IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 4, Ver. VI (Jul.-Aug. 2016), PP 84-90
www.iosrjournals.org

Finding New Trends in Public Twitter Streams using Link Anomaly Detection

Rajyalakshmi Golla[1], R Lakshmi Tulasi[2]
#1 Student of M.Tech (CSE) and Department of Computer Science & Engineering,
#2 Prof, Head of the Department in Computer Science & Engineering, Qis Institutute of Technology, Ongole, AP, India.

Abstract: *Social Network is a site where individual's vocation and share data identified with the present occasions everywhere throughout the world. This specific conduct of users made us concentrate on this rationale that handling these substance may lead us to the extraction the present point of enthusiasm between the users. It additionally functions admirably even the substance of the messages are non-printed data. The algorithm demonstrate that the proposed notice peculiarity based methodologies can identify new themes at any rate as right on time as content inconsistency based methodologies, and now and again much prior when the point is inadequately recognized by the printed substance in the posts. In this, we concentrate on informal organizations, for example, Facebook and Twitter, which increasing more significance in our everyday life. Since the data traded over informal organizations are testing test beds for the investigation of information mining. Specifically, we are keen on the issue of recognizing developing subjects from social streams, which can be utilized to make mechanized "breaking news", or find concealed market needs or underground political developments. Contrasted with traditional media, online networking can catch the soonest, unedited voice of conventional individuals. Subsequently, the test is to recognize the rise of a theme as ahead of schedule as could be allowed at a moderate number of false positives we are distinguishing rising points from informal community streams in light of observing the specifying. Conduct of users. Our fundamental supposition is that another (rising) theme is something individuals have a craving for talking about, remarking, or sending the data further to their companions. All the aforementioned ponders make utilization of literary substance of the records, yet not the social substance of the reports. The Twitter Stream (joins) has been used here.*

Keywords: *Authentication, aggregation, anomaly-detection, social network, burst detection.*

I. Introduction

The news information is becoming immensely continuously so the new ideas are getting added to the web. The center is towards the new subjects which can be found by mapping a portion of the beforehand talked about or distributed information. Social networking stages have advanced a long ways past detached assistance of online social communications. It is the need of a hour to break down the data content in online social networking (news articles, websites, tweets and so forth.). It permits business to comprehend general sentiment about strategies and items. In a large portion of these cases, information focuses show up as a surge of high dimensional element vectors. We return to the issue of web taking in of points from online networking content in certifiable modern organization situations. On one hand, the measurements of approaching information focuses is adjusted by the subjects powerfully and on the other h and, early identification of new patterns is vital in numerous applications. Past strategies propose online nonnegative grid factorizations structure. This system is predominantly used to catch the development and rise of topics in unstructured content under a novel worldly regularization structure. An advancement calculation is created for this system furthermore to stream Twitter information. Rising topics are quickly caught by the past framework. Additionally past framework can track the current themes after some time while keeping up worldly consistency and can be expressly arranged to tie the measure of data being introduced to the client. This model is utilized to gauge the abnormality of future client conduct. Utilizing the proposed likelihood model, we can quantitatively gauge the curiosity or conceivable effect of a post reflected in the saying conduct of the client. A frequency based approach for the most part relies on the frequencies of (printed) words happening in the social posts. This expels the verbal and descriptive word like words and considers just the nonverbal parts of the post. Word recurrence is figured for every word which will be taken primarily for extraction of the point. The restriction is that an expression recurrence based methodology could experience the ill effects of the equivocalness brought about by equivalent words or homonyms (plurals).It can't be connected when the substance of the messages are generally non-printed data. For eg "good life relies on upon liver", where liver might be organ or living individual, so there will be an uncertainty issue. We can't make a difference the system when the substance is nonliterary data. Social information mining has a few difficulties like identification of subjects, blasts, topic designs from content, exception location and change focuses. To defeat these difficulties there are assortments of techniques and

models have been proposed. In spite of the fact that, finding the inconsistency is again the testing errand. The new system is composed from a bringing together perspective that a subject structure in a content stream is displayed utilizing a limited burst model.

