ISSN: 2321-7782 (Online)
e-ISJN: A4372-3114
Impact Factor: 7.327

Volume 6, Issue 5, May 2018

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Trend Analysis Using Modified K-means Clustering

Kajal Khurana[1]
M.Tech Scholar
Department of Computer Engineering
YMCA UST
Faridabad-121006, India

Payal Gulati[2]
Assistant Professor
Department of Information Technology
and Computer Application
YMCAUST
Faridabad-121006, India

Abstract: ***With the emergence of the Internet, social networks have become the most popular way of communication. The community of users participating in these social networks tends to share their common interests, at the same time giving rise to what are known as social trends. A social trend reflects the voice or opinion of mass users on a certain topic which, for some reason, becomes popular thereafter. Amongst the various popular social networks, Twitter is the fastest growing micro-blogging site and is generating enormous volumes of data every minute. Therefore it becomes difficult to select credible and relevant tweets needed for analyzing the trends. This paper presents an approach for relevant extraction of tweets and thereafter clustering the tweets for efficient analysis of ongoing trends.***

Keywords: Social Network; Twitter; Clustering; K-means; Trends.

I. INTRODUCTION

The magnitude of media content over the social network and micro blogging sites has reached an unprecedented level. Now-a-days millions of users participate in social awareness streams and threads to spread information and their opinions over the network. Twitter [1] is one such very popular online social networking and micro-blogging service, which enables users to share and express their views. The topics discussed on twitter, which gets a lot of attention become a “trend". Using the “#” hashtag, Twitter has made events and trends easier to capture and retweet count; RT determines the popularity of the trends. For example, when Sridevi passed away in February, 2018, #LetHerRestInPeace trends on Twitter, when people appeal social media to end the speculation about the demise of Sridevi[2]. Detecting these trends in online social networks has become an ongoing research area for both the industry and the research community. Therefore this motivated the work. This work proposes an approach to extract, cluster, rank and analyze the emerging topics /tweets that appear in the stream at a particular time instance. The proposed method initially extracts the tweets (# tags, related text, RT count) related to user defined event (input keyword) using Twitter API. Further pre-processing of extracted tweets is done and thereafter modified K-means clustering is proposed to cluster and rank the trends.

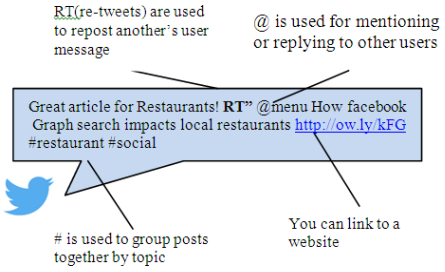

Fig.1 A Sample Tweet

II. Related Work

Ramage et al.[3] characterize the tweets using labelled Latent Dirichlet Allocation(LDA) method that maps the content of the Twitter feed into dimensions. These dimensions correspond roughly to substance, style, status, and social characteristics of posts. He used labeled LDA to discover trends in language usage (what words end up together in topics).

Cataldi et al. [4] try to identify emerging topics in the Twitter network based on term frequency and users authority. The set of emerging topics are found by creating topic graphs which links the emerging terms with their relative co-occurent terms.

Mathioudakis et al. [5] identify the trending topics on Twitter in real time using bursty keywords (keywords that suddenly appear in tweets at an unusually high rate) and their co-occurrences.

In the paper [6], Weng et al. find the influential users in Twitter by taking both topical similarity and link structure between the users. For topic identification they use Latent Dirichlet Allocation (LDA) which uses bag of words concept.

Chen et al. [7] study the problem of recommending tweets using different approach, one of them is based on topic. For topic identification they use TFI-DF technique which also used bag of words.

Qing Chen et al.[8] proposed to use Wikipedia search by identifying the important words in a tweet, which the authors called, "the trend," and then do a Wikipedia search based on the trend name . When a search is done with Wikipedia, it could potentially return hundreds of records. The top five to six search records are then selected, and the short snippet in the search result is added to the tweet, which is then used as an input record.

Swit Phuvipadawat et al.[9] used the hashtag #breakingnews and collected all the tweets with that hashtag . Once the tweets were collected, they used Stanford Named Entity Recognizer to classify the keywords into proper nouns. Once this step was completed, similar tweets were grouped together with a number and a group label so that every message would belong to a particular group.

