Elysium Journal of Engineering Research & Management

P-ISSN: 347-4408
E-ISSN: 2347-4734

A PROBABILISTIC APPROACH FOR PREDICTING ANOMALIES IN SOCIAL STREAMS

Abinaya.B1, Chellamaal.P[2]

[1]PG Student, Department of Computer Science and Engineering,
JJ College of Engineering and Technology,
Anna University, TamilNadu, India.
Mail id:abinayasker.12@gmail.com
[2]Assitant Professor (SE.G), Department of Computer Science and Engineering,
JJ College of Engineering and Technology,
Anna University, Tamil Nadu, India.
Mail id:chellachella@rediffmail.com

***Abstract*— Basic presumption is that a new (emerging) topic is something people feel like discussing, commenting, or forwarding the information further to their friends. The proposed approach, spot the emergence of topics in a social network stream. It focus on the social aspect of the posts reflected in the mentioning behavior of users instead of the textual contents. A probability model is proposed that captures both the number of mentions per post and the rate of occurrence of mentioned. The detection of emerging topics is the most vital renewed interest in the fast growth of social networks.**

To detect the anomalies in the social network and the detection is based on the links between the users that are generated dynamically. It has been classified through replies, mentions and retweets. A probability model is used to capture the mentioning behavior of a social network user, and detect the emergence of a new topic from the anomalies measured through the model. It aggregates the anomaly scores based on the reply/mention relationships in social network posts. The real data sets gathered from Twitter and implement using the technique called SDNML, burst detection and Bayesian.

Index Terms— **Topic detection, anomaly detection, social networks, sequentially discounted normalized maximum-likelihood coding, burst detection**

I. INTRODUCTION

Communication over social networks, such as Twitter and Facebook, is gaining its importance in our daily life. Since the information exchanged over such social networks are not only texts but also URLs, images, and videos, they are challenging test beds for the study of data mining. In specific, we are interested in the problem of detecting emerging topics from social streams, which could be used to create automated "breaking news", or discover hidden market needs or underground political movements. Comparing conventional media to social media, it is possible to capture the earliest, unedited voice of ordinary people. Thus the challenge is to detect the emergence of a topic as early as possible at a moderate number of false positives.

We are concerned in detecting trends topics from social network streams based on mentioning behavior of users. Basic presumption is that a new (emerging) topic is something people feel like discussing about, commenting about, or forwarding the information further to their friends. Early approaches for topic discover have mainly been concerned with the frequencies of words. In this method, initially, the social network is shown in a graph, and then similarity among users, then the graph is divided into smaller. Afterwards, all the similar profiles to the real profile are collected, then strength of relationship is calculated, and less strength of relationship will be verified by mutual friend system. In this study, in order to evaluate proposed method, all steps are applied on a dataset of Facebook, Twitter, Google+, and lastly this work is compared with two previous works by applying them on the dataset. Along with probability model can capture the normal mentioning behaviour of a user, this probability model consists of both the number of mentions per post and the frequency of users occurs in the mentions. Then probability model is used to measure the anomaly of future user behaviour. Using this model, quantitatively measure the novelty or possible impact of a post reflected in the mentioning behaviour of the user.

Aggregate the anomaly scores from the different users and apply to the change-point detection technique based on the Sequentially Discounting Normalized Maximum Likelihood (SDNML) coding. This technique can detect a change in the related dependence structure in the time series of aggregated anomaly scores, and pin-point to detect the emerging topic. To show that this approach can detect the emergence of a new topic at least as fast as using the best term that was not obvious at the moment.

II. RELATED WORK

In a broad range of private enterprise dealing with text streams, including social network, knowledge management, and stream monitoring services, it is a major issue to find topic trends and examine their dynamics in real-time [8]. For instance, it is desired in the social stream area to grasp a

