ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

Online Hybrid Tweet Summarization and Query Reformation for Linguistic features using Hadoop

G.G.Sreeja [1], R.P.Narmadha [2]

P.G. Student, Department of Computer Science & Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India[1]

Assistant Professor(Senior), Department of Computer Science & Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India[2]

ABSTRACT: Tweet Segmentation is usually constructed to cluster the tweets with predefined filtering rules and criteria's but normal hybrid segmentation fails during tweets are being created and shared at an unprecedented rate with enormous amount of data noises and redundancy. Hence, an efficient tweet stream clustering algorithm named as an online algorithm allowing for effective clustering of tweets with only one pass over the data. This algorithm employs two data structures to keep important tweet information in clusters. The first one is a novel compressed structure called the Tweet Cluster Vector (TCV). TCVs are considered as potential sub-topic delegates and maintained dynamically in memory during stream processing. The second structure is the Pyramidal Time Frame (PTF) which is used to store and organize cluster data at different moments. Given a timeline range, the summarization system may produce a sequence of time stamped summaries to highlight points where the topic/subtopics evolved in the stream. Summarization represents a set of documents by a summary consisting of several sentences. Intuitively, a good summary should cover the main topics (or subtopics) and have diversity among the sentences to reduce redundancy. In general, tweet summarization has taken the temporal feature of the arriving tweets into consideration and the TCV-Rank summarization algorithm is applied to generate summaries. Experimental on the synthesis dataset demonstrated the efficiency, flexibility and cluster formation for topic evolution.

KEYWORDS: online algorithm, Tweet Cluster Vector, Pyramidal Time Frame, TCV-Rank summarization, generate summaries.

I. INTRODUCTION

Micro blogging services such as twitter are a web service that allows the subscriber to share and disseminate short messages to other subscribers of the service in terms of opinions, News updates, and Promotional contents and as other valuable sources [6]. It is used to gather audience opinion for different important topics. Micro-blogging acts as platform for sharing status updates, marketing Product etc. The short nature of updates allows users to post news items quickly, reaching its audience in seconds. The Performance of Service is high due to retrieval of information is low. Micro blogging has noticeably revolutionized the way information is consumed. It has empowered citizens themselves to act as sensors or sources of information that could lead to consequences and influence, or even cause, media coverage. People now share what they observe in their surroundings, information about events, and their opinions about topics from a wide range of fields. Moreover, these services store various metadata from these posts, such as location and time. Aggregated analysis of this data includes different dimensions like space, time, theme, sentiment, network structure etc., and gives researchers an opportunity to understand social perceptions of people in the context of certain events of interest.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

Classification of the short text messages is open research in the online social network and micro blogging sites. The research challenges in the classification of short text messages are due to large volumes of data produced everyday with noisy and short nature of tweets carrying user sentiments and opinions. For data Analysis and summarization, large of redundant data and noisy data has to be eliminated. Clustering technique has been a solution to generate the data summarization using data structures and ranking techniques to the text feature and linguistic features. The topic evolution methods are incorporated to each cluster to generate summaries based time variations and volume variations.
To overcome the challenges, Tweet Summarization has been done from the segmented tweet data. Summarization is the set of documents which contains the most important generated topics that has several sentences. This provides a precise meaning to the tweets which helps different users to understand the tweets very clearly and also most of the times repeated topics will be shown to the users. The same type of tweets may be tweeted by the users at different arbitrary time durations.

Data Classification Technique for Text Message

Many short message classification techniques rely on the local and global context with respect to the linguistic features of the message which are as follows.

- **K- Nearest Neighbour:** KNN-based classifier is used to conduct word-level classification, leveraging the similar and recently labeled tweets.

- **Conditional Random Field (CRF):** It is class of statistical modeling method often applied in pattern recognition and machine learning, where they are used for structured prediction. Whereas an ordinary classifier predicts a label for a single sample without regard to "neighbouring" samples, a CRF can take context into account; e.g., the linear chain CRF popular in natural language processing predicts sequences of labels for sequences of input samples. CRF's are a type of discriminative undirected probabilistic graphical model. It is used to encode known relationships between observations and construct consistent interpretations. It is often used for labelling or parsing of sequential data, such as natural language text or biological sequences and in computer vision.