Karray et al. [10] detects the topic from the set of tweets (size n) and represent each topic using only one tweet. A tweet which is similar to a set of tweets and dissimilar to the rest of the tweets is a good topic representative.

III. Proposed Work

This paper proposed the ranking mechanism to cluster the tweets which is the modification of K-means clustering algorithm.

This section determines

- How the tweet data-set is collected?
- How clustering is performed on tweet text?
- How to analyze the trend?

A. Crawl Manager :-

The crawler is used to extract the data from multiple social networks. From Twitter, an application can extract data by using APIs. Twitter provides various kinds of APIs like Search APIs, REST API, and the Streaming APIs for different purposes. This system used Tweepy, an easy to use, open source, Python library for accessing the Twitter API. Tweepy requires Basic Authentication, so OAuth is the only way to use Twitter API. To access and collect Twitter data, consumer key, consumer secret, access token, and access token secret keys are required. The crawl manager extracts the real-time data based on user-generated keyword for a specific period of time and stored in the dataset.

B. Data Preprocessing :-

Now, the data in the datasets is processed. The first and the foremost step is tokenization which splits the stream of data into smaller units called tokens. Thereafter, abbreviations, emoticons, hyperlinks, and the punctuations are removed. Another step is to remove the stop words as well as transform all letters to lower case. Another simplification carried out on dataset is stemming. It is the process of reducing derived words to their root words. Porter algorithm is used for stemming.

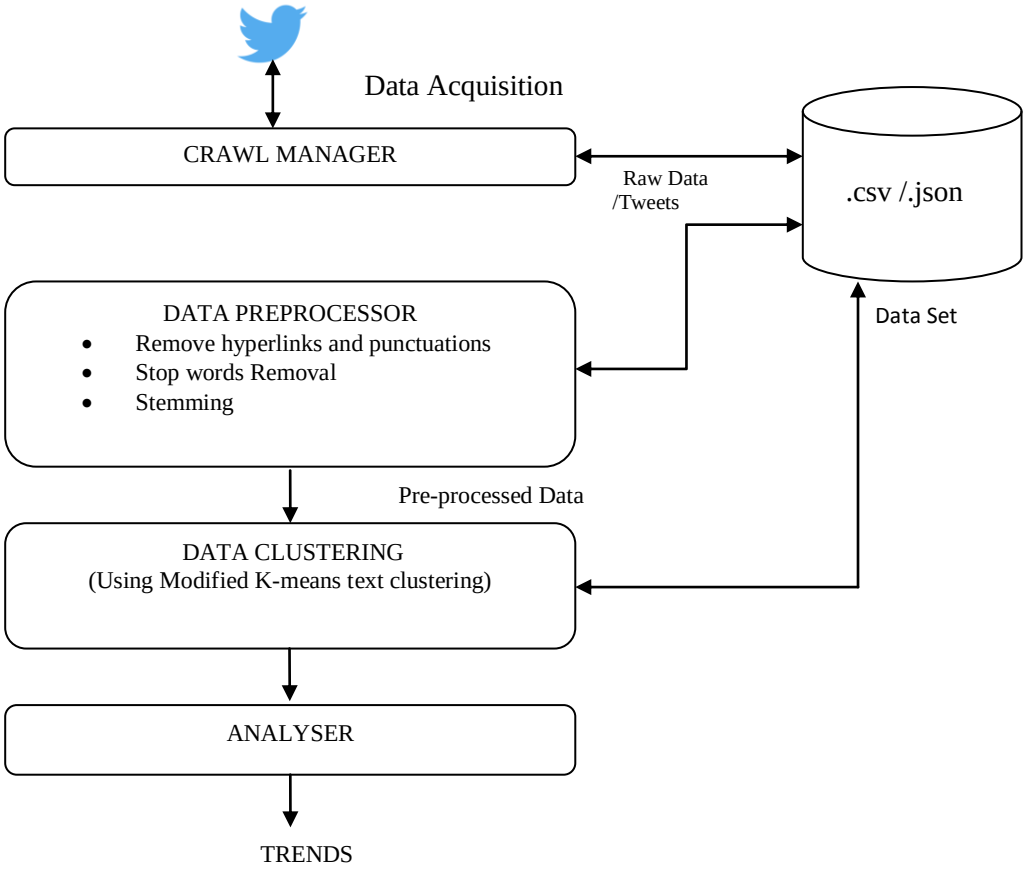

Fig.2 System Architecture

C. Data Clustering:-

In a stream of tweets extracted, many tweets are in fact re-tweets of another tweet. Re-tweets of the same tweet should always belong to the same cluster. To perform this task, this model used modified K-means clustering which is an unsupervised machine learning algorithm for text clustering.

