IOP Conf. Series: Materials Science and Engineering **263** (2017) 042081 doi:10.1088/1757-899X/263/4/042081

Twitter data analysis: temporal and term frequency analysis with real-time event

Garima Yadav, Mansi Joshi and R Sasikala
School of Computer Science and Engineering, VIT University, Vellore – 632014, India

E-mail: sasikala.ra@vit.ac.in

Abstract. .From the past few years, World Wide Web (www) has become a prominent and huge source for user generated content and opinionative data. Among various social media, Twitter gained popularity as it offers a fast and effective way of sharing users' perspective towards various critical and other issues in different domain. As the data is hugely generated on cloud, it has opened doors for the researchers in the field of data science and analysis. There are various domains such as 'Political' domain, 'Entertainment' domain and 'Business' domain. Also there are various APIs that Twitter provides for developers 1) Search API, focus on the old tweets 2) Rest API, focuses on user details and allow to collect the user profile, friends and followers 3) Streaming API, which collects details like tweets, hashtags, geo locations. In our work we are accessing Streaming API in order to fetch real-time tweets for the dynamic happening event. For this we are focusing on 'Entertainment' domain especially 'Sports' as IPL-T20 is currently the trending on-going event. We are collecting these numerous amounts of tweets and storing them in MongoDB database where the tweets are stored in JSON document format. On this document we are performing time-series analysis and term frequency analysis using different techniques such as filtering, information extraction for text-mining that fulfils our objective of finding interesting moments for temporal data in the event and finding the ranking among the players or the teams based on popularity which helps people in understanding key influencers on the social media platform.

1. Introduction

With the increase usage of social networking sites, the real-time events are easily tracked world-wide by any number of users. Before, using printed social media, it was difficult to each out many people and propagates the information instantly. But, with the introduction of social media websites like Facebook, Twitter, sharing of information to the World generated lots of content. Using such platforms, people are becoming more responsive and opinionative about the things happening which lead to the research in field of text mining and data analysis.

In our paper, we are specifically focusing on Twitter micro-blogging platform where each posted sentences are known as 'Tweets' and the important term it includes is 'Hashtag' that holds the meaning of the tweet with just few words. The information diffusion of such tweets is done based on their level of concerns such as global or local. For example, the natural calamities occurring majorly

like tornado, earthquakes are taken as global-level concern; whereas state-level elections may affect only to the national-level which makes it local-level concern. This shows the deployment of information depends upon the cause of the event, its concern and nature and also the target population [1]. There are various domains that people follow like 1) Political domain, where number of people are supporting which party and their views about other parties 2) Entertainment domain, where people rate their movie interest or rate on their favourite sports and 3)Business domain, where people's reviews matter on different products. In this paper, we are focusing on Entertainment domain especially 'Sports' and fetching real-time tweets related to cricket based on hashtag 'IPLT20' to which we are performing time-series analysis and term frequency analysis using different techniques.

The whole implementation process is divided majorly in 3 phases: 1) Collecting data from Twitter 2) Pre-processing of data and 3) Text-mining. Twitter is providing 3 APIs for developers that are: 1) Search API, focus on the old tweets 2) Rest API, focuses on and allow to collect the user profile, friends and followers 3) Streaming API, which collects tweets in a real time as they happen. From which, we are accessing Streaming API in order to fetch real-time event data. This large amount of data is stored to the MongoDB for the completion of first phase. In the second phase, pre-processing of the data is done by filtering some of the relevant fields from the MongoDBjson document. On this filtered data, time-series analysis is performed that shows the peak timings where the users have tweeted more according to the happening event. The third phase is an important one for ranking of the trending player in terms of popularity among users and for that text-mining is performed.

