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.5, Special Issue 2, April 2017

An International Conference on Recent Trends in IT Innovations - Tec'afe 2017

Organized by

Dept. of Computer Science, Garden City University, Bangalore-560049, India

Automatic Summarization of Public Events using Real-time Twitter Data-Stream

Muruganantham A[1], Banu Shree M[2], Geetha N[3]

Associate Professor, Department of Computer Science (PG), Kristu Jayanti College, Bengaluru, India[1]

MCA Student, Department of Computer Science (PG),Kristu Jayanti College, Bengaluru, India[2,3]

ABSTRACT- The radical growth of information in the web leads to prolific increase of research in the field of information retrieval. Information retrieval is the process of searching and extracting the required information from the web. The search engines of present era presents infinite hyperlinks of information, which suffers from redundancy and irrelevance. Therefore there is a need for Summarizing information available in the web. Text Summarization presents large text data into a shorter version without changing its meaning, which is easily readable, understandable and complete. The steps towards summarization are ranking sentences in order of importance and extraction of the most significant sentences. In this paper we summarized the report for the event Aero India 2017, Asia's Premier Air Show. We collected tweets related to the respective event through Twitter API and summarized the tweets to a short report that includes the important links of the most viewed images and videos also.

KEYWORDS-Data mining; Event Summarization; Social Media; Twitter API; Aero India 2017.

I. INTRODUCTION

With the multitude growth of contents and increased usage of the internet there is a need to represent each web document by its digest to save time and effort for searching the precise information. Therefore, a two-fold problem is encountered: searching for relevant documents through an overwhelming number of documents available, and absorbing a large quantity of relevant information. Text Summarization has become a vital and appropriate tool for assisting and interpreting text information in today's fast-growing information age. The goal of automatic text summarization is condensing the source text into a shorter version and preserving its overall meaning. In recent years, need for summarization can be seen in various purpose and in many domain such as news articles summary, email summary, short message of news on mobile, and information summary for businessman, government officials, researchers online search through search engine to receive the summary of relevant pages found, medical field for tracking patient's medical history for future treatment etc [1]. Text summarization on any event refers to the summarization of any concept of interest that gains the attention of the populace. Examples of real-world events range from global catastrophes such as earthquakes, political protests, launch of new consumer products, social occasion etc. The easiest way to extract data related to an event for summarizing purpose is known as Event summarization. It is based on social networks and it is proposed to get a summarized output of events based on temporal features from Twitter stream [2]. Social media services such as Twitter generate phenomenal volume of content for most real-world events on a daily basis. Twitter has increasingly become a critical source of information. People report the events they are experiencing or publish comments on a wide variety of events happening around the world, ranging from the unexpected natural disasters, regional riots, to many scheduled events, such as sports games, political debates, local festivals, and even academic conferences. The Twitter data streams thus cover a broad range of events and broadcast this information in a live manner [3]. Summarizing the collected tweets through the Twitter API can also be implemented using data mining classifying methods. This classifying technique uses natural language processing, text analysis or text mining and machine learning techniques to identify and extract subjective information from the source data set.

Natural Language Processing (NLP) is a branch of computer science that deals with analyzing, understanding and generating the languages that humans use naturally in order to interface with computers. The base-level of NLP lie in a number of sectors like computer science, artificial intelligence, and computational linguistics etc. Natural language processing systems take strings of words (sentences) as their input and produce structured representations capturing the meaning of those strings as their output. The Significance of NLP includes various provenance of thesis, such as

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.5, Special Issue 2, April 2017

An International Conference on Recent Trends in IT Innovations - Tec'afe 2017

Organized by

Dept. of Computer Science, Garden City University, Bangalore-560049, India

Machine Translation, Natural Language text processing and summarization, Information Extraction, Speech recognition, Sentiment Analysis, Artificial intelligence, Expert systems and so on. Text Mining is a domain in data mining. Text mining is used to extract interesting information or knowledge or pattern from the unstructured texts that are from different sources. There are many techniques used in text mining such as information extraction, information retrieval, query processing and clustering.