P-ISSN: 347-4408
E-ISSN: 2347-4734

current topics in online user profess every day and to track a new topic as soon as it become visible. A topic defined here is a seminal event or activity detection and detection of topics have been studied in the area of topic detection and tracking (TDT) [1]. In this factor, the job is to either find a new document into one of the known topics or to detect none of the known categories. Another, secular structure of topics have been modeled and examined through dynamic model selection [8], temporal text mining [6], and factorial hidden Markov models [4]. Another type research is distressed with the notion of "bursts" [3] in a stream of documents. The Bursts was modeled by Kleinberg using the time varying Poisson process with a hidden discrete process that controls the firing rate. Based on the change in the momentum a physics-inspired model of bursts of topics was developed by him and Parker [2]. Recently, the textual emotion mining [10] assigns emotion for the words from text document and based on the preference level it achieve emotion for the whole document. The social media such as blogs and social networks has raised interest in sentiment analysis. SentiStrength [7] used to extract positive and negative sentiment strength from short informal text. The main novel contribution of this work are: a machine learning approach to optimize sentiment term weightings; procedure for extracting sentiment from repeated letter non-standard spelling in informal text; and a related spelling correction method. A novel approach to notice compromised user accounts [5] in social networks, it does not depend on the presence of URLs in messages. As a result, it can detect a broad range of malicious messages, including scam message that contains telephone number and instant messaging contacts.

All the above mentioned make use of word content of the documents, but not the social content of the documents. The social content i.e. link has been used in the study of citation networks. However, citation networks are often analyzed in a stationary setting. The originality of the current paper lies in focusing on the social content of the documents (posts) and in combining this with a change-point analysis.

III. P R O B L E M RESEARCH & DISCUSSION

This section presents about the different research area and literature view on Probability model, Link-Anomaly Score, Dynamic Threshold Optimization (DTO) which will demonstrate the pervious view and challenges besides our proposal.

3.1 Probability Model

In this subsection, it describes the probability model that we used to capture the normal mentioning behavior of a user and how to train the model**.**

There are two types of infinity we have to take into account here.

1. **The first is the number k of users mentioned in a post.**
2. **The second type of infinity is the number of users one can possibly mention.**

Conditional Probability Formula

$$P(B/A) = \frac{P(A \text{ and } B)}{P(A)}$$

Then estimate predictive distribution based on the mentions given by the users which is in the dataset using Bayesian formulas.

Joint Probability Distributions
k=Modulo of V
m=Summation of number of mentions in the training set

Predictive Distributions

m is the number of total mentions

P(V/T)=mv/
m

Number of mentions to user v in the dataset t

Pseudocode (for finding error rate)

```
function train( i)
   { Instances++
   if (++N[$Klass]==1) Klasses++

  for(i=1;i<=Attr;i++)

        if (i !=
        Klass)

       if ($i !~
         /\?/)

        symbol(i,$i,$Klass)

}

function symbol(col,value,klass) {
   Count[klass,col,value]++;
}
```

Pseudocode Explanation

For each
attribute,

For each value of the attribute, make a rule as follows: count how often each class appears find the most frequent class make the rule assign that class to this attribute-value

Calculate the error rate of the rules

Choose the rules with the smallest error rate

For PDF (probability density function) calculation

P-ISSN: 347-4408
E-ISSN: 2347-4734

The probability density function for the normal distribution is defined by the mean and standard Dev (standard deviation)

Given:

- **n**: the number of values;
- **sum**: the sum of the values; i.e. sum = sum + value;
- **sumSq**: the sum of the square of the values; i.e. sumSq = sumSq + value*value

Then:

```
function mean(sum,n)  {

return sum/n

    } functionstandardDeviation(sumSq,sum,n)
{ returnsqrt((sumSq-((sum*sum)/n))/(n-1))
   }

functiongaussianPdf(mean,standardDev,x) {
pi= 1068966896 / 340262731; #: good to 17 decimal places
return 1/(standardDev*sqrt(2*pi)) ^
(-1*(x-mean)^2/(2*standardDev*standardDev))
```

}

Challenges:

In past, some recent topic model-based methods have been proposed to discover and summarize the evolutionary patterns of themes in temporal text collections. However, the theme patterns extracted by these methods are hard to interpret and evaluate. To produce a more descriptive representation of the theme pattern, we not only give new representations of sentences and themes with named entities, but however, sentence-level probabilistic model based on the new representation pattern are not satisfied. Compared with other topic model methods, this approach only gets each topic's distribution per term, but also generates candidate summary sentences of the themes as well. Consequently, the results are not easier to understand and can be evaluated using the top sentences produced this probabilistic model. Experimentation with the new proposed methods on the sample social network dataset shows that the proposed methods are useful in the discovery of evolutionary theme patterns.