There are 4 techniques to perform Text-mining: 1) Information extraction: It includes tokenization, identifying each entities and segmentation. This technique is performed when one has to need to deal with large number of unstructured data and the application is dynamically driven event. 2) Categorization: It comes under supervised learning and includes pre-processing indexing and dimension reduction and finally classification based on given input-output example. 3) Clustering: It is used for grouping similar kind of documents or contents that provides more meaningful data by creating vector of each topic. 4) Visualization: It creates visual hierarchy that makes the user interface really simple by providing scaling and zooming feature to the document [2]. Text-mining makes it easy to analyse different named entities and relationship among them [3]. Among these 4 techniques, we have used 'Information Extraction' for term frequency analysis that gives us the words count for frequently occurred terms. From this analysis, we can get the most popular player or team for that IPL match.

Our paper is structured in this format. The first section gives an introduction of the paper; the second section is literature survey that shows the amount of work carried out in this filed; the third section contains proposed methodology that includes the detailed description of the techniques used for the implementation; for the performed information diffusion analysis, the fourth section contains final result analysis using graphs and in the fifth section conclusion and future work is mentioned.

2. Literature survey

Nikou et al. [4] have proposed an approach which is used to overcome the challenges appeared when we are using statistical model for detecting the events by using frequency deviation of words frequency with respect to time .They focused on multi-word event as it provides much rich information as compared to single word event .also they taken into consideration the disadvantage when using multi event word is redundancy so to overcome this problem they proposed an approach to find non- redundant event. Further they showed the frequency of tweets with respect to detected event graph for analysis. In our model, frequency analysis is improved as it shows the no of max tweets for real time event.

Koichi et al. [5] proposed a scheme which can detect real time tweets for the events happening in the whole world. They showed the problem faced during real time tweets detection as 1) for quantifying the words accurately and 2) Evaluation of words dynamically. As a solution they proposed two methods 1) the extended hybrid TF-IDF and 2) remarkable word detecting method which accurately detects the events. In our model we are using a text mining techniques which are easy

compare to above proposed method in cost of implementation and it gives the accurate result for the real time tweets also.

Keigo Amma et al. [6] proposed a scheme which classifies the words according to morphological analysis of tweets. They analysed frequently occurring word using quantization method. Also, shows the relational graph and stream graph for events in different domain to convey which topic is trending recently. In our model we focused on time series analysis graph and frequency term analysis graph to show the recent trending topics as well as trending famous personalities.

Liza et al. [7] developed an application which focuses on emotions of twitter user by using six emotions namely sadness, happiness, anger, disgust, fear and surprise. They used the text mining techniques for Pre Processing, processing and validation. In pre-processing they conducted the activities such as cleansing, negation conversion and tokenization, thereafter in processing phase they performed classification using naive bayes algorithm. In our model we used information extraction as a text mining method to extract the relevant information from real time tweets. In the above proposed work they have not considered the real time tweets.

Halima Banu et al. Developed a system to analyse which are the topics trending now-a-days. They used topic based sentiment classification which summarizes the public views over selected trending topics, alongwith that they have generated extractive sub summaries over time using novel sub topic detection approach. There system uses foreground dynamic topic model for finding trending topic by avoiding noisy data. Then they extracted the most significant tweets using graph based approach which uses the salient features of tweets [8]. While in our work, we are showing which topics are trending by time series analysis graph and which player or team is trending or rank of a particular player or team by term frequency analysis graph.

Detecting trends with respect to time series is important in most of area such as market analysis and in system monitoring, some properties of trends are similar in different domain while some are domain specific. Tijl De Bie et al. [9] proposed a probabilistic model that models the time series accurately which shows the results in terms of peaks on top over an exponential graph. They showed the number of tweets addressed per day to a particular twitter user. While in our work we are showing the time series analysis for different topics as well as different users to show which player, artist or team is trending compare to others by showing max no of tweets with respect to time in graph representation.

Maryam Hasan et al. [10] proposed a method to measure the public emotion which is used to predict the important moments during a particular public event. They developed a full stack architecture that performs real time emotion analysis. Also they used a supervised learning approach for classifying tweets. The authors shoed the changes of different emotion classes during different public occasions or events. They have aggregate the emotion class and showed the emotion changing patterns over time. In our proposed work we are showing time series and frequency series graph for different events over real time tweets.