This paper is organized as follows. In Section II, we review the prior works on Twitter text analysis and summarization. In Section III, we describe the data used for constructing the text summarization for the event Aero India 2017. The details of our Twitter text summarization methodology is presented in Section IV. We describe our experimental results in Section V. Finally, we conclude our work and illustrate potential directions for future work in Section VI.

II. RELATED WORK

Text Summarization methods are divided into two types: Extractive and Abstractive summarization. Abstractive summarization understands of document, finding the new concepts and providing summary in few words. In Extractive summarization method select important sentences, paragraphs etc. from the original text document and concatenating them into shorter form. Anil Kumar et al [4] have presented the Automatic Text Summarization using Extractive techniques with the help of Genetic Algorithm (GA). They have worked on Single-document summarization using extraction method. Genetic Algorithms are a way of solving problems by mimicking the same processes. Therefore, GA is used to specify the weight of each text feature. These approaches are applied on a sample of some English articles and have been used in the feature extraction criteria. K Sathiyamurthy et al [5] proposed an approach for summarizing an event detection based on social networks and semantic query expansion. Events are categorized by a group of similar keywords, documents etc. They have approached a relationship between query expansion and semantic web to increase or improve the precision of event detection. They have given a solution based on hidden state representation of an event using Hidden Markov model. Sarda A.T and Kulkarni A.R [6] have proposed an approach for summarizing articles using neural networks and rhetorical structure theory. A neural network is trained to learn the relevant characteristics of sentences by using back propagation technique to train the neural network which will be used in the summary of the article. After training neural network is then modified to feature fusion and pruning the relevant characteristics apparent in summary sentences. The modified neural network is used to summarize articles and combining it with the rhetorical structure theory to form final summary of an article. Rhetorical Structure Theory provides a combination of features that useful in several kinds of discourse studies. Vishal Gupta and Gurpreet Singh Lehal [7] have performed a survey on extractive techniques of text summarization. To examine and interpret the text linguistic methods are applied. It mainly focuses on extraction of relevant sentences from the document and achieving low redundancy summary. They have summarized the text using regression for estimating feature weights. The experimental result is characterized into intrinsic or extrinsic techniques.

AartiPatil et al [8] investigated on sentence extraction based single Document summarization. They have described that the last six decades, the problem of text summarization has been approached from many different perspectives, in various domains and using various paradigms. With the help of graph based algorithm they have calculated the importance of each sentence in document to generate summary. This gives an indexing weight to the document terms to compute the similarity values between sentences. Jayabharathy et al [9] have presented an analytical framework for Multi-Document Summarization to analyze the performance of the existing techniques. It is mainly based on sentence-level semantic analysis and non-negative matrix factorization. The sentence similarity is calculated by using the semantic analysis and the similarity matrix is constructed. Then the symmetric matrix factorization process is used to group the similar documents into clusters.

Sudhanshu Shiwarkar et al [10] have performed an Automatic Twitter Summarization for novel speech act guided summarization. Speech act recognition technique and its corresponding data sets are used for extraction of keywords and phrases by using round robin algorithm in order to generate template based summaries. They have automatically recognized the speech tweets in a multi-class classification problem based on word-based and symbol-based features from free resources. Ganesh Mane and Anita Kulkarni [11] have performed twitter event summarization using phrase reinforcement algorithm and NLP features to summarize the event specified by the user. They have produced the summary with the help of PRA. The idea behind creating the desired summary is to parse the “raw” summary and build dependencies between the dependent and governor words in each summary. They performed parts of speech tagging and obtain lists of governing and dependent words. The NLP parser was used to build the lists of governor and dependent words. They showed the PR Algorithm can be improved by taking governor-dependency relationships among the constituents. Ruifang He et al. [12] provided a methodology to model temporal context globally and locally,

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.5, Special Issue 2, April 2017

An International Conference on Recent Trends in IT Innovations - Tec'afe 2017

Organized by

Dept. of Computer Science, Garden City University, Bangalore-560049, India

and proposed a novel unsupervised summarization framework with social-temporal context for Twitter data. To assess the proposed framework, they manually labeled the real-world Twitter dataset. The experimental results from the dataset demonstrated the importance of social-temporal context in Twitter summarization.

