International Journal of Engineering &Technology, 7 (3.12) (2018) 205-208

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Identification of Trending Topics Using Periodically Collected Twitter Data

Ranjini V[1], Soundarya R[2], S Karthika[3], S Mohanavalli[4], Srividhya[5]

[1,2,3,4,5]Department of Information Technology, SSN College of Engineering, Old Mahabalipuram Road, Kalavakkam, Tamil Nadu
**Corresponding Author Email: [1]ranjiniv14090@it.ssn.edu.in, [2]soundaryar14110@it.ssn.edu.in, [3]skarthika@ssn.edu.in, [4]mohanas@ssn.edu.in, [5]srividhyav@ssn.edu.in*

Abstract

Social media is an interactive personal tool to articulate an individual's cognizance. This project involves one such micro blogging platform, Twitter. Trends can simply be defined as the frequently mentioned topics throughout the stream of user activities. Mining twitter data for identifying trending topics provides an overview of the topics and issues that are currently popular within the online community. Therefore, the most effective and suitable methodology should be implemented to identify the short term high intensity discussion topic. The trigrams or higher order n-grams are used to determine the trending topic. Twitter Streaming API is used to collect data from the Twitter accounts using API keys and the formatted tweets are stored in a non SQL database. Subsequent steps include data cleansing followed by stemming. The processed data is subjected to trend prediction algorithms like DB Scan, Frequent Pattern Mining, Trees(fuzzy/inductive/decision), Soft frequent pattern mining and empirical statistics such as Frequency metric, TF-IDF, Normalized term frequency and Entropy based on the key parameters to identify the most trending event within a period of time. Thus, the trending topics can be detected with a reasonably close approximation to the expected outcome. This can be used in detecting and predicting events for an early warning system (or) prediction tools and also artificially intelligent services like web search system or recognition systems.

Keywords: *Trend detection, TF-IDF, preprocessing, Twitter, gram.*

1. Introduction

Twitter is a micro-blogging and social-networking website, where users are socially connected around the globe. With the advancements in networking, the entire human population is brought together under a single cap. In Twitter people convey their messages in simple and short posts called as tweets that can be 280 characters or less. According to recent study, there are more than 330 million users who contribute nearly 340 million tweets per day [10]. A user can follow any other twitter user just by clicking the 'follow' button. As a follower the person can see all the updates posted by that person. The tweets can focus on a simple topic or can be based on any particular event. Hashtag is used in twitter by the users while tweeting about any real time event making it easy for other users to follow about a specific topic. The users tend to post tweets on certain topics actively compared to some others that are less likely talked. If the number of tweets on a particular event over a period of time is more, it can be said as trending topic at that time. Hence Twitter data can be periodically collected for a marked time span and analysed for detecting the trending event that has happened at that time period. Twitter's streaming data provides quite a huge amount of tweets that can be mined. But a lot of content from Twitter consists of personal and unwanted information that cannot be used for Twitter trend detection which are the noisy ones and other bursty tweets (sudden increase in number of tweets on a topic at specific period). Therefore, the collected raw tweets are subjected to pre-processing where the unimportant content are filtered out.
Different methodologies and algorithms have been discussed and the trending topics have been detected for a particular time period. Though there are several researches on Twitter trend detection this paper deals on expanding the unigram/bigram techniques to determine trigrams or higher order n-grams to deal with all unseen and unknown words' combination and identify the trending topics from periodically collected twitter data using aggregation of state of art algorithms which is less studied about.

2. Related Works

The researches on trend detection mainly focus on number of tweets posted about the emerging topic. The goal of this research is to detect the most trending topics in a given time period using the aggregation of the state of art algorithms to deal with higher order grams. There are many previous studies dealing with trend detection. James Benhardus discussed different unigram and bigram techniques like relative normalized term frequency analysis and inverse document frequency analysis to detect trends in the collected data. The results are then evaluated with the standard metrics and compared with the real world trending events [1]. A simple model has been proposed for finding bursty topics from microblog's that considers both temporal and user's personal information. The unwanted information are filtered out by different techniques to effectively analyse the data. The result is then compared it with the outcome of the standard LDA [3], [9].

Luca Maria Aiello, et al. proposed a topic detection method that uses consistent ranking methods and preprocessing methods including tokenization, stemming and aggregation [7]. Ishikawa, Shota, et al. introduced a novel detection scheme for hot topics on Twitter by ranking the number of topic's tweets. Data from a local geographical region is collected for a specified time period and classifies the tweets according to the keywords found in them. They also finds the semantic fluctuations in the collected data that may prove as a challenge for pre-processing [2], [8]. 'Eventradar', a real time local event detection scheme that detects social events currently happening around the user with DBSCAN algorithm. The scheme focuses on local events detected that were left as unrelated by previous systems that considered only global events in the given time period [4]. The work done on analyzing new topics in Twitter by Rong Lu and Qing Yang monitors keywords and finds trend momentum to predict the trends and discuss the changes in trends using tendency indicators in technique analysis of stocks[5], [6].

