ISSN 2348 – 8034
Impact Factor- 3.155

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

TOPIC DETECTION FROM ONLINE SOCIAL MEDIA

Ms. Jasmin Jamadar[*1] and Prof. Rajesh Phursule[2]
[*1]Department of Computer Engineering, Imperial College of Engineering & Research, Wagholi, Pune, India
[2]Department of Computer Engineering, Imperial College of Engineering & Research, Wagholi, Pune, India

ABSTRACT
From last few decades there is wide spread usage of social network platforms such as twitter or other micro blogging systems which contains huge amount of timely generated data. Tweeter is fastest means of information sharing where user shares event/news which take place in front of them. Thus Tweeter act as news portal where news reaches to the people within fraction of seconds. Extracting valuable information in timely manner is important because this wealthy information is useful for companies, government agencies and health organizations. Topic detection is the new research area in data mining and knowledge discovery where extracting useful and valuable information from timely generated online streams is the new challenge. In this article we survey the different algorithms used for trending topic and event detection using social media data and proposes new system for topic detection from social media.

Keywords: *Twitter, Topic Detection, Term Aging, Term Co-occurrences, Burstiness.*

I. INTRODUCTION
Topic detection and topic modelling is the new research area in data mining and information retrieval which has attracted several communities in last few decades. Social media mostly used for online conversation. Some conversations are user specific and large part of conversations are responses which triggered by events e.g. political events (presidential polls), disasters (earth quack, terrorist attack), protest and marches. Twitter, Facebook, Tumblr are the popular microblogging system where people share their experience, exchange opinions in terms of short messages. Among these Twitter is the powerful means of fastest information service in the world. Billions of messages appear daily reacting to the real world event and cultural event. It is popularly known as best news paper as it informs the web community about emerging topics and events. Extracting valuable information in timely manner has gained the popularity because this wealthy information is useful for companies, government agencies and health organizations. Extracting valuable information from huge amount of data is challenging due its informal structure and diversity.
Numerous research paper have concentrated on presenting methods and systems for extracting emerging topics from social media but the proposed approach is not able to distinguish between a news event and irrelevant hot topics. In this sense, we propose a future work in which we distinguish them through the entity and event detection algorithms.

II. RELATED WORK
Mario Cataldi [1] studied the role and topology of Twitter. Authors analysed the user's interest by extracting content of his generated tweets then using social relationship, directed graph of online community having active users is designed. Authors' monitored streams generated by the entire network and studied the behaviour of each term using aging model. Term based aging model is used to compute the burstiness of each term using life status (i.e. Burstiness value) of each term authors ranked terms. Then using keyword graph based approach minimal set of terms is retrieved which connects the extracted terms. Considering the time frame of active user tweets authors analysed the content of tweets, this information is used to represent the topic. The proposed approach is not able to distinguish between a news event and irrelevant hot topics.
Luca Maria Aiello [2] presented a comparative study of six different topic detection methods on three Tweeter data collections authors observed the factors like structure of data, pre-processing techniques of data, noise in the data and user activity over the time, which greatly affect quality of the final result. Authors tested six methods using three different classes namely, probabilistic model (Latent Dirichlet Allocation), Classical Topic detection and Tracking (document- Pivot approach), and feature pivot methods and proposed the new algorithm which combine the n-gram with df-idf topic ranking algorithm which give best result among the other state-of-art techniques. As an extension

to this work authors suggested that this technique would be used to detect the most interesting topic within events and thus able to notify only more related topic/events which occurs.

Marta Arias [3] proposed a method that includes pre-processing of Tweeter data, building sentiment indicator and sentiment index from Tweeter data. Authors built the machine learning models i.e. neural network, linear models, Support vector model, several experiments has been conducted involving models with sentiment index and each of these model is trained with and without Tweeter data and performance comparison of each model is made using these experimental results authors developed the decision tree which represents the information which they termed as summary tree. Using summary tree they proved that forecasting prediction with Tweeter data is accurate as compared to forecasting prediction without Tweeter data. In the proposed work authors found that nonlinear models i.e. SVM and neural network performs better as predictor, whereas linear regression model unable to exploit any sentiment indices.

Hassan Sayyadi [4] suggested KeyGraph method for topic detection and tracking. KeyGraph is directed graph that represents document collection as a keyword concurrences graph. Authors then applied community detection algorithm to form the cluster of keywords into community. Each community forms grouping of keywords that represents the topic. Authors compared the accuracy and execution time of KeyGraph with other topic modelling technique like LDA-GS on spinn3r dataset.