III. DATA PREPARATION

Event summarization is a procedure of summarizing a data interconnected to an event based on sequential features from Twitter stream which generate the output in the form of summarized report. Summarization techniques can be used to create overviews that capture key facts related to events. Twitter turn out to be the most trendy platform for continuous real-time deliberations. This leads to enormous amount of information related to the real-world, Event detection on Twitter has gained consideration as one of the most popular domains of interest. People use acronyms, post with spelling mistakes, use emoticons and other characters that express special meanings. The challenge of micro blogging is the wide range of topic that is covered. Extracting summarized report from a piece of text such as a tweet, a review or an article can provide us with valuable insight about the author's emotions and perspective: whether the text is subjective (meaning it's reflecting the author's opinion) or objective (meaning it's expressing a fact). For our research we acquired real time data stream using Twitter hash tags (e.g., #AeroIndia2017,#airshow, #bengaluru, #smartfighter, #aerospace) [13]. We collected 100 manually annotated Twitter data (tweets) using Rapidminer. A twitter connection is established through Rapid miner to access the tweets. Tweets of English language and most recent and popular are fetched. An example set consisting of 100 record set from the Twitter API which comprises the tweet text, the tweet ID, the number of re-tweets, the date of creation, the language, the geo-location, the used source of the tweet, and user information are downloaded and consolidated as a text document. Out of 11 attributes from the dataset, Tweet text is taken for Event Summarization.

IV. METHODOLOGY

Text Analysis has the capability to analyze large quantities of unstructured text and detects lexical and linguistic usage patterns to extract meaningful and useful information. It involves information retrieval, information extraction, and data mining techniques, visualization and predictive analytics. An event is an arbitrary classification of a space or time region. An event [14] might have actively participating agents, passive factors, products, and a location in space or time. Events have several properties: i) They are of large scale (many users experience the event), ii) they particularly influence people's daily life (for that reason, they are induced to tweet about it), and iii) they have both spatial and temporal regions (so that real-time location estimation would be possible). Such events include social events such as large parties, sports events, exhibitions, accidents, and political campaigns. Event summarization framework is performed to extract relevant representative tweets from a time-ordered sample of tweets to generate a coherent and concise summary of an event. The basic task in text summarization is performed in order to summarize the events i.e summarizing the tweets related to an event. The methodology followed in Text Summarization of Event is given in Fig. 1.

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.5, Special Issue 2, April 2017