K-means[11] attempt to split the dataset into fixed number of clusters by selecting K sample points as the initial cluster center. This algorithm works well with the numbers. But this paper deals with the clustering of tweet text (text clustering) which is a different task. To get the numbers from text, it performs feature extraction using term frequency-inverse document frequency (TF-IDF). It is a way to convert the textual information into Vector space model (VSM). Before running K-means on a set of tweets, the tweets have to be represented as mutually comparable vectors using the TF-IDF score. It provides a weight to each tweet that describes the importance of a keyword in that tweet.

$$TF = \frac{number\ of\ occurence\ of\ a\ keyword\ in\ a\ tweet}{total\ number\ of\ words\ in\ tweet}$$

$$IDF = \log\left(\frac{number\ of\ tweets\ in\ data\ set}{number\ of\ tweet\ containing\ keyword}\right)$$

The high TF of a word means that the user mentions that word frequently, indicating higher interest, while high IDF of a word means that other users also mention this word, indicating that the word can better distinguish one user from other users. Inverting the document frequency by taking logarithm assigns a higher weight to rarer terms.

$$TF - IDF\ score = TF \times IDF$$

The modified algorithm uses the concept of re-tweet count (attribute of tweet) with the existing concept. The reason is very simple. If the tweet is not relevant, its re-tweet count is less as compared to the relevant tweets. So it also plays an important role in clustering the tweets.

Modified K-means Algorithm:

Let K is the keyword dataset

K = $\{k_1, k_2, k_3, \ldots\ldots\ldots\ldots, k_n\}$

R_i represents no. of re-tweets for keyword k in timeslot i.

t_i represents no. of tweets for keyword k.

Input:- Processed dataset of Twitter

Output:- Cluster set of data

1. Apply K-means clustering algorithm to the dataset using modified TF-IDF score by incorporating re-tweets count, rank of the keyword is found

$$R(ki) = \frac{\text{Ri} - \text{Ri} - 1}{\max(\text{Ri}, \text{Ri} - 1)}$$

$$Score(ki) = ti \times \log(1 + \frac{\text{R(ki} + 1)}{\text{ti} - 1})$$

2. In a timeslot, a keyword is related to number of tweets and re-tweets, normalization of score can be done.

$$NS(NKi) = \frac{\text{Score(ki)} - \min(\text{Score}(< \text{K} >)}{\max\big(\text{Score}(< \text{K} >)\big) - \min(\text{Score}(< \text{K} >))}$$

The above algorithm outputs the cluster set of data. Now, these clusters are ranked using already proposed ranking methods. Ranking the clusters gives us the relevancy score which provides an indication about the trends.

IV. OUTCOME

The graph in Fig. 3 shows the analysis of trend for a particular interval of time.

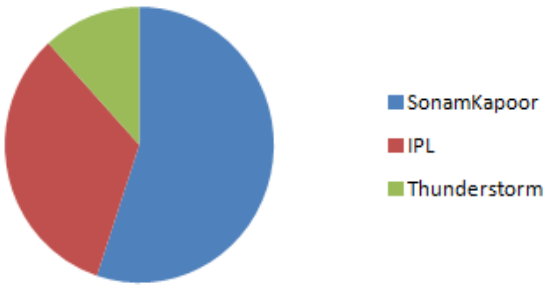

Fig.3 Trend Analysis

V. CONCLUSION

Twitter is generating enormous volumes of data every minute leading to difficulty in selecting credible and relevant tweets needed for analyzing the trends. The aim of this paper is to extract relevant tweets, re tweet count and thereafter apply clustering

on tweets for efficient analysis of ongoing trends. Modified K means clustering algorithm is proposed in this paper which proves to be better and efficient way of clustering tweets. Finally on the basis of relevancy score ongoing trends in a stipulated time instance are analyzed.