Mariam Adedoyin-Olowe et al. [11] proposed a method transactional –based rule change mining (TRCM) to detect events in particular public event for example they have taken English FA cup finals of year 2012. They used TRCM to extract the Hashtag keywords of tweets during different time slots. They also showed the changing Hashtag keywords over time using association rule mining approach. They showed the graph of statistics of tweets during every goal in the game. They showed the performance based on the game minutes in three sets such as 1) ALL MINUTES, where they haven't performed any filtration 2) PEAK MINUTE, where only the minutes from best performing peak selection method is included. 3) EVENT MINUTE, where only the minutes during which the event takes place.

3. Proposed methodology

Twitter data analysis is done by collecting tweets using streaming API and text mining technique. Basic steps for twitter data analysis is shown in Figure 1.

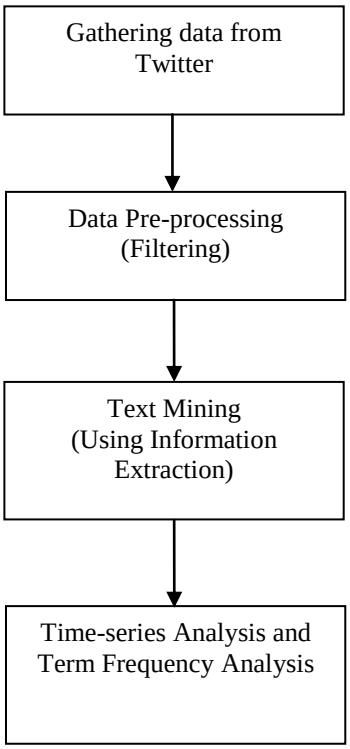

Figure 1. Steps of Twitter Data Analysis

3.1 Gathering twitter data:

In order to collect the tweets from twitter we need create an app that interacts with the twitter API. Basically API stands for application programming interface, many services provides API's to developers to interact with their services. In our proposed work we are using streaming API to collect the tweets. In order to access streaming API we need information about the 4 keys as follows:

1) Consumer key
2) Consumer secret key
3) Access token
4) Access token secret

After accessing the twitter streaming API will connect it to python library using Tweepy and then we will download the real time tweets.

3.2 Data pre-processing:

As we are collecting raw data in the form of tweets which is highly contains noisy, inconsistent and unstructured data. Quality of data affects the results, so in order to improve the quality we apply data pre processing on tweets. Data pre–processing methods are divided mainly in four categories:

1) Data cleaning
2) Data integration
3) Data Transformation
4) Data Reduction

In our work we are using data cleaning for filtering out the redundant and inconsistent data.

3.3 Text mining:
It is the process of extracting the interesting pattern or data from large data set. In our proposed work we have taken into consideration the large no of tweets over which we are applying text mining method. Various methods are there as follows:

1) Information Extraction
2) Categorization
3) Clustering
4) Visualization

We are focusing on image extraction which mainly converts the unstructured data into structured data. As tweets are coming from different sources there is high probability that data is raw and unstructured so we need some mechanism which overcome the above problem, information extraction is one such mechanism. Also, this technique is used for dynamic event detection.

3.4 Time series analysis and term frequency analysis:
Based on the data extracted after text mining we are applying time series analysis to figure out at particular time which event is getting high no of tweets and term frequency analysis to figure out rank of particular player or team which is trending.

4. Result analysis
We have fetched large number of real-time tweets based on IPL-T20 cricket match and performing time-seriese analysis and term frequency anaysis on that using R software and Python. The study shows the need of this anaysis and explains the results generated using graphs performed in python.

