www.ijecs.in
*International Journal Of Engineering And Computer Science ISSN:2319-7242*
*Volume 4 Issue 3 March 2015, Page No. 10918-10923*

# A Probabilistic Approach for Anomaly Detection In Social Streams

***S. Sundara Selvi, Mr. K.Durairaj***
MCA Final Year
Veltech Technical University
Avadi,Chnnai-62
E-Mail-sundara2812@gmail.com
Asst.Prof IT dept
Veltech Technical University
Avadi,Chnnai-62
E-Mail-durairajk@veltechuniv.edu.in

## Abstract

This Project is used to measure to abnormality of future user behavior of users It proposed a possibility model that can capture the normal mentioning behaviour of a user, which consists of both the number of mentions per post and the frequency of users occurring in the mentions. Conventional-term-frequency-based approaches may not be appropriate in this context, because the information exchanged in social-network posts include not only text but also images, URLs, and videos. Our basic assumption is that a new (emerging) topic is something people feel like discussing about, commenting about, or forwarding the information further to their friends. It shows that this approach can detect the emergence of a new topic At leastas fast as using the best term that was not obvious at the moment. 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 demonstrate our technique in several real data sets we gathered from Twitter. The proposed link-anomaly based method can detect the emergence of the topics earlier than keyword frequency based methods.

***Keywords: Topic Detection, link-anomaly , burst detection, social network.***

## I. Introduction

This Project Conventional approaches for topic detection have mainly been concerned with the frequencies of (textual) words.

We are interested in detecting emerging topics from social network streams based on monitoring the mentioning behaviour of users.

The information exchanged over social networks such as Facebook and Twitter is not only texts but also URLs, images, and videos, they are challenging test beds for the study of data mining. Our basic assumption is that a new (emerging) topic is something people feel like discussing about, commenting about, or forwarding the information further to their friends.

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.

It cannot be applied when the contents of the messages are mostly non-textual information. The "words" formed by mentions are unique, requires little prepossessing to obtain (the information is often separated from the contents), and are available regardless of the nature of the contents.

## II. Problem Statement

- Conventional approaches for topic detection have mainly been concerned with the frequencies of (textual) words.
- We are interested in detecting emerging topics from social network streams based on monitoring the mentioning behaviour of users.
- The information exchanged over social networks such as Facebook and Twitter is not only texts but also URLs, images, and videos, they are challenging test beds for the study of data mining.
- Our basic assumption is that a new (emerging) topic is something people feel like discussing about, commenting about, or forwarding the information further to their friends.
- 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.
- It cannot be applied when the contents of the messages are mostly non-textual information.
- The "words" formed by mentions are unique, requires little prepossessing to obtain (the information is often separated from the contents), and are available regardless of the nature of the contents.

***The Disadvantages are as following.***

- 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.
- It cannot be applied when the contents of the messages are mostly non-textual information.
- The "words" formed by mentions are unique, requires little prepossessing to obtain and are available regardless of the nature of the contents.
- The keyword was manually selected to best capture the topic.
- The sparsity of the keyword frequency seems to be a bad combination with the SDNML method.
- A drawback of the keyword-based dynamic thresholding is that the keyword related to the topic must be known in advance.

## III. Proposed System

- Using this project, we can quantitatively measure the novelty or possible impact of a post reflected in the mentioning behaviour of the user.
- This project is used to measure the anomaly of future user behaviour.
- It proposed a probability model that can capture the normal mentioning behaviour of a user, which consists of both the number of mentions per post and the frequency of users occurring in the mentions.
- It aggregate the anomaly scores obtained in this way over hundreds of users and apply a recently proposed change-point detection technique based on the Sequentially Discounting Normalized Maximum Likelihood (SDNML) coding.
- This technique can detect a change in the statistical dependence structure in the time series of aggregated anomaly scores, and pinpoint where the topic emergence is.
- The effectiveness of the proposed approach is demonstrated on two data sets we have collected from Twitter.
- The proposed link-anomaly based method can detect the emergence of the topics earlier than keyword frequency based methods.