An International Conference on Recent Trends in IT Innovations - Tec'afe 2017

Organized by

Dept. of Computer Science, Garden City University, Bangalore-560049, India

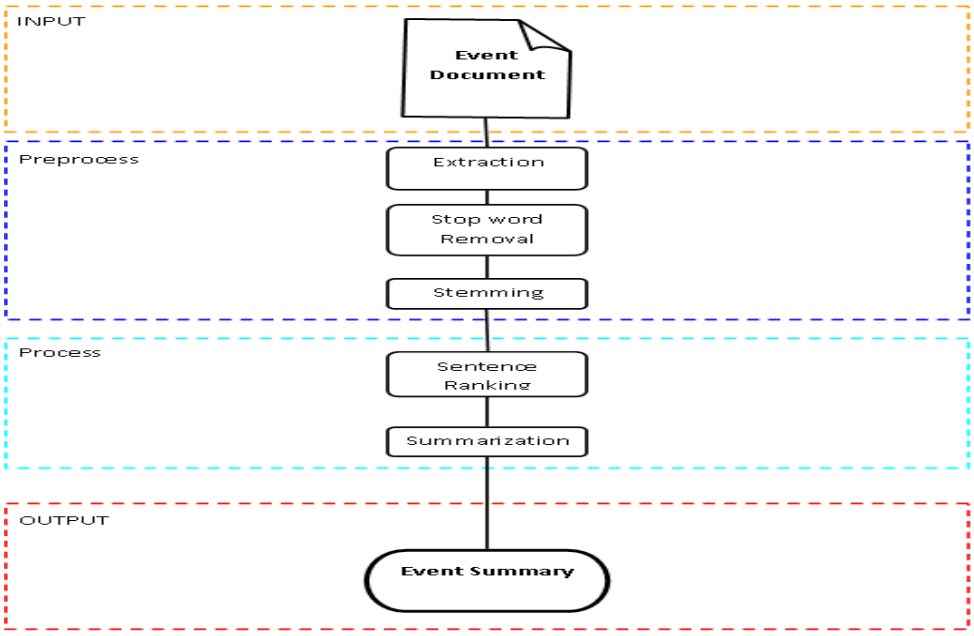

Fig-1 Event Summarization

Preparing the data for summarization is the crucial step [15] in Text Summarization. The retrieved data undergoes preprocessing that includes Text Extraction, Stop Word Removal, Stemming, Sentence Ranking, Sentence extraction and Summarization. The Extraction process tokenizes the file content into individual word.Stop Words are frequently occurring words with no semantics and aggregate relevant information. The most common examples of stop words in text documents are the, in, a, an, with, etc. These stop words can be removed as they are irrelevant to identify most essential sentences. A filter list is created for those words and that are removed from the text vector. Stemming method is performed in order to obtain the radix of each word. For example, the words autograph, graphic, epigraph, demographic all can be stemmed to the word “graph”. This method saves time and memory space. After removing Stop words and Stemming, Sentence Ranking process is performed. Sentence Ranking is a Text Summarization method used for ranking the top most sentences of the document. The outcome of the Sentence ranking is used to extract the top ranked sentences and remove the similar sentences to give the final summary.

V. EXPERIMENTAL RESULTS

The stated methodology is implemented using Rapid miner tool. In our paper we used AYLIEN text analysis extension to summarize from the text. The 100 tweets are filtered according to the location and given to the summarize operator. The connection for the Twitter API and Aylien API should be established to perform the analysis. The Aylien Summarize operator considers each tweet as a single document. Thus the 100 tweets are converted to single document. Then it is summarized to a glimpse of the event. The summarized output includes the frequent tweet text, links of frequently visited images and videos.

The Fig. 2 shows the input dataset which are fetched from Twitter API for the event Aero India 2017, Asia's Premier Air Show in Bengaluru. Fig. 3 shows the Summarized output (important links of the images and videos) of the event Aero India 2017 and Fig. 4 shows the Top most ranked sentences during the summarization of Asia's Premier Air Show event.

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.5, Special Issue 2, April 2017