4.1 Time-series analysis:
It helps in understanding the interesting events occurred in temporal data. The Figure 2 below shows the graph generated at particular times to the volume of the tweets. High peak of the graph shows that during a particular time an interesting phase (such as the batsman hits the six or gets out) has come and hence the users tweeted more, So that the person can understand and simply can jump to that particular event rather than seeing it in whole afterwards. In our graph, at 17:00 the maximum peak level is seen when the #srh team won against #dd. The streaming data has been taken from Twitter api. The streaming API gives developers to access global stream data. The sample stream data for the IPL-T20 cricket match is given in Tabel – 1.

Table 1. Twitter data of IPL-T20 cricket match

Text	User	created_at
RT @ChrisGayleFdn: .@henrygayle talks about his '10,000 T20 runs' bat going up for auction for the foundation! https://t.co/t4qeZKiZrqâ€¦	colombo58	19-04-2017 14:08
RT @ChrisGayleFdn: .@henrygayle talks about his '10,000 T20 runs' bat going up for auction for the foundation! https://t.co/t4qeZKiZrqâ€¦	Sriniva73760848	19-04-2017 14:10
I have voted on my SRH line-up, have you? https://t.co/z6sKukVveM @IPL #VIVOIPL	Rods_William	19-04-2017 14:12
psl is fierce, and in for it all. https://t.co/VHbiaGD0Z7 #IPL #IPLFANTASY #FANTASYLEAGUE via @ipl	askarizaidi72	19-04-2017 14:18

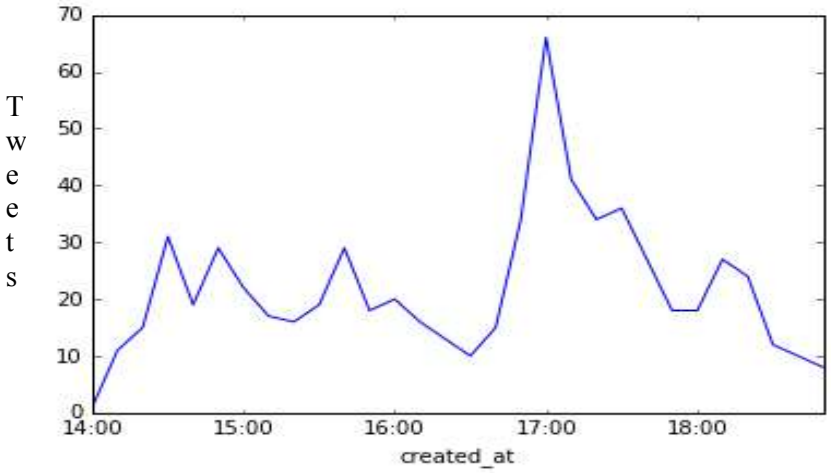

Figure 2. Time Series Analysis

4.2 Term frequency analysis:

It gives number of words count with respect to texts in whole document. From, term frequency analysis we can check the words that are commonly and most frequently used by users and become trendy for that real-time event. From these terms, we can also see the ranking of players by popularity. It is an important factor for the people to understand the key influencers on social media platform while searching. That is why, this analysis is carried out. Our analysis shows the graph of trending 25 terms with respect to count as shown Figure 3. Also we can analyse that for that event players named 'Samson' and 'Iyer' played really well and were trending in terms of popularity.