Data Summarization-Topic Generation Models

The Methods are designed and implemented to produce the timelines automatically against the volume variation and time variation data's. Summarization represents a set of documents by a summary consisting of several sentences. Intuitively, a good summary should cover the main topics (or subtopics) and have diversity among the sentences to reduce redundancy. Summarization is extensively used in content presentation.

II. RELATED WORK

Named Entity Recognition (NER) and Entity Linking (EL) [1] were used to find different names which are present in the tweet. The author K.L.Liu [5] used to identify the attitude or opinions of the tweets. [6] extracted how the automatic opinion summarization poses a greatest challenge to the summarization system by considering the users opinions which are posted in the twitter. The author C.Li [3], proposed NER to discover the named entities automatically. [7] [2] demonstrates how to understand the precise meaning of short text messages, order of words and the phrases present in the tweet.
The main aim of this paper is to collect the different set of documents which is full of summaries. It is one of the obvious ways of generating summaries from the segmented tweet data. Initially, the tweets are partitioned and pre-processed for removing the unnecessary words through stop word removal and stemming process. Parts Of Speech (POS) tagging [4] will be done for the pre-processed tweets which will be highly useful for generating different clusters. Then, number of times repeated words must be calculated and kept separately. This will be maintained for generating most important topics that has been tweeted in the twitter. Through this method, tweets can be understandable to everyone easily.

Problems in Existing Solution

Tweet segmentation has to be used for segmenting the tweets based on different class labels contain in the tweets. Unsupervised learning models has to be constructed in order to learn the tweets. By splitting tweets into meaningful

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

segments, the semantic or context information is well preserved and easily extracted by the downstream applications. [10] HybridSeg is proposed segmentation technique which finds the optimal segmentation of a tweet by maximizing the sum of the stickiness scores of its candidate segments. Experiments on two tweet data sets show that tweet segmentation quality is significantly improved by learning both global and local contexts compared with using global context alone. Through analysis and comparison, it has been shown that local linguistic features are more reliable for learning local context compared with term-dependency [5].

Proposed Solution

Microblogging services, such as Twitter, have become popular channels for people to express their opinions towards a broad range of topics. The tweets may can contain any misspellings, ill-formed words, grammatical errors which allows the people to misunderstand the meaning of the tweets. So, the sentences can be segmented based on the POS tagging. The term frequency of the word will be estimated which shows how important that a word is for the document and it helps to establish a cluster. Then tweets are classified into different class labels based on term frequency and linguistic features.Among the different clusters formed, optimal clusters will be identified by using TCVie; it generates a sub-topic delegates.

III. FLOW OF SEGMENTATION & SUMMARIZATION

Tweet Segmentation

Initially as in fig 1, tweets are preprocessed for removing the words except the technical words through stop word removal and stemming process. After preprocessing, POS tagging will be done for the preprocessed tweets. Then consideration of local and global context for calculating the term frequency. Feature selection is provided for checking whether the tweet occurred is local context or global context. Term frequency is nothing but the number of times repeated words present in the document will be shown by considering the threshold values.ie; whether it is minimum or maximum. If the threshold value is less than 50 then it is considered as a minimum segmentation ie; less number of times repeated words and if the threshold value is between 50-100 then it is considered as maximum segmentation ie; most number of times repeated words in the document. As an application, high accuracy can be achieved in named entity recognition by applying segment-based POS tagging.