3.2 Link-Anomaly detection

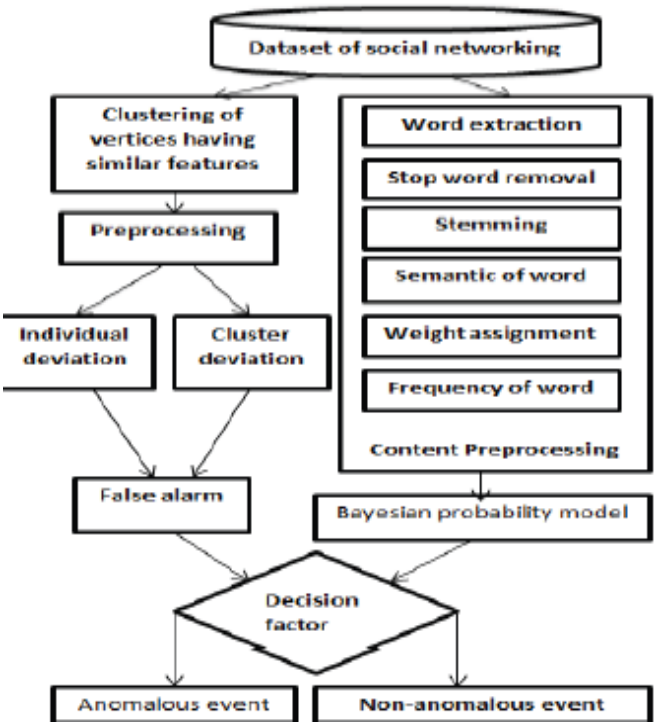


Fig 1. Flow diagram of link anomaly detect

Challenges:

a) Detection of User Cluster with Suspicious Activity Group of users with suspicious activities has to be identified using anomaly detection shown as Fig.

b) Approach to detect suspicious profiles on social platforms Aim of a dynamic approach is to alert the users of Smartphone users about suspicious profiles located in his or her close circle of contacts on a given social network.

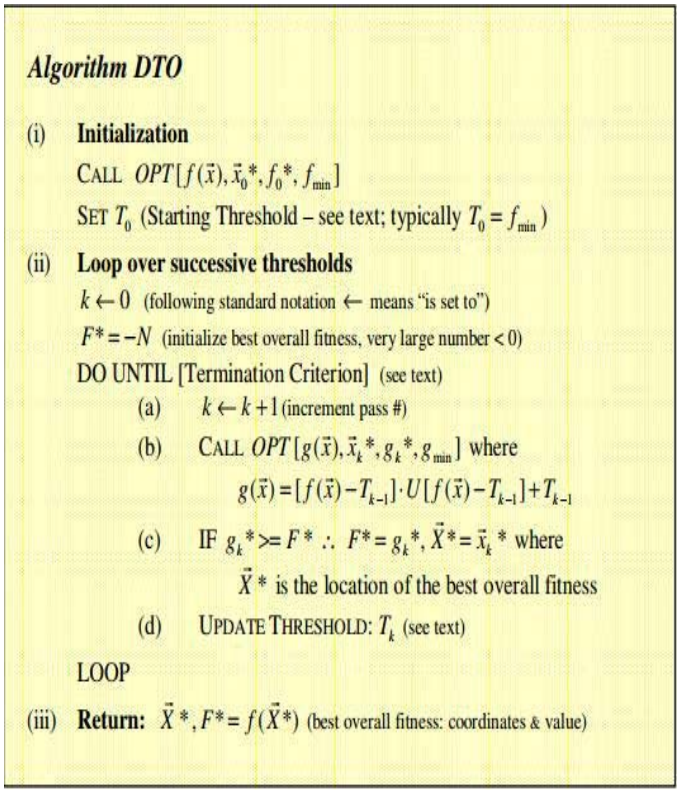
c) Detection of Random Link Attacks Malicious users create false identities and used it to communicate with innocent users. While detecting random Link Attack mining social networking graph which is extracted from user interaction in communication network is important

d) Threat Detection through Graph Learning and Psychological Context

e) Detection of Emerging Topics via Link-Anomaly Detection in Social Streams Main focus is on detecting emerging topics from social network streams based on monitoring the mentioning behaviour of users.

