International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 10 | Oct-2016 www.irjet.net p-ISSN: 2395-0072

Traffic detection from user's status update messages in twitter

Sruthi sathyanandan [1, 2] Dhanya Sreedharan

[1]*Sree Buddha college of engineering, Alappuzha, India,*

[2]*Sree Buddha college of engineering, Alappuzha, India*

Abstract - Nowadays social networking sites play a major role in providing information to public. This paper discusses the social networking site twitter. Twitter as a social networking site can be used as a source of information for event detection with reference to traffic congestion and road accidents. In this paper we consider a real time monitoring system for the detection of traffic in metropolitan cities. For this approach, the system fetches information from the status update messages of twitter. The system thus fetches each tweets from the update messages. Then the tweets are processed in a way to extract useful information from the tweet. Each tweets are subjected to various data mining process and thus the useful information is mined from those data.

Keywords:-Part of speech tagging, Natural language processing, Status update messages.

1. Introduction

Social networking sites are used as a source of information in event detection for traffic congestion and road accidents. Today social networking sites like Facebook, twitter have gained a high popularity through their real time information broadcasting channel. People mainly use social networking sites to report (public or personal) real life events around them to gain a public attention and also to express their genuine opinion in a particular topic. These messages are shared public among all the users in a social networking sites. Thus each updated messages in these sites gain a high popularity.

Our paper discusses twitter as a social networking site to provide a real time information to public. Each update messages shared in social networking sites are called as status update messages (SUM's).Each such Status update messages may contain meta-informations.The main information include geographical coordinates, timestamp, user's information, links to various resources, mentions and hashtags. In general a collection of status update messages in twitter provides a great deal of useful information. But these messages has to be processed very carefully. In general we can consider users in social networking sites as social sensors and status update messages got from twitter can be considered as sensor information.

Today, social networking sites are used for detection of various events such as traffic congestion, natural disasters. An event can be termed as a an occurrence in a specified space and time.Anyways event detection from the social networking sites is a hard job because often the SUM's will be in the form of unstructured and irregular texts. So we can to be very careful while processing the data. Often the SUM's may contain grammatical errors and sometimes it may contain meaningless information.

So, in order to extract useful information from these texts we use different methods like machine learning, statistics and NLP for the extraction of useful information. The main difficulty arising during the mining process is the vagueness of natural language. If the data is correctly analysed during the mining process we can get a great deal of useful information on a particular topic or an event.

The text mining process can be explained as following. Firstly the document containing the data to be mined is converted into a structured form. During the mining process lot of other operations can be performed. Some of the commonly used operations include indexing and statistical techniques, linguistics analysis by using the NLP techniques, textual feature extraction, part of speech tagging. Among these available methods feature extraction is a major task as well as it is considered as the most important task among the whole mining process. Then the machine learning algorithms,

decision trees, neural networks are applied to the text in the structured form. Then we build the classification and regression models. Then we can modify the results obtained by modifying certain parameters and by repeating the whole process.

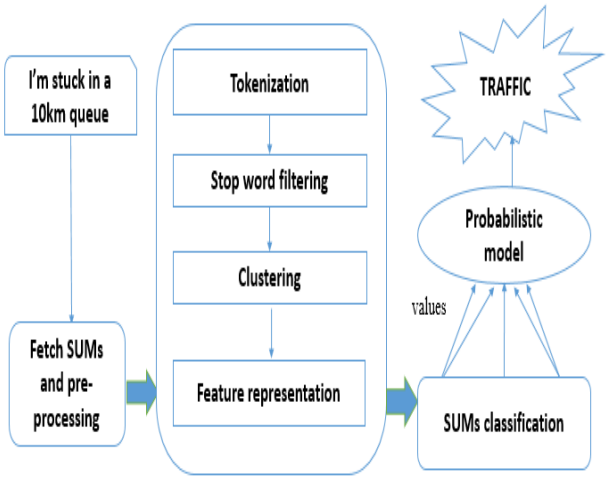

Fig 1.Architecture for detecting traffic from SUMs

The architecture for traffic detection mainly consists of four modules. They are i) Fetch of SUMs, ii)Expansion of SUMs,iii)Classification of SUMs and iv)Probabilistic model. The system fetches SUMs from twitter to process each tweet by applying the suitable text mining process and finally assigns the most suitable class label to each of the SUMs. Most common tools used for developing the system are twitter's API,ii)Twitter4 and iiii)Java API.The main purpose of twitter's API is to provide direct access to the stream of tweets.Twitter4j is a java library that is used as a wrapper for twitter's API.The java API is provided by the weka(Waikato Environment for knowledge analysis).Java API is mainly used in data pre-processing and for text mining elaboration.

I.Fetch of SUMs and preprocessing:

This module extracts raw tweets from the twitter based on certain search criteria. For example it may consider the keywords in the text, geographical coordinates. Each fetched tweet contains timestamp, user id, geographical coordinates, and text of tweet. It may contain additional as well as important information like links, hashtags, mentions and some special characters. After the tweets are fetched then the SUMs are pre-processed in order to gain the most useful information and all the meta-information are removed. Finally a case folding operation is applied to each texts thus obtained in order to convert all the characters into lowercase. Finally we get the output as a set of strings.