An International Conference on Recent Trends in IT Innovations - Tec'afe 2017

Organized by

Dept. of Computer Science, Garden City University, Bangalore-560049, India

Row No.	Text
1	RT @LockheedMartin: Discover advanced weapons capabilities & other upgrades for #F16 block 70: https://t.co/3IFrSFHGvY #AeroIndia2017 https…
2	RT @VishalJolapara: My latest □ from #AeroIndia2017
3	RT @VishalJolapara: My latest □ from #AeroIndia2017
4	RT @LockheedMartin: Discover advanced weapons capabilities & other upgrades for #F16 block 70: https://t.co/3IFrSFHGvY #AeroIndia2017 https…
5	RT @VishalJolapara: My latest □ from #AeroIndia2017
6	RT @LockheedMartin: Discover advanced weapons capabilities & other upgrades for #F16 block 70: https://t.co/3IFrSFHGvY #AeroIndia2017 https…
7	My latest □ from #AeroIndia2017
8	RT @vkthakur: Why doesn't anyone remember me, asks the Kargil hero! Mirage 2000 at #AeroIndia2017 https://t.co/WWOkLk3rFr
9	RT @SpokespersonMoD: To touch the sky with glory.....someday!
10	RT @LockheedMartin: Discover advanced weapons capabilities & other upgrades for #F16 block 70: https://t.co/3IFrSFHGvY #AeroIndia2017 https…
11	RT @SpokespersonMoD: To touch the sky with glory.....someday!
12	RT @vkthakur: Why doesn't anyone remember me, asks the Kargil hero! Mirage 2000 at #AeroIndia2017 https://t.co/WWOkLk3rFr
13	RT @LockheedMartin: Discover the #F16 Block 70 for India, the newest generation of Fighting Falcon. https://t.co/bAkpFM8Su1 #AeroIndia2017…
14	RT @SpokespersonMoD: To touch the sky with glory.....someday!
15	RT @SpokespersonMoD: To touch the sky with glory.....someday!
16	RT @SpokespersonMoD: To touch the sky with glory.....someday!
17	RT @LockheedMartin: Discover advanced weapons capabilities & other upgrades for #F16 block 70: https://t.co/3IFrSFHGvY #AeroIndia2017 https…

Fig.2 Input dataset of the Event Asia's Premier Air Show 2017.

ICYMI: Check out our recap video of the incredible technology, cooperation & air power on display at https://t.co/tkSF2nS2W5 Inaugurated

https://t.co/CyzsswFJyF" RT @LockheedMartin: Discover advanced weapons capabilities & other upgrades for #F16 block 70: https://t.co/3IFrSFHGvY https… /: Entretien avec l'ambassadeur de France @FranceinIndia.

https://t.co/oa0kWhmgPX RT @LockheedMartin: Discover the #F16 Block 70 for India, the newest generation of Fighting Falcon.

https://t.co/eWNgIdUcoO Tejas glass cockpit display at https://t.co/xKPHSUoODT RT @akananth: Air Cmde Tejbir Singh, AOC, AFS Yelahanka, real unsung hero

Fig.3 Summarized output of the Event Asia's Premier Air Show 2017.

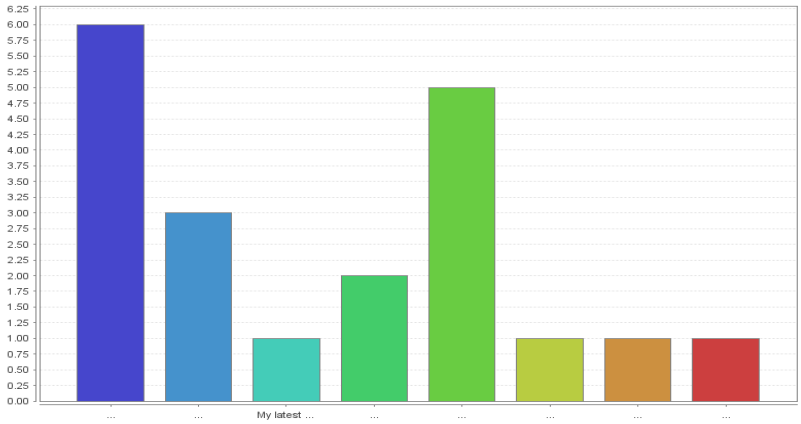

Fig.4 Top most Ranked Sentences of the Event Asia's Premier Air Show 2017.