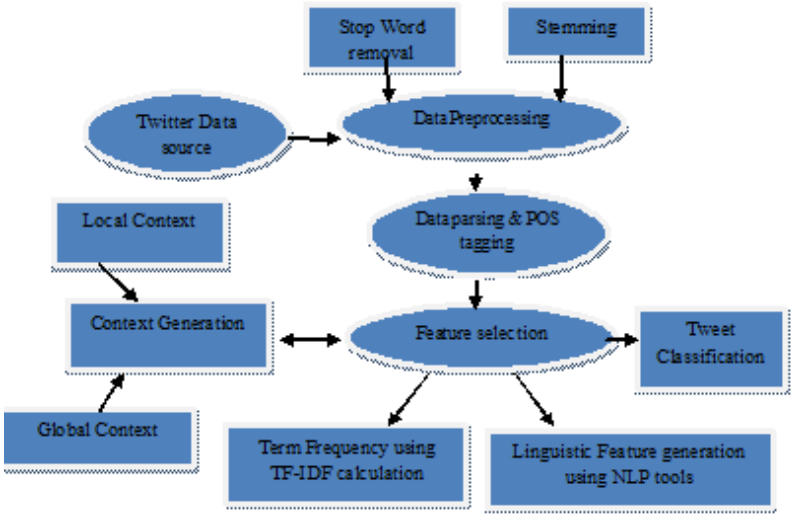

Fig 1: System Architecture for Tweet Segmentation

Tweet Summarization

Fig 2, is described as follows. The term frequency calculations for repeated words has been done and formed into different classes. This will be then clustered based on the different types of tweets and different operations like merging,

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

insertion and deletion of tweets can be done. From the formed clusters, the most repeated topics will be considered. Then, the greedy algorithm is applied to select representative tweets to form summaries. This algorithm first computes centrality scores for tweets kept in TCVs, and selects the top-ranked ones in terms of content coverage and novelty. The generated summaries can be of 2 forms. One is online summaries based on volume variations and the other one is online summaries based on arbitrary time durations.

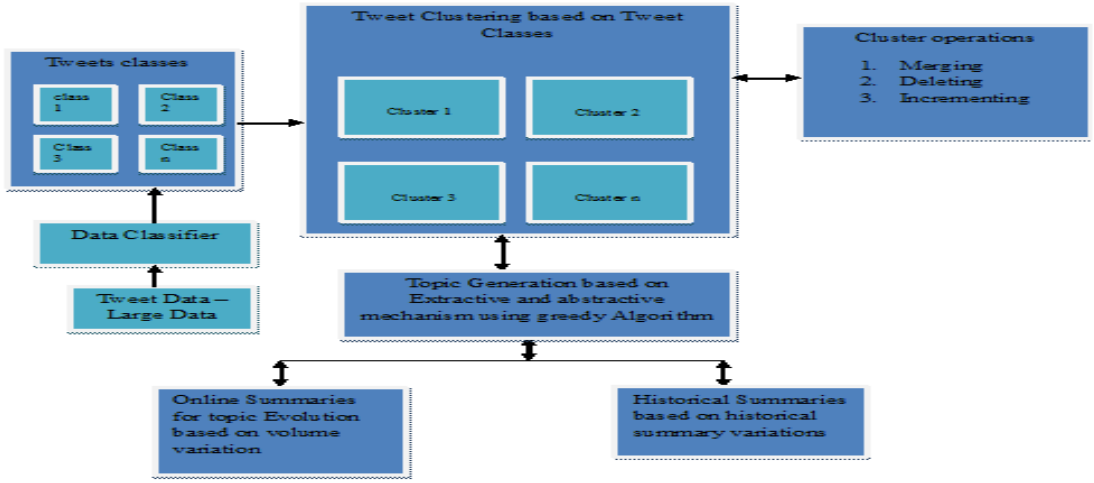

Fig 2: System Architecture Diagram for Tweet Summarization

IV. EXPERIMENT

Real time dataset has been considered for analyzing the tweets and generating the summaries. Our experiment consists of following 7 parts which can be described as follows:-

Preprocessing the Tweets

The Hadoop file system is configured with database software which as temporary storage media and to ease data retrieval mechanism before establishing Hadoop cluster in the Hadoop file system, in this module, preprocessing the tweet data using the stop word removal and stemming process will be performed.

Labelling based on the POS tagger

POS tagger is used to tag the data after parsing mechanism; the parsed data is labelled using parts of speech based on the data placement. It is further used for text mining and estimating the term frequency of the files. POS tags of the constituent words in segments. The segments that are likely to be a noun phrase (NP) are considered as named entities.