II) Elaborating SUMs:

Here a set of pre-processed SUMs are transformed. Here various methods are performed. In the very first itself we perform stop word filtering, tokenization, clustering,, feature representation, SUMs classification are done.

a) Tokenization: It transforms a set of characters into a stream of units called as tokens. Removal of punctuation marks and other non-textual characters are also removed in this step.

b) Stop-word filtering: Here all the stop words that provides very little or no information are eliminated. Example of stop words include words like and, is, or, was etc.

c)Clustering: k-means clustering is used here.It is a method of partitioning data to 'k' subsets and each data element is assigned to closest cluster based on distance of data element from the center of the cluster.

d) Feature extraction: Then corresponding to each SUM its corresponding vector of numeric features are extracted. Then corresponding weights are applied to each relevant stem.

III) Classification of SUMs:

Here each SUM is assigned to class label regarding whether it is related to traffic events.The output of this module forms a collection of N labelled SUMs.

IV) Probabilistic model:

It assigns label to each document based on their ranking and their relevance to a given search query.

2. Literature survey

[1]This paper presents a qualitative investigation to various factors that classifies tweets to useful or not useful. The investigation was carried out based on some search criteria. The study found 16 features that

makes a tweet useful. Mostly tweets that are not useful shows only 2 or 3 of these features. Later these tweets can be assigned various weights based on certain search tasks.

[2]This paper explains about how to distinguish twitter messages about the real life events and non-real life events. Main focus on the messages is done through online identification. The paper identifies each events and its related twitter messages using a clustering technique. The clustering technique typically groups together the similar texts. Then a computation is done based on the revealing features. Each cluster thus helps in determining which cluster is related to each particular events.

[3]This paper explains about a technique that automatically detects events by identifying their text features and temporal components. The events are identified by focusing on three basic components like i) Extraction scheme for representing an event ii)a storage mechanism to store the patterns and iii)hierarchical clustering process to identify the relevant features. The results are a set of relevant events.

[4]This paper makes an investigation of the real-life interaction of events like earthquakes. This paper proposes an algorithm to monitor each tweets and detects a target event. For detecting the relevant events a classifier of tweets is considered based on certain keywords in a tweet. Then a probabilistic model for each target events are created.

[5]This paper presents a segment based event detection system for each tweets. The system firstly detects bursty tweet segments and then clusters the event segments into various events considering the frequency of occurrence of each tweet and content similarity. Then each tweets are divided into non-overlapping segments. Then burst segments are identified within a fixed time based on their frequency patterns.

[6]This paper presents the effectiveness in using a filtered stream of tweets from twitter to automatically identify the events of certain interest from the web. The paper presents a new idea of tagging events with certain keywords from a stream of tweets. The paper concludes that we can identify and tag the most appropriate events by using a filtered stream of tweets.

[7]This paper presents a way of collecting information by environmental and social sensors. Then the data is pre-processed and analysed using various data mining techniques. Then at last optimal routes are found. The paper promotes the flexibility of transport system by providing a number of relevant features. The text mining stages applied to the SUMs are SUM fetch, tokenization, stemming, pos tagging, sentence analysis. The mining stages explained in the paper pre-elaborates and filters the useful information.

[8]This paper explains a method for supporting drivers to drive efficiently by showing the optimum routes along with some driving information's related to traffic jams, weather reports etc. The paper proposes a method of extracting real-time information using the social media. The information gained from the social media is subjected to various text based classification methods. Finally a system is developed to provide information regarding important events for the drivers and then the results are evaluated.

[9]This paper aims in finding real-life occurrences and classifies events based on their type. The paper classified the data mining techniques from different fields according to the type of event, task detection and their method of detection and discusses the most commonly used features. Finally, the system points the need for public benchmarks and evaluates the performance of detection approaches and their features.

[10]The paper presents a way to detect new tweets, analyse tweets based on the spatial and temporal pattern, identifies the importance of various events.The system explained here works for both offline processing and online computing. The paper developed an efficient focused crawler (CDE) to classify and rank each twitter streams and then the location is predicted from the tweet. The system is developed to detect and analyse millions of tweets and users.

[11]The paper presents an online method for detecting real-world events from the twitter. The method adopted combines various textual, frequency components that represent some interesting aspects of an events.The algorithm proposed in this paper uses

different detection components which represents the semantic aspects of the event.

[12]This paper presents an approach based on Hadoop, since Hadoop can be used for processing huge data. This paper implements machine learning concepts also. The paper does real time investigation in three ways. The system analyses the number of tweets associated with various target events. A probabilistic module is used to extract events from tweets to predict location of various events to categorize tweets to positive and negative class. Finally the system shows a message to registered users.