VI. CONCLUSION

In this paper, an approach to summarize an event effectively using social media data is presented. The method is applied on real-time data stream available in the form of tweets and post. It effectively filters unwanted data and

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.5, Special Issue 2, April 2017

An International Conference on Recent Trends in IT Innovations - Tec'afe 2017

Organized by

Dept. of Computer Science, Garden City University, Bangalore-560049, India

gives summarized results. The efficiency of the system can be improved by including sentiment analysis on the event tweets. The user tweets are likely to change with respect to timeline. The analysis we performed during the days of Airshow. The post event analysis with respect to timeline can also be performed in future to extract the opinions of the participants which could be used as feedback of the event.

REFERENCES

[1] Rafeeq Al-Hashemi, "Text Summarization Extraction System (TSES) Using Extracted Keywords", International Arab Journal of e-Technology, Vol. 1, No. 4, June 2010.
[2] Deepayan Chakrabarti, and Kunal Punera, " Event Summarization using Tweets", Yahoo! Research701 1st Avenue Sunnyvale, CA 94089, 2011.
[3] Alexander Pak, Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", Proceedings of the International Conference on Language Resources and Evaluation,1320-1326, LREC 2010, 17-23 May 2010.
[4] Anil Kumar, Jyoti Yadav and Seema Rani, " Automatic Text Summarization Using Regression Model (GA) ", International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE), Vol. 3, Issue 5, May 2015.
[5] K.Sathiyamurthy, G.Shanmugavalli and Udayalakshmi, " Event Detection And Summarization Based On Social Networks And Semantic Query Expansion", International Journal on Natural Language Computing (IJNLC),Vol. 3, No.5/6, December 2014.
[6] Sarda A.T and Kulkarni A.R, "Text Summarization using Neural Networks and Rhetorical Structure Theory", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 6, June 2015.
[7] Vishal Gupta and Gurpreet Singh Lehal, " A Survey of Text Summarization Extractive Techniques", Journal Of Emerging Technologies In Web Intelligence, Vol. 2, No. 3, August 2010.
[8] Aarti Patil, Komal Pharande, Dipali Nale and Roshani Agrawal, "Automatic Text Summarization", International Journal of Computer Applications (0975 – 8887), Volume 109 – No. 17, January 2015.
[9] Jayabharathy, Kanmani and Buvana, "An Analytical Framework for Multi-Document Summarization", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011.
[10] Sudhanshu Shiwarkar, Indraneel Deshmukh, Nikhil Patil and Akash Thanke, "Automatic Twitter Summarization", International Journal of Advanced Research inComputer Science and Software Engineering, Volume 5, Issue 10, October-2015.
[11] Ganesh Mane and Anita Kulkarni, " Twitter Event Summarization Using Phrase Reinforcement Algorithm and NLP Features", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 5, May 2015.
[12] Ruifang He, Yang Liu, Guangchuan Yu, Jiliang Tang, Qinghua Hu andJianwu Dang, "Twitter summarization with social-temporal context", World Wide Web (2017) 20: 267 - 290.
[13] Deepali K. Gaikwad and C. Namrata Mahender, "A Review Paper on Text Summarization", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 3, March 2016.
[14] Freddy Chong Tat Chua and Sitaram Asur, "Automatic Summarization of Events From Social Media", Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, 2012.
[15] Namita Mittal, Basant Agarwal, Himanshu Mantri, Rahul Kumar Goyal and Manoj Kumar Jain, "Extractive Text Summarization", International Journal of Current Engineering and Technology, Vol.4, No.2 (April 2014).

BIOGRAPHY

Muruganantham A is an Associate Professor, Department of Computer Science (PG), Kristu Jayanti College, Bangalore. He is pursuing Ph.D. in Computer Science from M.S. University, Tirunelveli, TamilNadu. He has 19 years of teaching experience; his research interest is web mining and related areas.

Banu Shree M is an IV Semester MCA Student, Department of Computer Science (PG), Kristu Jayanti College, Bangalore. She interested in doing research on data mining and related area.

Geetha N is an IV Semester MCA Student, Department of Computer Science (PG), Kristu Jayanti College, Bangalore. She interested in doing research on data mining and related area.