Nargis Parvin [5] described novel method and its implementation for the detection of trending topics. Authors designed novel and context sensitive algorithm 'TrendMiner' for detecting trending topics in microblog post streams. This work has several limitations; first, the current approach does not consider the order of the words in trending topics. Second, when the cluster size increases, the precision of the TrendMiner method decreases.

Xiangmin Zhou [6] proposed a novel framework to detect composite social events over streams, which fully exploits the information of social data over multiple dimensions. First a graphical model called location-time constrained topic (LTT) is implemented to capture the content, time, and location of social messages. Using location-time constrained topic a social message is represented as a probability distribution over a set of topics by inference, and the similarity between two messages is measured by the distance between their distributions. Events are then identified by conducting efficient similarity joins over social media streams. Variable dimensional extendible hash over social streams is utilised to accelerate the similarity join.

Erich Schubert [7] suggested a significance measure that can be used to detect emerging topics early, long before they become \hot tags", by drawing upon experience from outlier detection. Secondly, by using hash tables in a heavy-hitters type algorithm for establishing a noise baseline also showed how to track even all keyword pairs using only a fixed amount of memory. Finally the detected co-trends are aggregated into larger topics using clustering approaches, as often as a single event will cause multiple word combinations to trend at the same time.

Feng Chen [8] has focused on Non-Parametric Heterogeneous Graph Scan (NPHGS), a new approach that considers the entire heterogeneous network for event detection: author first modeled the network as a "sensor" network, in which each node senses its "neighbourhood environment" and reports an empirical p- value measuring its current level of anomalousness for each time interval (e.g., hour or day). Then, they efficiently maximize a nonparametric scan statistic over connected sub graphs to identify the most anomalous network clusters. Finally, the event represented by each cluster is summarized with information such as type of event, geographical locations, time, and participants. The limit of NPHGS is it can't provide rich domain knowledge naturally.

Tim Althoff [9] studied across three major online and social media streams, Twitter, Google, and Wikipedia, covering thousands of trending topics during an observation period of an entire year, results indicate that depending on one's requirements one does not necessarily have to turn to Twitter for information about current events and that some media streams strongly emphasize content of specific categories. Author further presented a novel approach for the challenging task of forecasting the life cycle of trending topics in the very moment they emerge. This fully automated approach is based on a nearest neighbour forecasting technique exploiting author's assumption that semantically similar topics exhibit similar behaviour. The proposed model is unable to explicitly detect and exploit seasonality as well as incorporate global changes in viewing statistics.

III. PROPOSED FRAMEWORK

We initially analyze user interests by extracting and formalizing the content of her generated tweets. We then model the social community as a directed graph of the active authors based on their social relationships, and calculate their authority by relying on the well-known PageRank algorithm. Therefore, we monitor the stream of information expressed by the entire network by studying the lifecycle of each term according to a novel aging model that also leverages the reputation of each author. We therefore select the set of most emerging keywords by dynamically ranking the terms depending on their life status (defined through a burstiness value). We represent each related to

each emergent term by constructing and analyzing a keyword graph which links the extracted emerging terms with all their co-occurrence keywords. At this point, in order to personalize the list of retrieved emerging topics, we analyze the temporal time frames in which the user has been active and analyze her generated content to estimate the user's interests according to this temporal information. This time-aware information is finally used to highlight the topics that best match the interests of the user. Various steps of proposed work are as follows:

3.1 Pre-processing

Dropping common terms: stop words

Sometimes, some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded from the vocabulary entirely. These words are called stop words. We remove unwanted words or stop words from data.

Stemming

Stemming usually refers to a heuristic process that chops off the ends of words, including the removal of derivational affixes. We also apply stemming operations on every word.

3.2 Term aging model

Weight - TF/IDF: Term frequency counts the number of occurrences of term in the document whereas inverse document frequency is high for rare words in document collection and low for frequent terms.

Reputation Calculation: Calculate the reputation for every user using followers and there weighted words.

Term Burstiness Values: The burstiness value of a term indicates its actual contribution (i.e., how much it is emergent) in the corpus of tweets. Our idea is that the temporal information associated to the tweets can be used as distinguish function in that sense.