3. Framework

This research follows a step by step approach for collecting and analyzing data from twitter's real time streaming data. First, the huge quantity of data is collected using Twitter Streaming API. It contains the original tweets along with other details such as id, source attributes, time of tweet, retweets and users personal details. The tweet itself contains names, urls, Unicode characters, emoticons, punctuations, hashtags and stop words. Hence in the second step all these unnecessary information are removed. The preprocessed and cleansed data is subjected to different methodologies and state of art algorithms for detecting the list of trending topics over a period of time. In each of the methods, a threshold value is set that shows the rapid increase in the frequency of tweets focusing a particular topic. Then the outputs of all the methods are aggregated to produce the final result of trending topics. Fig.1 represents the framework of these steps.

Architecture Diagram

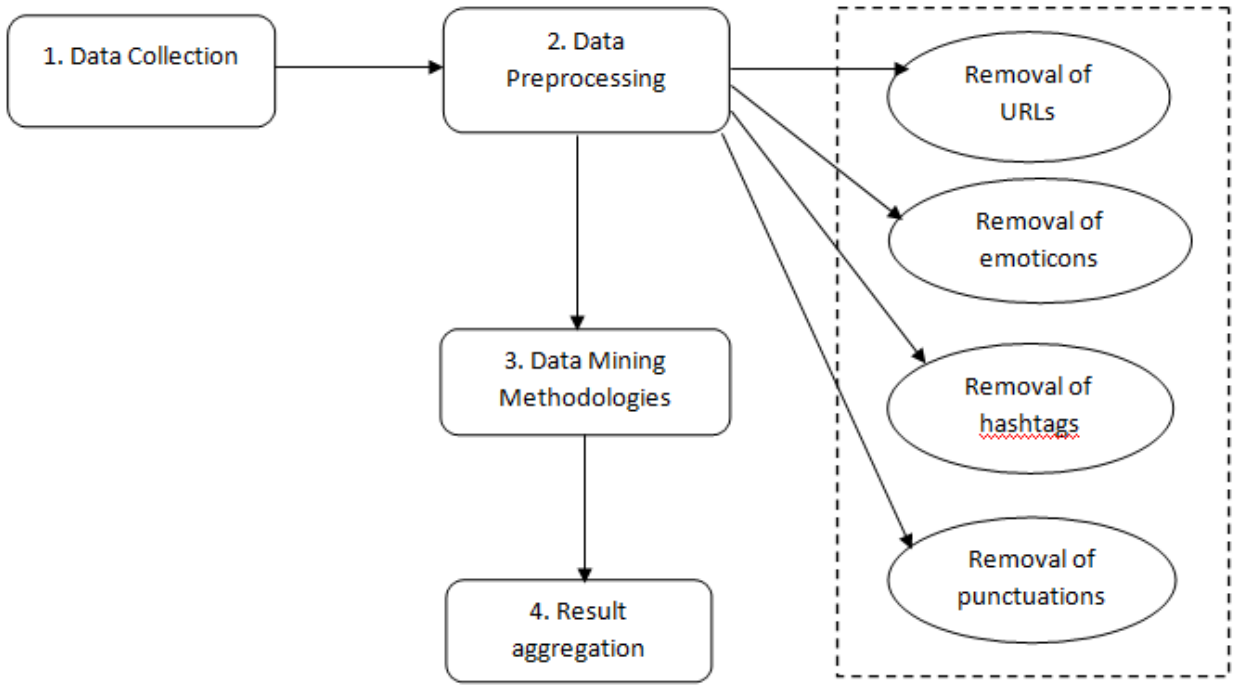

Fig.1: Framework diagram showing the steps involved in the research

3. a Methodology

The multiple methodologies used are discussed below. These methods were implemented using the tweets collected from the Twitter Streaming API. The various statistical methods were chosen to identify the trending topic because there is no one effective algorithm or methodology that could find out the most trending topic. This can be explained with various tradeoffs such as robustness, recall, precision etc. The trend detection is usually based on the key factor which is the product need or the business need. The collected tweets are cleansed through pre-processing steps with rules according to the requirements which is followed by stemming.

The first method used is the raw frequency. This method is a brute force approach but it identifies the trending topic by counting the number of time a particular word occurs in the document of tweets. This method cannot be used autonomously considering its draw backs. Hence it is used as the threshold criterion and a trace back criterion for the other methods. This sometimes will produce higher frequency for the terms that are least required but frequently used like "I" "we" "rt" (retweet) etc., thus when such miscellaneous words occur the stop word list should be updated accordingly to remove such unwanted high frequency results.