II. Related Work

In this paper [1] client finds the developing subjects from the interpersonal organizations. As the data traded in the informal communities post incorporates the content, as well as pictures, URLs and video thusly traditional term recurrence based methodologies may not be proper in this setting. In view of the reacts from several users in interpersonal organizations post is utilized to distinguish the development of new themes. In this paper likelihood model is proposed to catch various notice per post and the recurrence of users happening in the notice. The burden is all the algorithm displayed in the paper was led disconnected. In this paper [2] model choice in Gaussian direct relapse utilization of the standardized greatest probability which postures disturbing in light of the fact that the standardization coefficient is not limited. The point of the model choice is not used to pick the right model, but rather it is utilized to minimize future expectation mistakes. SNLS is a best strategy with a little edge. In [5] creator has checked the event of points in a flood of occasions. There are a few calculations, delivers altogether different results to screen the event of subjects. Kleinberg's blasted model and Shasha's blasted model are utilized to screen. It functions admirably to track theme blasts of MeSH terms in the bio logical Literature; it can likewise be utilized for determining approaching blasts and energy based subject elements burst model have a critical favorable position. The inconvenience is Hierarchical structure merits more prominent consideration on burst. In this paper [6] Normalization produces standardized most extreme probability (NML) dissemination. Successive standardized most extreme probability (SNML) is simpler to register and incorporate an arbitrary procedure. SNLS is the best strategy, except for the littlest example sizes. AIC, BIC, PLS, SNLS strategies is utilized to appraise the request of an AR Model. BIC is known not an inclination to think little of as opposed to overestimate the request. Likewise, it is not very astounding that AIC, which from the earlier supports more perplexing models than the other criteria, wins for the littlest example size. The issue was considered identifying with gatherings of information where every study inside a gathering is a draw from a mix model. Yee Whye Tech, Michael I, Jordan, Matthew J, Beal, and David M. Blei [7] has speaks to various leveled Dirichlet process in the term of the stick breaking process that gives irregular measures expressly, a chinese eatery handle that is alluded as "Chinese eatery establishment" depicts a representation of peripheral's as far as a urn model and representation of the procedure regarding a definition of three MCMC inspecting plans for back surmising. In this technique to the issue sharing bunches among numerous related gatherings is a nonparametric Bayesian methodology.

III. Proposed System

We focus on emergence of topics signaled by social aspects of networks. Specifically, we focus on mentions of users— link between users that are generated dynamically through replies, mentions, and retweets. We propose a probability model of the mentioning behavior of a social network user, and propose to detect the emergence of a new topic from the anomalies measured through the model. Aggregating anomaly scores from hundreds of users, we show that we can detect emerging topics only based on the reply/mention relationships in social-network posts. We propose a probability model that can capture the normal mentioning behavior of a user, which consists of both the number of mentions per post and the frequency of users occurring in the mentions. Using the proposed probability model, we can quantitatively measure the novelty or possible impact of a post reflected in the mentioning behavior of the user. We aggregate the anomaly scores obtained in this way over. The flow of the system is given in Fig.1.In this system, the main challenge is to detect the anomaly detection. Dataset is collected from the real world data from twitter. In that dataset, it contains the attributes such as username, friend list, followers, screen name, and last tweet date etc.. This is the input process of our project. After the dataset insert into the database, we eliminate the null or unwanted values in the next process. Eliminating the null or unwanted values in the dataset. It is called as dataset pre-processing.

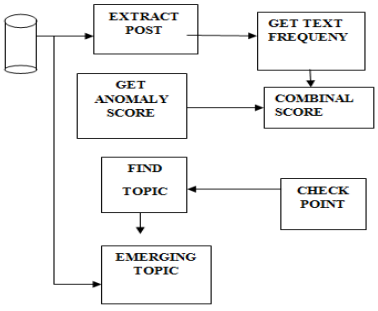

Fig 1. Proposed System Architecture

Data Aggregation

After the dataset has been pre-processed, then aggregating the data from the database. In the data aggregation is processed based on the mentions, replies and retweets in the dataset. Through this, we can easily identify who posts the mentions, replies and mentions in the network. In this, we are aggregated based on the description of user's posts information.

Probability Estimation

The probability estimations consists of two types of distributions. They are predictive distribution and joint probability distribution.

a) Predictive distribution is estimated based on the estimating the probability values based on the mention and mentions in the network.

b) Joint probability distribution. It is estimated based on the number of users in the network and number of users who posts the mentions in the network.