***The advantages are as following.***

- The proposed method does not rely on the textual contents of social network posts.
- It is robust to rephrasing.The probability model that captures both the number of mentions per post and the frequency of mentionee.
- It can be applied to case where Topics are concerned with the information exchanged are not only texts but also images, URLs, and videos.
- The proposed link-anomaly based approach performed even better than the keyword-based approach on "NASA" and "BBC" datasets

## *IV. Algorithm as Proposal*

**Dynamic Threshold Optimization (DTO)**

As a final step in our method, we need to convert thechange-point scores into binary alarms by thresholding.Since the distribution of change-point scores may changeover time, we need to dynamically adjust the threshold toanalyze a sequence over a long period of time. In this subsection, we describe how to dynamically optimize thethreshold using the method of dynamic threshold optimization.

**Algorithm 1** Dynamic Threshold Optimization (DTO) [13]

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

**Initialization:** Let $q_1^{(1)}(h)$ 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**

**Kleinberg's Burst-Detection Method**

In addition to the change-point detection based on SDNML followed by DTO described in previous sections,we also test the combination of our method with Kleinberg's burst-detection method [2]. More specifically,we implemented a two-state version of Kleinberg's burstdetection

Model hierarchical structure. The burst-detection method is basedon a probabilistic automaton model with two states, burst state and nonburst state.

**Scalability of the Proposed Algorithm**

The proposed link-anomaly-based change-point detection ishighly scalable. Every step described in the previoussubsections (Step 1-Step 6) requires only linear time against the length of the analyzed time period. Computation of thepredictive distribution for the number of mentions (4) can be computed in linear time against the number of mentions.Computation of the predictive distribution for the mention probability in (5) and (6) can be efficiently performed usinga hash table. Aggregation of the anomaly scores fromdifferent users takes linear time against the number of users, which could be a computational bottle neck but canbe easily parallelized. SDNML-based change-point detection requires two swipes over the analyzed time period.Kleinberg's burst-detection method can be efficiently implemented with dynamic programming.

## *V. System Architecture*

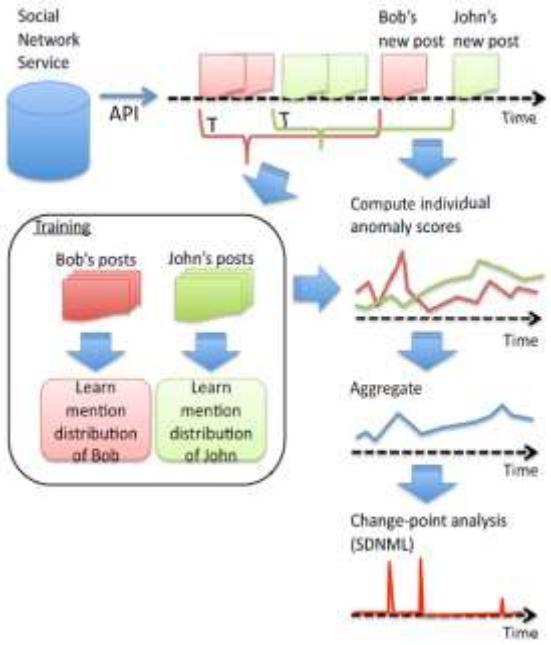


***Fig 1: System Architecture***

We describe the probability model that we use to capture the normal mentioning behaviour of a user and how to train the model. We characterize a post in a social network stream by the number of mentions k it contains, and the set V of names (IDs) of the mentionees (users who are mentioned in the post). There are two types of infinity we have to take into account here. The first is the number k of users mentioned in a post. Although, in practice a user cannot mention hundreds of other users in a post, we would like to avoid putting an artificial limit on the number of users mentioned in a post. Instead, we will assume a geometric distribution and integrate out the parameter to avoid even an implicit limitation through the parameter. The second type of infinity is the number of users one can possibly mention

COMPUTING THE LINK-ANOMALY SCORE

In this subsection, we describe how to compute the deviation of a user's behaviour from the normal mentioning behaviour modeled In order to

compute the anomaly score of a new post x = (t, u, k, V ) by user u at time t containing k mentions to users V , we compute the probability with the training set T (t) u , which is the collection of posts by user u in the time period [t−T, t] (we use T = 30 days in this paper). Accordingly the link-anomaly score is defined as the two terms in the above equation can be computed via the predictive distribution of the number of mentions, and the predictive distribution of the mentionee.