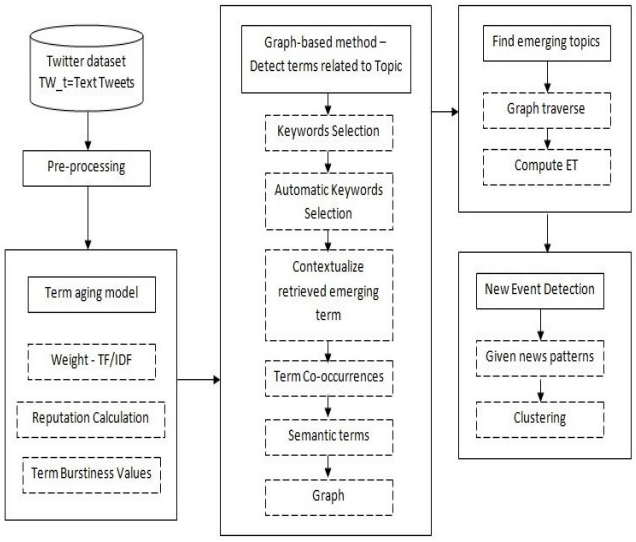

Figure 1: Proposed Framework

3.3 Detect terms related to Topic

Keyword selection: The first approach for the selection of emerging terms relies on a user-specified threshold parameter.

Automatic keyword selection: In order to cope with this, we provide a completely automatic model that works as follows.

1. The system first ranks the keywords in descending order of burstiness value previously calculated in.
2. It computes the maximum drop between consecutive entries and identifies the corresponding drop point.

3. It computes the average drop (between consecutive entities) for all those keywords that are ranked before the identified maximum drop point.

4. The first drop which is higher than the computed average drop is called the critical drop.

Contextualize retrieved emerging term: Considering the given corpus of tweet, tweets are extracted within the time interval, in this step we study the semantic relationships that exist among the keywords in it, in order to retrieve the topics related to each emerging term.

Term Co-occurrences: We get various terms and calculate the co-occurrences for each word.

Graph: We create the graph by using User, word, TF co-occurrences etc.

3.4 Find Emerging Topics

We traverse the graph and calculate emerging trends. Also get the news patterns and find the news events from twitter.

3.5 Clustering

Finally we get all these parameters and create clusters of news events.

IV. CONCLUSION

The related work presented in the paper showed that Twitter is the ideal scenario for the study of real-time information spreading phenomena. All, the proposed approach is not able to distinguish between a news event and irrelevant hot topics (e.g., discussions on facts about celebrities). Considering possible future works on this direction, we aim at how to distinguish them through an event and entity detection technique, taking into account their presentation to the user (separately and when requested). Finally, we would aim at evaluating the existing correlation between the popularity of a user (intended as the capacity to influence other users to post related tweets, measured in some way) and the information spread. In detail, our goal would be to investigate on this aspect and analyze its impact in research approaches for topic detection and tracking techniques on information networks.

REFERENCES

1. *Mario Cataldi, Luigi Di Caro and Claudio Schifanella, "Personalized Emerging Topic Detection Based on a Term Aging Model", ACM Transactions on Intelligent Systems and Technology, Vol. 5, No. 1, Article 7, December 2013.*
2. *Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, "Sensing Trending Topics in Twitter", IEEE Transactions on multimedia, Vol. 15, NO. 6, October 2013.*
3. *Marta Arias, Argimiro Arratia, and Ramon Xuriguera, "Forecasting with Twitter Data", ACM Trans. Intell. Syst.Technol. 5, 1, Article 8 ,December 2013.*
4. *Hassan Sayyadi and Louiqa Raschid, "A Graph Analytical Approach for Topic Detection", ACM Trans. Internet Technol. 13, 2, Article 4 December 2013.*
5. *Pervin, N., Fang, F., Datta, A., Dutta, K., and Vandermeer, "Fast, scalable, and context-sensitive detection of trending topics in microblog post streams", ACM Trans. Manage. Inf. Syst. 3, 4, Article 19, January 2013.*
6. *Xiangmin Zhou, Lei Chen, "Event detection over twitter social media streams" ,© Springer-Verlag Berlin HeidTelberg 2013.*
7. *Erich Schubert, Michael Weiler, Hans-Peter Kriegel," SigniTrend: Scalable Detection of Emerging Topics in Textual Streams by Hashed Significance Thresholds", ACM Transactions on Intelligent Systems, August 2014.*
8. *Feng Chen, Daniel B. Neill," Non-Parametric Scan Statistics for Event Detection and Forecasting in Heterogeneous Social Media Graphs", ACM, August 2014.*
9. *Tim Althoff Damian Borth Jörn Hees Andreas Dengel, "Analysis and Forecasting of Trending Topics in Online Media Streams", ACM MM'13, October 21–25, 2013.*