To implement the Probability Estimation Values based on the classification of mentions and replies and retweets from the pre-processing data set. First find out the Probability density function, by using number of mentions user v in the dataset and total number of mentionees in dataset. Then we estimate predictive distribution using the equation (1)

Predictive Distributions = mv/mentioness (1)

Number of mentions user v in the dataset is that total number user mentions the post.

Burst Detection

Burst detection is nothing but a detecting the anomaly which is based on the time series. In this, id, url, joined date and last date are the important parameters to detect. The burst-detection method is based on a probabilistic automaton model with two states, burst state and non-burst state.

Pseudo code for Burst Detection

```
-----------------------------------------------------------------
        for each Data
                Select join date and last tweet date
                Calculate burst detection
  BT=join date-last tweet date   if  BT ≥0
                        return burst state
                else
  return Non burst state  end for
-----------------------------------------------------------------
```

Pseudo code for Anomaly Detection

```
-----------------------------------------------------------
Input     :  Twitter dataset
Output    :   Anomaly detection
-----------------------------------------------------------
For all record
        Pre-process the data
 Identify mentions, replies, retweets.
Calculate joint probability distribution
 k=modulo of mentions.
Calculate Predictive distribution
Number of mentions to the user v in the dataset t.
Calculate Burst Estimation process.
Difference Value between join date and last tweet date.
Classify using Bayesian rule,
Calculate decision rule.
return Anomaly Score Aggregation end for
-----------------------------------------------------------
```

Dynamic Threshold Optimization (DTO) Algorithm

Algorithm 1. Dynamic Threshold Optimization (DTO) [19].

Given: $\{Score_j \mid j = 1, 2, \ldots\}$: scores, N_H: total number of cells, ρ: parameter for threshold, λ_H: estimation parameter, r_H: discounting parameter, M: data size

Initialization: Let $q_1^{(1)}(h)$ (a weighted sufficient statistics) be a uniform distribution.

for $j = 1, \ldots, M-1$ **do**

Threshold optimization: Let l be the least index such that $\sum_{h=1}^{l} q^{(j)}(h) \geq 1 - \rho$. The threshold at time j is given as

$$\eta(j) = a + \frac{b-a}{N_H - 2}(l+1).$$

Alarm output: Raise an alarm if $Score_j \geq \eta(j)$.

Histogram update:

$$q_1^{(j+1)}(h) = \begin{cases} (1-r_H)q_1^{(j)}(h) + r_H & \text{if } Score_j \text{ falls into the } h\text{th cell,} \\ (1-r_H)q_1^{(j)}(h) & \text{otherwise.} \end{cases}$$

$$q^{(j+1)}(h) = (q_1^{(j+1)}(h) + \lambda_H)/(\textstyle\sum_h q_1^{(j+1)}(h) + N_H\lambda_H).$$

end for

IV. Deriving Link-Anomaly Score

We compute the link anomaly score for each post separately. Anomaly score is defined as the user's deviation from the post. The comments are either good or bad weather related to the post are determined by using link anomaly score. Accordingly, the link-anomaly score is defined by the following diagram. **Step1**: Compute anomaly score of a new post x=(t,u,k,v) K-mention,v-user,u-user,t-time.

Step2: Find s(x) $\mathbf{s(x)= -\log(p(k|T_u^{(t)}} \prod_{v \in V} P(v|T_u^{(t)})$

$\mathbf{=-\log(p(k|T_u^{(t)}} \sum_{v \in V} \log P(V|T_u^{(t)})$

Step3: By using training set which consist of both number of user and mention compute anomaly score.

Step4: Finally we aggregate the anomaly score obtained for the post.

V. Results and Outcome

The description of the data set used in this work is tabulated in Table 1

Dataset Used: Twitter Data Set

Total number of Data : 138

Total number of Attributes : 15

Tables

Table 1: Attributes In Dataset

S.No	ATTRIBUTES	DESCRIPTION
1.	ID	Id for the user
2.	Name	Name of the user
3.	JOINDATE	Join Date on twitter
4.	LASTTWEET DATE	Last Tweet Data on Twitter
5.	LANGUAGES	Languages used in Twitter
6.	SCREEN NAME	Twitter Name
7.	PROTECTED	Security process

First the Data set are browsed from the system and data are inserted in to database Then pre-process the data , the data contain null or missed values are eliminated from the database. After Pre-process Data we have 88 data. After pre-process unwanted data are removed from the dataset, and values are updated in the database. To calculate the Mentioness in the Data is that total number of user in the dataset. Here count the total number of mentioness, mentions, replies and retweets is tabulated in Table 2.

