



## Survey on Link Anomaly Detection for Textual Stream in Online Social Network

**Chandan M G**

PG Scholar, Dept. of CSE,  
NMAMIT, Nitte, India

**Chandra Naik**

Assistant Professor, Dept. of CSE,  
NMAMIT, Nitte, India

---

**Abstract**— *Link Anomaly detection is one of the most important topics in social network. Many of the social networks such as Facebook, Google+, LinkedIn, or twitter require an effective and efficient framework to identify deviated data. Anomaly detection methods are typically implemented in social stream mode, and thus cannot be easily extended to large-scale problems without sacrificing computation where the user's link is generated dynamically (replies, mentions, and retweets). A new approach model i.e. probability model, this model to the capture normal linking behavior of a social network users, and propose to detect the trending topic from the social networks through the probability model. We collect anomaly score from the different user. And aggregated score feed to change-point analysis or change-point detection, or with burst detection, finally show that to detect trending topics only based on the reply/mention in social network posts. Our technique to collect number of real data from real time twitter account.*

**Keywords**— *Topic Detection, Anomaly Detection, Social Networks, SDNML, Burst Detection.*

---

### I. INTRODUCTION

Now-a-days social networking sites are becoming the main communication media among individuals and organizations. Social network are web-based application, social network individuals to construct a public or semi-public profiles, articulate list of other users with whom they share a connection, and traverse and view their list of connections and those made by others in the system. People share their personal information, photos, videos, URLs, ideas on these sites. People live in contact with their family, friends and colleague. However leakage of personal information creates security problem, cyber bullying, spreading the hatred messages etc. Malicious users may cause many severe issues like De-Anonymization attack, neighbourhood attack, profile cloning, social phishing, spam attacks and many more. Hence development of reliable anomaly detection in social networking sites is extremely important [4]. In this sense, mention means like a language, this mention contain number of words equal to the number of user in a social media.

We are interested in detecting trends topics from social network streams based on mentioning behaviour of users. 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 [2]. Early approaches for topic discover have mainly been concerned with the frequencies of words. In this method, first, the social network is shown in a graph, and then similarity among users, this graph is divided into smaller communities [7]. Afterwards, all of 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 finally 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 [2]. Using this model, quantitatively measure the novelty or possible impact of a post reflected in the mentioning behaviour of the user.

We 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 our 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 [2].

### II. PROBLEM STATEMENT

In a wide range of commercial areas dealing with text streams, including social network, knowledge management, and stream monitoring services, it is an important issue to discover topic trends and analyse their dynamics in real-time. For example, it is desired in the social stream area to grasp a new trend of topics in online user claims every day and to track a new topic as soon as it emerges. A topic is here defined as a seminal event or activity detection and detect of topics have been studied in the area of topic detection and tracking. In this context, the main task is to either a new document into one of the known topics or to detect the none of the known categories. Alternative, temporal structure of topics that have been modeled and analyzed through dynamic model selection, temporal text mining, and factorial hidden Markov models. Another type research is concerned with the notion of “bursts” in a stream of documents. 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 utilized in the study of citation networks. However, citation networks are often analyzed in a stationary setting. The novelty of the current paper lies in focusing on the social content of the documents (posts) and in combining this with a change-point analysis [2].

**Major problems are mentioned below:**

- 1) 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
- 2) Suspicious activities are not been monitored when user information are stored in database called user history Current and previous information is used to update user information which alerts about suspicious user.
- 3) To focus is on detecting emerging topics from social network streams based on monitoring the mentioning behaviour of users.

**III. PROBLEM RESEARCH & DISCUSSION**

This section present 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,

**Probability Model**

**Literature View:**

Earlier, S. Morinaga and K. Yamanishi “Tracking dynamics of topic trends using a finite mixture model”, ” In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’04), 2004 and following to that G. Fung, J. Yu, H. Liu, P. Yu. Time-dependent event hierarchy construction. In Proceedings of the 13<sup>th</sup> ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (KDD’07), 2007”.

**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.

**Link-Anomaly detection**

**Literature View:**

H. Gao, Jun Hu, T. Huang, J. Wang and Y. Chen,”Security issues in online social networks”, IEEE Internet Computing Journal, vol. 15, no. 4, pp. 56-62, 2011and K. Gani, H. Hacidand R. Skraba, ”Towards Multiple Identity Detection in Social Networks”, In Proceedings of the 21st ACM international conference companion on World Wide Web, pp. 503-504, 2012. K. Hanumantha Rao, G. Srinivas, Ankam Damodhar, M. Vikas Krishna, “Implementation of Anomaly Detection Technique Using Machine Learning Algorithms,” International Journal of Computer Science and Telecommunications, volume 2, Issue 3, June 2013.

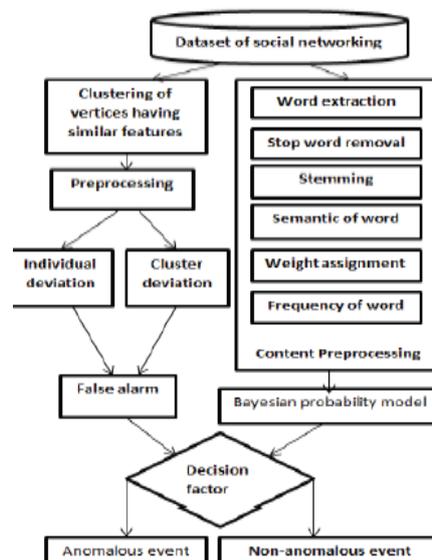


Figure 1: Anomaly detection

**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 1.
- 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[5][6].
- 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 [7].
- d) Threat Detection through Graph Learning and Psychological Context [8].
- 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.