Calculating the Term Frequency

It is used for Class formation to global Context and local Context. It is a numerical statistic that is intended to reflect how important a word is to a document in a collection in file system. It is often used as a weighting factor in information retrieval and text mining. Term frequency value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the twitter parsed data collection, which helps to adjust for the fact that some words appear more frequently in general. Term frequency estimation is used to establish a cluster and to return a record based on the query input.

Classification of Tweet data based on different class

In this module, classify the tweet into different classes based on term frequency and Linguistic features of the tweet data. Many named entities and common phrases are preserved in tweets for information sharing and dissemination. Although many tweets contain unreliable linguistic features like misspellings and unreliable capitalizations, there exist tweets composed in proper English. The normalization of the data is carried out using different techniques

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

- **Length normalization**

As the key of tweet segmentation is to extract meaningful phrases, longer segments are preferred for preserving more topically specific meanings. At first glance, it seems that applying maximum likelihood estimation is straightforward.

- **Right-to-left smoothing (RLS):**

Like most n-gram models, the follows the writing order of left-to-right. However, it is reported that the latter words in a n-gram often carry more information.

- **Bursty-based weighting:**

The estimation of local collocation can be combined with global context using a linear combination with a coupling factor. However, because tweets are noisy, the estimation of an n-gram being a valid segment is confident only when there are a lot of samples. Hence, global context in tweet segmentation has been preferred when the frequency of an n-gram is relatively small. Therefore, bursty-based weighting scheme for combining local collocation and global context has been introduced.

Establishing the Tweet Cluster Vector from Term Frequency

The clusters generated from the term frequency contains the most repetitive set of words which is the important tweet information considered as sub-topic delegates. A data structure is required to maintain this important tweet cluster information. The Tweet Cluster Vector (TCV) is the data structure which handles potential sub-topic delegates and maintained that dynamically in memory during stream processing.

Clustering of tweets based on the different time frame using PTF

An efficient tweet stream clustering algorithm called as online algorithm allowing for effective clustering of tweets with only one pass over the data. This algorithm employs two data structures to keep important tweet information in clusters. The first one is a novel compressed structure called the TCV. TCVs are considered as potential sub-topic delegates and maintained dynamically in memory during stream processing. The second structure is the Pyramidal Time Frame (PTF) which is used to store and organize cluster snapshots at different moments, thus allowing historical tweet data to be retrieved by any arbitrary time durations. Given a timeline range, the summarization system may produce a sequence of time stamped summaries to highlight points where the topic/subtopics evolved in the stream.

Generating the summaries of the cluster data for Topic Generation using Greedy Algorithm

Summarization represents a set of documents by a summary consisting of several sentences. Intuitively, a good summary should cover the main topics (or subtopics) and have diversity among the sentences to reduce redundancy. A greedy algorithm has been adopted to select representative tweets to form summaries, this algorithm first computes centrality scores for tweets kept in TCVs, and selects the top-ranked ones in terms of content coverage and novelty. To compute a historical summary where the user specifies an arbitrary time duration, two historical cluster snapshots from the PTF with respect to the two endpoints (the beginning and ending points) of the duration has to be retrieved. Then, based on the difference between the two cluster snapshots, the TCV-Rank summarization algorithm is applied to generate summaries.

V. PERFORMANCE EVALUATION

Performance Analysis for Tweet Segmentation

For the given time duration, first retrieve the corresponding tweet subset, and use the following three well-recognized summarization algorithms to get three candidate summaries. Tweet segmentation helps to preserve the semantic meaning of tweets, which subsequently benefits many downstream applications, e.g., NER. Through experiments, a segment-based named entity recognition method achieves much better accuracy than the word-based alternatives in terms of inter cluster similarity and execution time has been shown in fig 3.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

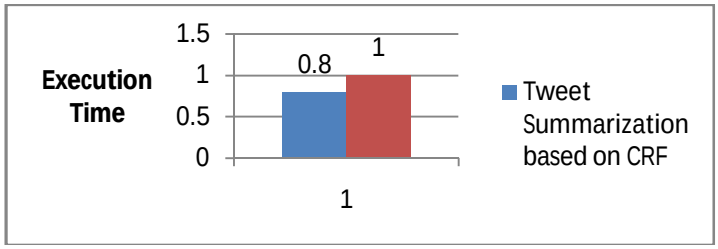

Fig 3: Performance evaluation of the Tweet Segmentation using PCA

Performance analysis on F-measure

Data classification helps to preserve the semantic meaning of tweets, which subsequently benefits many downstream applications, e.g., named entity recognition for further generation of Summary.