Table 2: Calculate Mentions Replies and Retweet

S.No	CLASSIFY DATA	COUNT
1.	Mentioness	88
2.	Replies	14
3.	Retweets	25
4.	Mentions	42

After classify, need to calculate the Predictive Distribution by using total number of mentionees and mentions in the data set is showed in Table 3.

Table 3: Predictive Distribution

S.No	DATA	DISTRIBUTION
1	Number of Mention to user V in the datas	0.477277

Dataset are cluster based on verified and the data taken from the cluster that are count the language from the data, count the cluster information from the cluster debases. To use this we assign A as cluster and B as languages to do Bayes Rule is showed in Table 4.

Table 4: Bayesian Value

S.No	BAYESIAN VALUE
1.	00081828
2.	0.146139
3.	0.006535
4.	0.006535
5.	0.029411

After calculate the Bayesian classification then attributes selection measure from the data using Information gain and Matrix estimation and finally aggregate the anomaly link based on the URL for each user that are separated in each cluster1 and cluster 2.Then finally estimate the anomaly score values and classify the instances is shown in the Table 5.

Table 5: Anomaly Instance

S.NO	INSTANCE	TOTAL
1.	Normal Instance	88
2.	Anomaly Instance	24

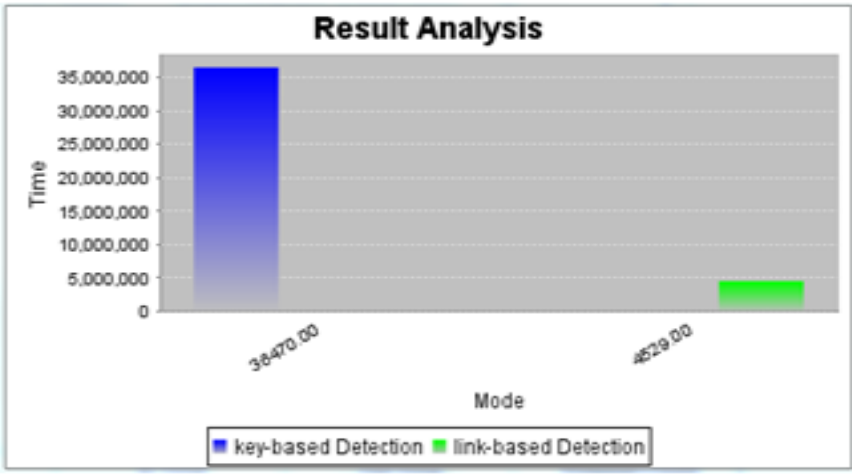

Fig 2: 14th Indian trend topics

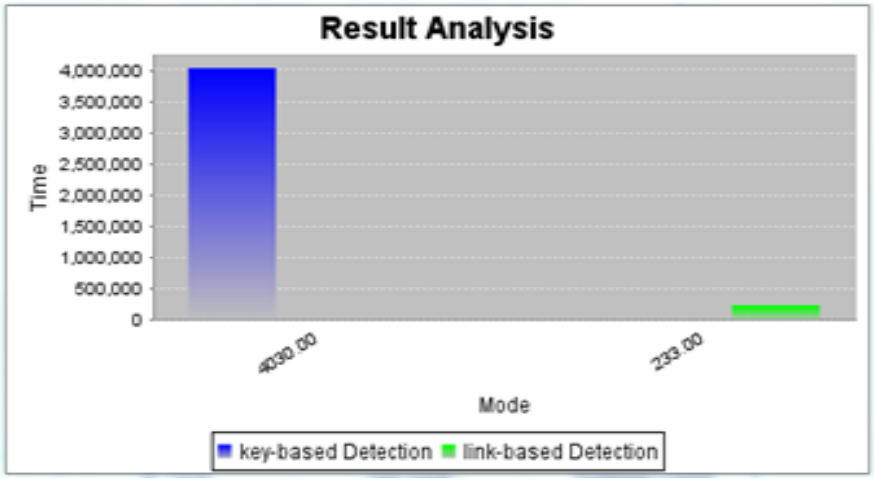

Fig 3: 8th ODI trends topics

VI. Conclusion

In this paper we are interest in detecting emerging topics from social network streams based on monitoring the mentioning behavior of users. Our basic assumption is that a new (emerging) topic is something people feel like discussing, commenting, or forwarding the information further to their friends. Conventional approaches for topic detection have mainly been concerned with the frequencies of (textual) words. A term-frequency-based approach could suffer from the ambiguity caused by synonyms or homonyms. It may also require complicated preprocessing (e.g., segmentation) depending on the target language. The proposed probability model determines both number of mentions per post and the frequency of the mentionee and this approach is used to detect the emergence of topics in a social network stream .We have put forward a probability model that captures both the number of mentions per post and frequency of mentioning .The text frequency based methods used to determine how many times the text gets repeated and from that the repeated words are considered We combined the proposed mentioned model with the SDNML change point detection algorithm to pin point the emergence topic ,the link anomaly based approach have detected emergence of the topic even earlier than the keyword based approach that use hand chosen keywords. It will be more effective when combining both text anomaly based and link anomaly based approach.

Future Scope