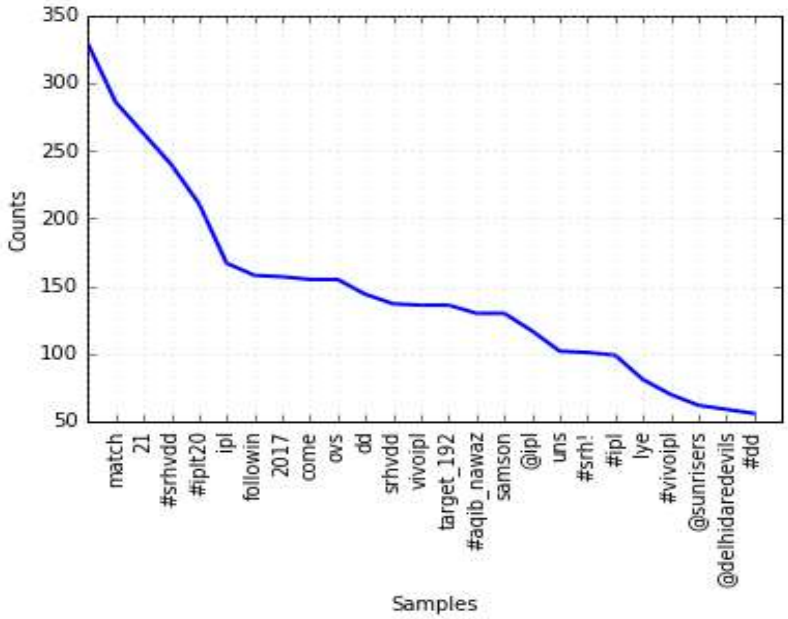

Figure 3. Term Frequency Analysis

5. Conclusion and future work

Data analysis plays vital role with the increase in number of users on each day on social media platform that eventually generates huge amount of data. Twitter data analysis helps in structuring the real-time data and understanding it in a better way using different techniques. This paper discussed about different twitter APIs used for accessing real time tweets, user profile etc. and text mining techniques to extract the relevant data from raw data of tweets, further we performed temporal and term frequency analysis on the data extracted from text mining mechanism. The result of which shows the most happening events of the match and the most frequent as well as popular keywords used by the users. For future work one can apply the proposed method in different domain such as politics and entertainment.

References

[1] Purva Pruthi, Anu Yadav, FarheenAbbasi and Durga Toshniwal, 2015How has Twitter changed the Event Discussion Scenario, A Spatio-Temporal Diffusion Analysis, IEEE International Congress on Big Data, IEEE
[2] Sonali Vijay Gaikwad, ArchanaChaugule and Pramod Patil 2014 Text Mining Methods and Techniques *International Journal of Computer Applications* (0975 – 8887) **85**(17)
[3] Sukanya M, Biruntha S 2012 Techniques on text mining International Conference on Advanced Communication Control and Computing Technologies, IEEE
[4] Nikougunnemann and Jurgenpfeffer 2015 Finding Non-Redundant Multi-Word Events on Twitter *International conference on advances in social networks and mining* IEEE
[5] Koichi Sato, Junbowang, Zixue Cheng 2016 Detecting Real-time Events using Tweets International Conference on Advances in Social Networks Analysis and Mining IEEE
[6] Keigo Amma, Shunsuke Wada, Kanto Nakayama, Yuki Akamatsu, Yuichi Yaguchi, and KeitaroNaruse 2014 Visualization of Spread of Topic Words on Twitter using Stream Graphs and Relational Graphs SCIS&ISIS IEEE
[7] Liza Wikarsa and SherlyNoviathithahir 2015 A Text Mining Application of Emotion Classifications of Twitter's Users Using Naive Bayes Method IEEE
[8] Halima Banu S and S Chitrakala 2016 Trending Topic Analysis Using Novel Sub Topic Detection Model International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics IEEE
[9] Tijl De Bie, JefreyLijffij, C´edricMesnage, Ra´ul Santos-Rodr´ıguez 2016 Detecting Trends in Twitter Time Series International Workshop on Machine Learning for Signal Processing IEEE
[10] Maryam Hasan, ElkeRundensteiner, Xiangnan Kong and Emmanuel Agu 2017 Using Social Sensing to Discover Trends in Public Emotion International Conference on Semantic Computing IEEE
[11] Mariam Adedoyin-Olowe, Mohamed MedhatGaber and Frederic Stahl 2014 Extraction of Unexpected Rules from Twitter Hashtags and its Application to Sport Events International Conference on Machine Learning and Applications IEEE