The TF-IDF stands for Term Frequency-Inverse Document Frequency. TF-IDF is better than the previous methodology. TF-IDF uses term frequency concept which alone might not yield the best solution because random word's frequency might be higher than the required term's frequency i.e., term frequency is not proportional to relevance. TF-IDF also used a concept called Inverse document frequency which abides by a rule rare vectors are more informative than the frequent ones.

This method is very much suitable for identifying trending topics using a value tf-idf index which is a weighting index. The tf-idf index reflects how important a word is from a collection of documents and also shows its relevance. The documents can either be tweets collected at certain time intervals or every tweet can be considered as a document depending on the implementation requirement. The TF-IDF score is always greater than 0.

Term frequency, $tf_{i,j} = n_{i,j}$ (1)

where $n_{i,j}$ is the number of times word i occurs in document j.

Inverse document frequency, $idf_i = \log\frac{D}{di}$ (2)

where *di* is the number of documents that contain word *i* and *D* is the total number of documents.

$TF\text{-}IDF = tf_{i,j} * idf_i$ (3)

The next numerical statistic used is entropy. The entropy is also an index that has been calculated for all the terms of the document that contains pre-processed tweets. The entropy illustrates the certainty of a vector which is used for the trend identification. The entropy weights are never a negative value as it is a probabilistic approach of identifying the certainty of vector occurrence.

$H_i = -\sum_j \frac{nji}{N} log\left(\frac{nji}{N}\right)$ (4)

where *nj,i* is the number of times word *j* occurs in the collection of tweets containing term i.

$N = -\sum_j n\,j,i$ which represents the total number of words. (5)

3. b Implementation

Data collection

The data collection is the foremost step of the trend detection from twitter data. This is done using the Twitter streaming API, a python wrapper. The twitter allows the users to have access to the public tweets along with its multiple attributes through various twitter API's. The streaming API is one such kind which is python data scrapper that is used here that allows the users to collect data that stream into this micro blogging platform. The API was used with the user credentials and 2,42,700 tweets were collected as a sample for further processing. The data is also grouped into documents of equal duration. Based on the collection time the streaming API retrieves different tweets as the data runs in real-time. The collected data can either be stored in JSON object format or it can be stored in any non SQL database for further parsing using scripts.

Data pre-processing

The data preprocessing or in other words data cleansing is done for the data collected from the Streaming API. In order to identify a trending topic in twitter data cleansing is an important step because the numerical statistics wholly relies on the vectors of the corpus collected in real-time from Twitter using the Streaming API. The cleansed data will thus be free from the frequently occurring unwanted words such as "a" "the" "at" "is" "in" "yes" etc., Also, the concept of using the root word by chopping off the unwanted trail off to the root word. This greatly improves performance. The collection of only English tweets has reduced the pre-processing to a great extent. The pre-processed data will be ready for the implementation of the above mentioned numerical methodologies' aggregation to identify the trending topics.

Pre-processing Rules

As discussed earlier the stop words are removed in the pre-processing stage. A stop word is a vector that has no meaning and it is also not very relevant to the trend detection methods. The following are the rules that are followed during the pre-processing step other than the stemming process.
Rule 1: Removal of URLs
Rule 2: Removal of Unicode characters
Rule 3: Removal of punctuations
Rule 4: Removal of hash tags
Rule 5: Removal of stop words

4. Result

The tweets collected using twitter API were preprocessed using the six above mentioned rules. Table.1 shows the implementation of rules on the raw tweets.

Table 1: Results after applying all the mentioned rules for preprocessing

Sample Tweet	Removal of URLs	Removal of Unicode characters	Removal of Punctuations	Removal of Hash tags	Removal of Stopwords
RT @NoelleFoley: WOMEN\u2019S ROYAL RUMBLE MATCH?!?!?!? OHHHH MYYY GODDDD!!!!!!!!! YES YES YES YES YES YES YES YES YES YES!!!! GOOSEBUMPS!!!!! #\u2026https://t.co/PUJ1RLrNVN	RT @NoelleFoley: WOMEN\u2019S ROYAL RUMBLE MATCH?!?!?!? OHHHH MYYY GODDDD!!!!!!!!! YES YES YES YES YES YES YES YES YES YES!!!! GOOSEBUMPS!!!!! #\u2026	RT :WOMEN ROYAL RUMBLE MATCH?!?!?!? OHHHH MYYY GODDDD!!!!!!!!! YES YES YES YES YES YES YES YES YES YES!!!! GOOSEBUMPS!!!!!#	RT WOMEN ROYAL RUMBLE MATCH OHHHH MYYY GODDDD YES YES YES YES YES YES YES YES YES YES GOOSEBUMPS	RT WOMEN ROYAL RUMBLE MATCH OHHHH MYYY GODDDD YES YES YES YES YES YES YES YES YES YES GOOSEBUMPS	WOMEN ROYAL RUMBLE MATCH GOD GOOSEBUMPS

The last column displays the pre-processed output on which the methodologies like raw frequency, Term Frequency - Inverse Document Frequency and entropy are applied. Table.2 is a comparison between the raw tweets and preprocessed tweets.