COMBINING ANOMALY SCORES FROM DIFFERENT USERS Combining Anomaly Scores from Different Users: we describe how to combine the anomaly scores from different users; The anomaly score in is computed for each user depending on the current post of user u and his/her past behaviour T (t) u . In order to measure the general trend of user behaviour, we propose to aggregate the anomaly scores obtained for posts x1, . . . , xn using a discretization of window size τ > 0 as follows: where xi = (ti, ui, ki, Vi) is the post at time ti by user ui including ki mentions to users Vi.

BURST DETECTION METHOD In addition to the change-point detection based on SDNML followed by DTO described in previous sections, we also test the combination of our method with Kleinberg's burst detection method . More specifically, we implemented a two-state version of Kleinberg's burst detection model. The reason we chose the two-state version was because in this experiment we expect no hierarchical structure. The burst detection method is based on a probabilistic automaton model with two states, burst state and non-burst state. Some events (e.g., arrival of posts) are assumed to happen according to a time varying Poisson processes whose rate parameter depends on the current state.

## VI. Experimental Results

We generated synthetic data sets over 20 days from100 users as we describe below. For each user, we assume that posts arrive from a Poisson process and draw the interpost intervals from an exponential distribution with a fixed rate. The number of mentions in each post is drawn from a geometric distribution with parameter. We generated two data sets. In the first data set, which we call "Synthetic100" For the burst-detection approach, we used the firing rateparameter of the Poisson point process

**Table 1.0 Proposed Change PIN Point Detection Anomaly in The parameter**

| P | youtube | Post | Detection | Pin Point | Data |
|---|---|---|---|---|---|
| 0.02 | 4 | 1 | 9 | 8 | 4 |
| 0.06 | 4 | 5 | 11 | 10 | 4 |
| 0.07 | 6 | 6 | 24 | 15 | 4 |
| 0.1 | 8 | 7 | 28 | 18 | 4 |

p-parameter

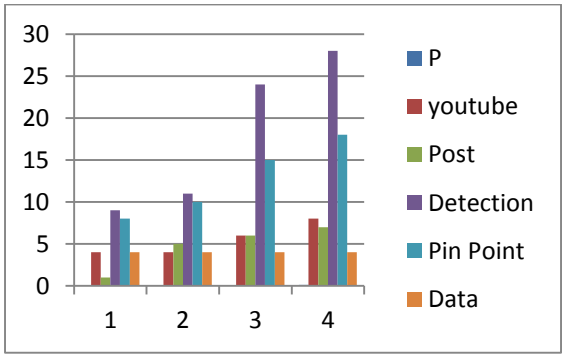


**Fig 2. PIN Point Detection Anomaly**

**CO NCLUSION:**

Our project have proposed a new approach to detect the emergence of topics in a social network stream. The basic idea of our approach is to focus on the social aspect of

the posts reflected in the mentioning behavior of users instead of the textual contents. We have proposed aprobability model that captures both the number of mentions per post and the frequency of mentionee. We have combined the proposed mention model with the SDNML change-point detection algorithm and Kleinberg's burst-detection model to pinpoint the emergence of a topic. Since the proposed method doesnot rely on the textual contents of social network posts, it is robust to rephrasing and it can be applied to the casewhere topics are concerned with information other than texts, such as images, video, audio, and so on.

## References:


[1] 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.

[2] D. He and D.S. Parker, "Topic Dynamics: An Alternative Model of Bursts in Streams of Topics," Proc. 16th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 443-452, 2010.

[3] T. Roos and J. Rissanen, "On Sequentially Normalized Maximum Likelihood Models," Proc. Workshop Information Theoretic Methods in Science and Eng., 2008.

[4] 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.

[5] 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.

[6 ] C. Giurc_aneanu and S. Razavi, "AR Order Selection in the Case When the Model Parameters Are Estimated by Forgetting Factor Least-Squares Algorithms," Signal Processing, vol. 90, no. 2, pp. 451-466, 2010.