Through experiments, the segment-based named entity recognition methods achieves much better accuracy than the word-based alternative in terms of inter cluster similarity and execution time has been shown in the table 1 and local linguistic features are more reliable than term-dependency in guiding.

Table 1: Comparison between the algorithms based on different measures

Classification technique	Cluster Size	Execution time	Precision
Hidden Markov model	4.15	80ms	1.00
Conditional random Field	3.75	55ms	0.80
Principle Component Analysis	2.00	45ms	0.45

Performance analysis based on cluster size

In figure 4, it is clear that optimal cluster formation based on the data classification technique in PCA is Proved to be efficient against the other existing classifier like conditional random field and Hidden markov Model. The proposed summarization technique is the good solution for continuous summarization which addresses the following three issues:

- **Efficiency**—tweet streams are always very large in scale, hence the summarization algorithm is highly efficient;
- **Flexibility**—it provides tweet summaries of arbitrary time durations.
- **Topic evolution**—it automatically detect sub-topic changes and the moments that they happen.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

Fig 4: **Cluster analysis of the hybrid classification**

VI. CONCLUSION

This project has been developed to provide the opinion summary which can reduce the execution time and make this method more reliable. The opinion summarization can be generated by considering both the global and local context. The basic modules for the tweet summarization like segmenting the tweets, identifying named entities using term frequency has been analyzed and performed in this phase. It is done by using different methods like data pre-processing, POS tagging and tweet classification. This can be helpful for the users to understand the tweets easily. Tweet segmentation provides much better accuracy than the word based alternative in terms of inter cluster similarity and execution time. Intuitively, a good summary should cover the main topics (or subtopics) and have diversity among the sentences to reduce redundancy. Rank summarization algorithm is applied to generate summaries. Experimental on the synthesis dataset demonstrated the efficiency, flexibility and cluster formation for topic evolution.

REFERENCES

[1] A. Sil and A. Yates, "Re-ranking for joint named-entity recognition and linking," in Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage.,2013, pp. 2369–2374

[2] B. Han and T. Baldwin, "Lexical normalisation of short text messages: Maknsens a #twitter," in Proc.49thAnnu. Meeting.Assoc.Comput. Linguistics: Human Language Technol., 2011, pp. 368–378.

[3] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee, "Twiner: Named entity recognition in targeted twitter stream," in Proc. 35th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2012, pp. 721–730.

[4] Chenliang Li, Aixin Sun, JianshuWeng, and Qi He, Member, "Tweet Segmentation and Its Application to Named Entity Recognition", IEEE transaction, VOL. 27, NO. 2, FEBRUARY 2015.

[5] K.-L. Liu, W.-J.Li, and M. Guo, "Emoticon smoothed language models for twitter sentiment analysis," in Proc. AAAI Conf. Artif.Intell., 2012, pp. 1678–1684.

[6] X. Meng, F. Wei, X. Liu, M. Zhou, S. Li, and H. Wang, "Entity centric topic-oriented opinion summarization in twitter," in Proc.18th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining, 2012, pp. 379–387.

[7] X. Wang, A. McCallum, and X. Wei, "Topical n-grams: Phrase and topic discovery, with an application to information retrieval," in Proc. IEEE 7th Int. Conf. Data Mining, 2007, pp. 697–702.

[8] Zhenhua Wang, LidanShou, Ke Chen, Gang Chen, and SharadMehrotra, "On Summarization and Timeline Generation for Evolutionary Tweet Streams", IEEE transaction, VOL. 27, NO. 5, MAY 2015.