VI. Future Work

In future, this algorithm is further modified by simultaneously using many features of tweets shown in Fig. 4 to efficiently cluster the tweets. The proposed system is not efficient for a large stream of data because with time, the number of clusters will keep growing. So another future work will include the extension of an algorithm by introducing cluster management.

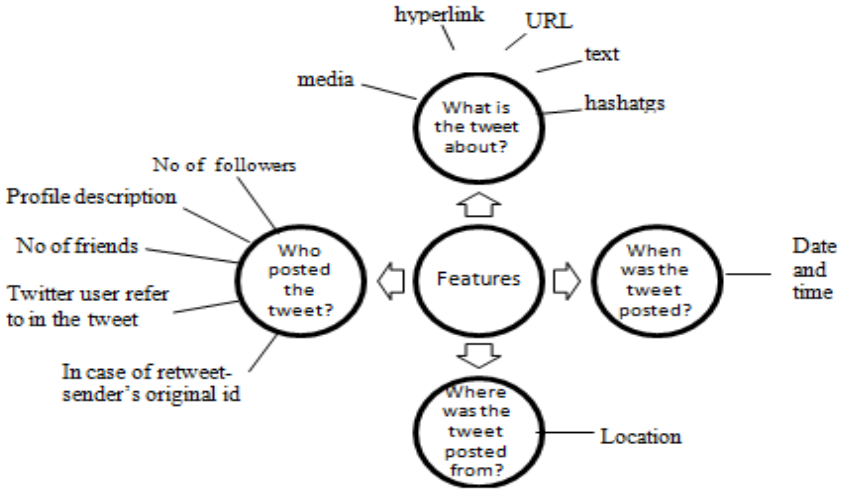

Fig. 4 Features used to categorize tweets

References

1. https://en.wikipedia.org/wiki/Twitter
2. www.twitter.com
3. D. Ramage, S. Dumais, and D. Liebling. Characterizing Microblogs with Topic Models. In ICWSM,2010.
4. M. Cataldi, L. Di Caro, and C. Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In Proceedings of the Tenth International Workshop on Multimedia Data Mining, MDMKDD '10, pages 4:1–4:10, New York, NY, USA, 2010. ACM.
5. M. Mathioudakis and N. Koudas. Twittermonitor: Trend detection over the twitter stream. Proceedings of the 2010 International conference on Management of data, pages 1155–1157, 2010.
6. J. Weng, E. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In Proceedings of th third ACM international conference on Web search and data mining, WSDM '10, pages 261–270, New York, NY, USA, 2010.ACM.
7. J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: experiments on recommending content from information streams. In Proceedings of the 28th international conference on Human factors in computing systems, CHI '10,
8. Qing Chen; Shipper, Timothy; Khan, Latifur; , "Tweets mining using WIKIPEDIA and impurity cluster measurement," Intelligence and Security Informatics (ISI), 2010 IEEE International Conference on , vol., no., pp.141-143, 23-26 May 2010 doi: 10.1109/ISI.2010.5484758
9. Swit Phuvipadawat and Tsuyoshi Murata. 2010. Breaking News Detection and Tracking in Twitter. In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 03 (WI-IAT '10), Vol. 3. IEEE Computer Society, Washington, DC, USA, 120-123. DOI=10.1109/WI-IAT.2010.205
10. Ahmed Elbagoury, Rania Ibrahim,Ahmed K. Farahat, Mohamed S. Kamel, and Fakhri Karray, "Exemplar-Based Topic Detection in Twitter Streams" In Proceedings of the Ninth International AAAI Conference on Web and Social Media,2015
11. Ranjana Agrawal, Madhura Phatak, "Document Clustering Algorithm Using Modified K-Means ", fourth international conference on advances in recent technologies in communication and computing(artcom-2012),pp: 294-296,IEEE, 2012
12. Rupesh Kumar Mishra, Knika Sain, Sakshi Bagri, "Text Document Clustering On The Basis Of Inter Passage Approach By Using K-Means", International Conference On Computing, Communication And Automation,(ICCCA- 2015), may 15-16, pp:110-113,IEEE,2015
13. "Introduction to K-Means Clustering with Twitter Data" http://rstudio-pubs static.s3.amazonaws.com/5983_af66eca6775f4528a72b8e243a6ecf2d.html
14. Sophie Baillargeon, Simon Halle, Christian Gagne, "Stream Clustering of Tweets" International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (2016 IEEE/ACM), 2016.