P-ISSN: 347-4408
E-ISSN: 2347-4734

3.3 Dynamic Threshold Optimization (DTO)



Algorithm DTO

(i) **Initialization**

CALL $OPT[f(\vec{x}), \vec{x}_0^*, f_0^*, f_{min}]$

SET T_0 (Starting Threshold – see text; typically $T_0 = f_{min}$)

(ii) **Loop over successive thresholds**

$k \leftarrow 0$ (following standard notation $\leftarrow$ means "is set to")

$F^* = -N$ (initialize best overall fitness, very large number < 0)

DO UNTIL [Termination Criterion] (see text)

(a) $k \leftarrow k+1$ (increment pass #)

(b) CALL $OPT[g(\vec{x}), \vec{x}_k^*, g_k^*, g_{min}]$ where

$$g(\vec{x}) = [f(\vec{x}) - T_{k-1}] \cdot U[f(\vec{x}) - T_{k-1}] + T_{k-1}$$

(c) IF $g_k^* >= F^* \therefore F^* = g_k^*, \vec{X}^* = \vec{x}_k^*$ where

$\vec{X}^*$ is the location of the best overall fitness

(d) UPDATE THRESHOLD: T_k (see text)

LOOP

(iii) **Return:** $\vec{X}^*, F^* = f(\vec{X}^*)$ (best overall fitness: coordinates & value)

Challenge

DTO appears to be an effective technique for adaptively changing the topology of the decision space in a multidimensional search and optimization problem. DTO should be useful with any search and optimization algorithm. Bounding DS from below removes local maxima, and as the threshold or "floor" is increased, more and more local maxima are eliminated. In the limit, DS collapses to a plane whose value ("height") corresponds to the value of the global maximum. In that case, DS contains no information as to the global maximum's location, but the maximum's value is known precisely. In order to preserve location information, the DTO threshold should not be set too high, thereby retaining enough structure for efficient DS exploration. There are many unanswered questions concerning how DTO should be implemented. For example, there almost certainly are better ways to set the threshold than the simple linear scheme used here. Thresholds that are progressively closer together probably will work better. Another question arises in connection with what optimization algorithm should be used. Even though DTO is algorithm-independent (i.e.Algorithm1), it may work best when different algorithms are combined to take advantage of their different strengths and weaknesses.

Naive Bayes

Naive Bayes classifiers are a popular statistical technique of email filtering. They typically use bag of words features to Identify spam email, an approach commonly used in classification. Naive Bayes classifiers work by correlating the use of tokens (typically words, or sometimes other things), with spam and non-spam emails and then using Bayesian inference to calculate a probability that an email is spam or not.

- Fast to train (single scan) and fast to classify
- Not sensitive to irrelevant features
- Handles real and discrete data
- Handles streaming data well

IV.P R O P O S E D MODEL

The probability model that capture the normal mentioning behaviour of a user that consists of both the number of mentions per post and the rate of occurrence of mentioned (who are mentioned in the post). This model is used to measure the anomaly score from divergent users. Using this model, it quantitatively measures the novelty or possible impact of a post displayed in the mentioning behaviour of the user. Aggregate the anomaly scores from divergent users and relate to change point detection technique based on the sequentially discounting normalized maximum-likelihood (SDNML) coding. This technique can spot a change in the related dependence structure in the time series of aggregated anomaly scores, and pinpoint the topic exposure is detected.

V. EXPERMENTAL RESULTS

Data sets is collected from Twitter. Each data set is associated with a list of posts in a service called Together; together is a collaborative service where people can tag Twitter/Facebook posts that are related to each other and organize a list of posts that belong to a certain topic. The aim is to determine whether the emergence of the topics recognized and collected by people are detected by the proposed system. Different data sets are selected each corresponding to a user organized list in Together. Collect posts from users for each data set that appeared in each list (participants).

Twitter Dataset

The dataset was crawled from geographic networks on Facebook. Geographic networks were used to group people together who lived in the same area. These networks allow, in default, anyone in the network to see all the post from all other members. Thus, it was easy, at the time, to collect millions of messages by creating a minimal number of profiles and join one of these geographic networks.