[13]This paper discusses the large scale study of data from multiple social networking sites.Studies shows that there is very high structural difference in data between networking sites and the web.Study reveals that there exists a much higher degree of local clustering in social networks.

[14]This paper discusses a way to detect traffic panels in form of road level images and identifies the information contained in them. Then these images are classified using naïve Bayes classifier. Finally text identification is applied on those images and a probabilistic model is created.

[15]This paper discusses about the challenges of event detection from twitter messages. These challenges are overcomed by using the clustering and classification techniques.

3. Conclusion

This paper discusses about the social networking site twitter. Twitter as a social networking site can be used as a source of information for event detection with reference to traffic congestion and road accidents. Considered as sensor information. Today, social networking sites are used for detection of various events such as traffic congestion, natural disasters. An event can be termed as a an occurrence in a specified space and time.Anyways event detection from the social networking sites is a hard job because often the SUM's will be in the form of unstructured and irregular texts. So we can to be very careful while processing the data. Often the SUM's may contain grammatical errors and sometimes it may contain meaningless information.

So, in order to extract useful information from these texts we use different methods like machine learning, statistics and NLP for the extraction of useful information. The main difficulty arising during the mining process is the vagueness of natural language. If the data is correctly analysed during the mining process we can get a great deal of useful information on a particular topic or an event.

Acknowledgements

We are thankful to our seminar guide Prof Dhanya Sreedharan and PG Coordinator Prof. Minu Lalitha Madhavu for her remarks, suggestions and for providing all the vital facilities like providing the Internet access and important books, which were essential. We are also thankful to all the staff members of the Department of Computer Science & Engineering of Sree Buddha College of Engineering, Alappuzha.

References

[1] J. Hurlock and M. L. Wilson, "Searching twitter: Separating the tweet from the chaff," in Proc. 5th AAAI ICWSM, Barcelona, Spain, 2011.

[2] Beyond Trending Topics: Real-World Event Identification on Twitter Hila Becker Columbia University hila@cs.columbia.edu Moor Naima 2011.

[3] P. Ruche and K. Kamalakar, "ET: Events from tweets," in Proc. 22nd Int. Conf. World Wide Web Brazil, 2013.

[4] Earthquake shakes twitter users: Real time event detection by social sensorsTakeshi Sakami the University of Tokyo Yayoi 2-11 16, Bunkyo-cu Tokyo, Japan sakaki@biz model.t.u- tokyo.ac.jp-2011

[5] Twevent: Segment-based Event Detection Tweets Chankiang Li, Auxin Sun, Anwitaman Datta School of Computer Engineering, Nanyang University, Singapore-2012

[6] Using Twitter to Detect and Tag Important Events in Live Sports James Lanagan and Alan F. Smeaton Centre For Sensor Web Technologies-2011.

[7] G. Anastasi et al., "Urban and social sensing for sustainable mobility in smart cities," in Proc. IFIP/IEEE Int. Conf. Sustainable Internet ICT Sustainability, Palermo, Italy, 2013, pp.

[8] T. Sakami, Y. Matsuo, T. Yanagihara, N. P. Chandragiri, and K. Nawaz, "Real-time event extraction

for driving information from social sensors, "in Proc. IEEE Int. Conf. CYBER, Bangkok, Thailand, 2012.
[9] F. Atene and W. Cherish, "A survey of techniques for event detection in Twitter, ComputIntel" vol. 31, no. 1, pp. 132–164, 2015.
[10] R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang, "TEDAS: A Twitter based Event detection and analysis system," in Pros IEEE ICDE, 2012, pp. 1273–1276.
[11] Getting There First: Real-Time Detection of Real-World Incidents on Twitter Milos Kirstie University of Konstanz Germany 2012 [12] Real time Tweet analysis for event detection & reporting system for Earthquake Grantham V.Shendge 1, Mangesh R.Pawar 2, Nikhil D.Patil 3, Pratik R.Pawar 4, Prof: Devdatta B.Bagul 5 1 2 3 4 5 B.E., Computer, BVCOERI, Maharashtra, India-2015

[13] A. Miscode, M. Macron, K. P. Gummed, P. Druschel, "Measurement and analysis of online social networks," in Proclus, 2007
[14] Traffic Panels Detection Using Visual Appearance A. Gonzalez, L.M. Bergama, J. Javier Yeses and J. Almaz´an-2011
[15] What's happening: A Survey of Tweets Event Detection Amina Laboratory Lyon, France, 2014

Sruthi Sathyanandan received B.Tech. Degree in Computer Science and Engineering from Kerala University, India. Pursuing M.Tech. Degree in Computer Science and Engineering from Kerala technological university.

Dhanya Sreedharan received B.Tech. Degree in Computer Science and Engineering from College of engineering Karunagappally (CUSAT), received M.Tech. Degree in Computer and information Technology from MS University, India. Currently, she is Assistant Professor in Dept of Computer Science and engineering, Sree Buddha College of Engineering, Kerala University, India