Table 2: Comparison between collected raw data and preprocessed data

Sample Data	Preprocessed Data
#new Digital Hair Straightener & Detangling Brush-Ceramic Iron Teeth Heated Straightening Comb-Professional Salon S\u2026 https://t.co/1izK9ceoWq	Digital Hair Straightener amp Detangling BrushCeramic Iron Teeth Heated Straightening CombProfessional Salon
RT @TheHarryNews: #New | Harry with a fan in NYC \n\nJuly 21, 2017 \u2022 \u00a9 @aurpitadeb https://t.co/uC8PrEBYrV	Harry fan NYC July
RT @DanGriffinWLWT: #NEW at 11: Two women from NKY are out tonight trying to make sure people without a home are staying warm and alive in\u2026	Two women NKY out tonight trying make sure people without home staying warm alive

For the sample tweet "*RT @NoelleFoley: WOMEN\u2019S ROYAL RUMBLE MATCH?!?!?!? OHHHH MYYY GODDDD!!!!!!!!! YES YES YES YES YES YES YES YES YES YES!!!! GOOSEBUMPS!!!!! #\u2026*". The preprocessed vectors of the sample tweet is "WOMEN ROYAL RUMBLE MATCH GOD GOOSEBUMPS".

Table 3: Values of the tweet vectors on applying the various methodologies

Tweet Vectors	Raw Frequency	TF-IDF	Entropy
women	2077	0.01238	2.584963
royal	932	0.00413	2.321928
rumble	833	0.00413	2.321928
match	2606	0.00975	2.321928
god	12	0.00125	1.251629
goose bumps	26	0.00125	2.947703

Comparing the entropy score to the TF-IDF score some vectors are more probable than others. The threshold set to the tf-idf score is 0.001 and to that of entropy, the threshold score is above 1.0. If these two scores cross the expected threshold, it tells us that each vector often occurs equally in the distribution. If they are far apart, then the vectors are discarded considering they don't contribute to the trending topic. In the above sample the tweets were collected using *#RAW* (a professional wrestling TV program). Table.3 shows the results on implementing the mentioned methods in the collected and preprocessed corpus.

5. Conclusion and Future Work

The trend identification using the twitter data is done by the aggregation of more than one empirical methodology which has quite obviously produced better results than using a single effective algorithm for trend detection. The future scope includes the machine learning algorithmic aggregation to identify the trending topic for n-grams efficiently.

References

[1] Benhardus, James, and Jugal Kalita. "Streaming trend detection in twitter." International Journal of Web Based Communities 9.1, 122-139, 2013.
[2] Ishikawa Shota, Arakawa Yutaka, Tagashira Shigeaki, Fukuda Akira. "Hot topic detection in local areas using Twitter and Wikipedia." ARCS Workshops (ARCS), 2012. IEEE, 2012.
[3] Diao Qiming, Jiang Jing, Zhu Feida, Lim E-Peng. "Finding bursty topics from microblogs." Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, 2012.
[4] Boettcher, Alexander, and Dongman Lee. "Eventradar: A real-time local event detection scheme using twitter stream." Green Computing and Communications (GreenCom), 2012 IEEE International Conference on. IEEE, 2012.
[5] Lu, Rong, and Yang Qing. "Trend analysis of news topics on twitter." International Journal of Machine Learning and Computing 2.3, 327, 2012.
[6] Roy Soma, Gevry David, and M. Pottenger William. "Methodologies for trend detection in textual data mining." Proceedings of the Textmine. Vol. 2. 2002.
[7] Aiello, Luca Maria, et al. "Sensing trending topics in Twitter." IEEE Transactions on Multimedia 15.6,1268-1282 2013.
[8] Li Rui, Hou Lei Kin, Khadiwala Ravi, Chen-Chuan Chang Kevin. "Tedas: A twitter-based event detection and analysis system." Data engineering (icde), 2012 ieee 28th international conference on. IEEE, 2012.
[9] Ahangama, Sapumal. "Use of Twitter stream data for trend detection of various social media sites in real time." International Conference on Social Computing and Social Media. Springer, Cham, 2014.
[10] https://en.wikipedia.org/wiki/Twitter , Retrieved January 2018.