The four data sets included a wide-spread discussion about a controversial topic ("Job hunting" data set), a quick propagation of news about a video leaked on YouTube ("YouTube" data set), a rumor about the upcoming press conference by NASA ("NASA" data set), and an angry response to a foreign TV show ("BBC" data set). In all the data sets, our proposed approach showed promising performance. In three out of four data sets, the detection by the proposed link-anomaly based methods was earlier than the text-anomaly-based counterparts. Furthermore, for "NASA" and "BBC" data sets, in which the keyword that defines the topic is more ambiguous than the first two data sets, the proposed link-anomaly-based approaches have detected the emergence of the topics even earlier than the keyword-based approaches that use hand-chosen keywords. All the analysis presented in this paper was conducted offline, but the framework itself can be applied online. We are planning to scale up the proposed approach to handle social streams in real time. It would also be interesting to combine the proposed link-anomaly model with text-based approaches, because the proposed link-anomaly model does not immediately tell what the anomaly is. Combination of the word-based approach with the link-anomaly model would benefit both from the performance of the mention model and the intuitiveness of the word-based approach.

References

[1]. Toshimitsu Takahashi, Ryota Tomioka, and Kenji Yamanishi, "Discovering Emerging Topics in Social Streams via Link-Anomaly Detection," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014.
[2]. B.G. Obula Reddy, Dr. Maligela Ussenaiah, "Literature Survey on Clustering Techniques," IOSR Journal of Computer Engineering, Volume 3, pp 01-12.
[3]. VARUN CHANDOLA, ARINDAM BANERJEE, VIPIN KUMAR, "Anomaly Detection: A Survey," A modified version of this technical report will appear in ACM Computing Surveys, September 2009.
[4]. Artur Silie, Lovro Zmak, Bojana Dalbelo, MarieFrancine Moens, "Comparing Document Classification using K-means Clustering".
[5]. A. Ghose and P. G. Ipeirotis, "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics," IEEE Trans. Knowl. Data Eng., vol. 23, no. 10, pp. 14981512. Sept.2010.
[6]. K.A. Kontogiannis, R. Demori, M. Galler, M. Bernstein,"Pattern matching for Clone and Concept Detection," Automated Software Engineering Volume 3, pp 77-108, 1996.
[7]. Genrikh Altshuller, "Concept Generation," Soviet patent investigator, 1950.
[8]. Prof. Nam Suh, "Axiomatic Design for Concept Generation," MIT.
[9]. Ankan Saha and Vikas Sindhwani: 2012," Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization,".
[10]. Victoria J. Hodge, "A survey of outlier Detection Methodologies," Kluwer Academic Publisher, Netherlands, 2004.

Authors:

Rajyalakshmi Golla is a student of Computer Science & Engineering from QIS Institute of Technology, She Presently pursuing M.Tech (CSE) in this college.

R Lakshmi Tulasi is a Professor, H.O.D of QIS Institute of Technology, Ongole. She received M.Tech from JNTUCEA. She is pursuing Ph.D. at JNTUH. She is a good Researcher in semantic web, Computer Networks. She attended Various National and International Workshops and Conferences.