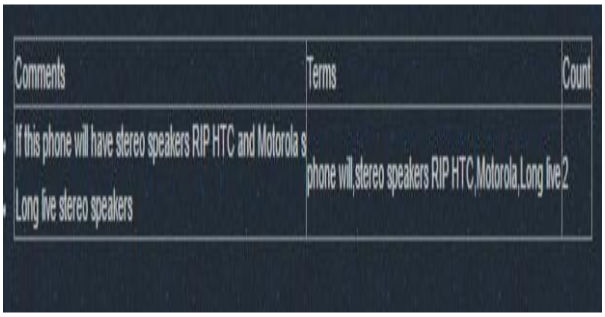
```
For each cluster $C_i$ in cluster C, initialize $k_i = \phi$
For each comment $comnt_i \in C_i$
Create 1-gram, 2-gram, 3-gram.
End for each loop
For each word $W_i$ in Wordlist W
If $W_i$ in 3-gram $_i$
If $k_i$ contains $W_i$ = = false
Add $W_i$ to $k_i$
Else if $W_i$ in 2-gram $_i$
Add $W_i$ to $k_i$
Else if $W_i$ in 1-gram $_i$
Add $W_i$ to $k_i$
  End
```

Figure 3Key-Term extraction algorithm

IV. IMPLEMENTATION

For testing the system, first awebpage is chosen. The website must contain article comments, for time being gsm arena is taken as our test webpage for article comment summarization.Intel Pentium Core i3 Personal laptop and 2GB memory is used for this experiment..

The proposed scheme is tested, by searching different article pages using web application. In the experiment, we use HTML agility pack tool to crawl article comments from a website. Each commentsare converted into N-GRAM's, then they are represented in term vector model as 0's & 1's. BatchSTS algorithm is used to group comments into batches, from batches Key term is extracted using Key-Term extraction algorithm. And thus batches are created according to Key terms, thus getting the summary. The results are as follows:



Comments	Terms	Count
• If this phone will have stereo speakers RIP HTC and Motorola s • Long live stereo speakers	phone will,stereo speakers RIP HTC,Motorola,Long live	2

Figure 4Summary of Article Comments

V. CONCLUSION

To provide the capability of comment stream summarization on articles, which can update clustering results with latest incoming comments in real time. Finally a web application is designed for presenting the summarized result.This paper helps to sort out comments on articles, with same content, so that a user can classify the category of article based on the clustered comments.

In the future work, article comments from the websites can be summarized. In this paper helps to summarize article comments only from a single webpage.

REFERENCES

[1] Cheng-Ying Liu, Ming-SyanChen,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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] M. F. Porter. "An Algorithm for Suffix Stripping. Program", pages 130–137, 1980.