Facebook Dataset

On average, receiving tweets from more than 500,000 distinct users per hour. Unfortunately, because of the API request limit, we were not able to generate profiles for all users that

P-ISSN: 347-4408
E-ISSN: 2347-4734

we saw in the data stream. Thus, as discussed in the previous section, we first cluster messages into groups that are similar. Then, starting from the largest cluster, we start to check whether the messages violate the behavioural profiles of their senders. We do this, for increasingly smaller clusters, until our API limit is exhausted. On average, the created groups consisted of 30 messages. This process is then repeated for the next observation period. To determine the weights that we have to assign to each feature, we applied proposed model to a labeled training dataset for both Twitter and Face book. While the Face book dataset contains the network of a user, Twitter does not provide such a convenient proximity feature. Therefore, we omitted this feature from the evaluation on Twitter.

VI. C O N C L U S I O N AND FUTURE WORK

In this paper, proposed a new approach to spot the exposure of topics in a social network stream. Instead of the textual contents, the basic scheme of our approach is to emphasis the social aspect of the posts reflected in the mentioning behavior of users. The number of mentions per post and the rate of occurrence of mentioned is captured by a probability model is proposed. The proposed model is integrated with the SDNML change-point detection algorithm and Kleinberg's burst- detection model to speck the exposure of a topic. Since the proposed method does not rely on the textual contents of social network posts, it is durable to rephrasing and it can be applied to the case where topics are concerned with information other than texts, such as images, video, audio, and so on.

REFERENCES:

[1] Allan James, Jaime Carbonell , George Doddington, Jonathan Yamron , and Yiming Yang, "Topic Detection and Tracking Pilot Study: Final Report," *Proc.DARPA Broadcast News Transcription and Under-standing Workshop*, 1998.

[2] Dan He and D.S. Parker, "Topic Dynamics: An Alternative Model of Bursts in Streams of Topics," *Proceedings 16th ACM SIGKDD International Conference in Knowledge Discovery and Data Mining,* pp. 443-452, 2010.

[3] Kleinberg. J, "Bursty and Hierarchical Structure in Streams," *Data Mining Knowledge Discovery,* vol. 7, no. 4, pp. 373-397, 2003.

[4] Krause.A, J. Leskovec, and C. Guestrin, "Data Association for Topic Intensity Tracking," *Proc. 23rd Int'l Conf. Machine Learning (ICML' 06),* pp. 497-504, 2006.

[5] Manuel Egele, GianlucaStringhini, Christopher Kruegel, and Giovanni Vigna, "COMPA: Detecting Compromised Accounts on Social Networks" *in Proceedings of Network & Distributed System Security Symposium (NDSS),* 2013.

[6] Mei.Q and C. Zhai, "Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining," *Proc. 11th ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining,* pp. 198-207, 2005.

[7] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai , Kappas A, "Sentiment Strength Detection in Short Informal Text" *Journal of the American Society for Information Science and Technology,* 61(12), 2544–2558, 2010.

[8] Morinaga. S and K. Yamanishi, "Tracking Dynamics of Topic Trends Using a Finite Mixture Model," *Proc.10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,* pp. 811-816, 2004.

[9] Sattikar A. A, Dr. R. V. Kulkarni, "Natural Language Processing For Content Analysis in Social Networking" *International Journal of Engineering Inventions ISSN: 2278-7461,*Volume 1, Issue 4 pp: 06-09, 2012.

[10] Sujitha.S, S.Selvi, "A Model of Textual Emotion Mining From Text Document" *Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT),*Vol.2,Special Issue 1, 2014.

[11] Y. Urabe, K. Yamanishi, R. Tomioka, and H. Iwai, "Real-Time Change-Point Detection Using Sequentially Discounting Normalized Maximum Likelihood Coding," *Proc. 15th Pacific-Asia Conf.Advances in Knowledge Discovery and Data Mining (PAKDD' 11),* 2011.

[12] J. Rissanen, T. Roos, and P. Myllyma¨ki, "Model Selection by Sequentially Normalized Least Squares," *J. Multivariate Analysis,* vol. 101, no. 4, pp. 839-849, 2010.

[13] C. Giurc_aneanu, S. Razavi, and A. Liski, "Variable Selection in Linear Regression: Several Approaches Based on Normalized Maximum Likelihood," *Signal Processing,* vol. 91, pp. 1671-1692, 2011.

[14] K. Yamanishi and J. Takeuchi, "A Unifying Framework for Detecting Outliers and Change Points from Non-Stationary Time Series Data," *Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,* 2002.

[15] S. Lee and J. Kim, "WarningBird: Detecting Suspicious URLs in Twitter Stream," *in Symposium on Network and Distributed System Security (NDSS),* 2012.