**Dynamic Threshold Optimization (DTO)**

**Literature View:**

Wang, J., "Particle Swarm Optimization with Adaptive Parameter Control and Opposition" J. Computational Information Systems, Vol. 7, No. 12, 2011, pp. 4463-4470 (<http://www.jofcis.com>) Hayes, B., "Quasirandom Ramblings," American Scientist magazine, vol. 99, July-August 2011, pp. 282-287 ([www.americanscientist.org](http://www.americanscientist.org)).

```

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  $g_k^* \geq F^* \therefore F^* = g_k^*, \vec{X}^* = \vec{x}_k^*$  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)
    
```

**Algorithm 1: DTO**

**Challenges:**

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 [10].

**IV. PROPOSED MODEL**

The probability model that consists of capture the normal mentioning behaviour of a user that consists of both the number of mentions per post and the frequency mentionee (who are mentioned in the post). This model is used to measure the anomaly score from different users. Using this model, quantitatively measure the novelty or possible impact of a post reflected in the mentioning behaviour of the user. And aggregate the anomaly scores from different users and apply to 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 pinpoint the topic emergence is detected. The effectiveness of the approach is collected four data sets from twitter.

The keyword-based approach can only be used when we are expecting a burst of tweets mentioning the prespecified keyword, which could happen if we were making an advertisement campaign or any other kind of manipulation. However, here it should be regarded as a sanity check, since we are interested in automatically detecting the emergence of a topic without any intervention. Therefore, our goal is to detect trends topics as fast as the keyword-based methods.

## V. EXPERIMENTAL RESULTS

We collected data sets from Twitter. Each data set is associated with a list of posts in a service called Together1; 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. Our goal is to evaluate whether the proposed approach can detect the emergence of the topics recognized and collected by people. We have selected different data sets each corresponding to a user organized list in Together. For each data set we collected posts from users that appeared in each list (participants).

### Twitter Dataset

The dataset was crawled from geographic networks on Facebook. Geographic networks were used to group together people that lived in the same area. The default privacy policy for these networks was to allow anybody in the network to see all the posts from all other members. Therefore, it was easy, at the time, to collect millions of messages by creating a small number of profiles and join one of these geographic networks [13].

Table 1: Test facebook and twitter dataset

Network & Similarity Measure	Twitter Text		Twitter URL		Facebook Text	
	Groups	Accounts	Groups	Accounts	Groups	Accounts
<b>Total Number</b>	374,920		14,548		48,586	
<b># Compromised</b>	9,362	343,229	1,236	54,907	671	11,499
<b>False Positives</b>	4% (377)	3.6% (12,382)	5.8% (72)	3.8% (2,141)	3.3% (22)	3.6% (412)
<b># Bulk Applications</b>	12,347		1,569		N/A	N/A
<b># Compromised Bulk Applications</b>	1,647	178,557	251	8,254	N/A	N/A
<b>False Positives</b>	8.9% (146)	2.7% (4,854)	14.7% (37)	13.3% (1,101)	N/A	N/A
<b># Client Applications</b>	362,573		12,979		N/A	N/A
<b># Compromised Client Applications</b>	7,715	164,672	985	46,653	N/A	N/A
<b>False Positives</b>	3.0% (231)	4.6% (7,528)	3.5% (35)	2.2% (1,040)	N/A	N/A

### Face book Dataset

On average, we received 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 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 [14]. 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. For Twitter, the weights for the features are determined from our labeled training dataset consisting of 5,236 (5142 legitimate, 94 malicious) messages with their associated feature values as follows: Source (3.3), Personal Interaction (1.4), Domain (0.96), Hour of Day (0.88), Language (0.58), and Topic (0.39) are shown in Table 1.

## VI. CONCLUSIONS

Thus we have discussed different techniques of anomaly detection. We have proposed the model of anomaly detection in social networking site by integrating two approaches first is link anomaly detection and second is text anomaly detection which will generate more accurate results through the following new approach to detect the trending topics in a social network stream It should be noted that although detecting fake proof can stop greater extent of deception in future, prevention is better than cure because it is enough for an attacker to observer users' detail once. Therefore, teaching users to prevent cloning attacks so that not accepts friend requests when they do not know the sender. It can be developed as Facebook application which each user can run it on his/her profile and also some fuzzy methods can be used to typed information in user's profiles. Proposed new approach i.e. Probability model captures both the number of mentions per post and the frequency of mentionee. Further feed into change-point detection algorithm and burst detection model to pin-point the trending of a topic.

## REFERENCES

- [1] Faraz Rashid, Reda Alhadj, "A Framework for Periodic Outlier Pattern Detection in Time-Series Sequences," IEEE Transactions on Cybernetics, Vol.44, No.5, May2014.
- [2] 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 2013.

- [3] Shenghua Bao, Shengliang Xu, Li Zhang, Rong Yan, Zhong Su, Dingyi Han, and Yong Yu, "Mining Social Emotions from Affective Text," IEEE Transactions on Knowledge and Data Engineering Vol. 24, No. 9, September 2012.
- [4] D. Boyd and N. B. Ellison, "Social Network Sites: Definition, History, and Scholarship," Journal Computer-Mediated Communication, vol.13, no. 1-2, Nov. 2007
- [5] Sharath Kumar A and Sanjay Singh, "Detection of User Cluster with Suspicious Activity in Online Social Networking Sites," Second International Conference on Advanced Computing, Networking and Security, 15-17 Dec. 2013
- [6] Charles PEREZ, Marc LEMERCIER, Babiga BIRREGAH, "A dynamic approach to detecting suspicious profiles on social platform," IEEE International Conference on communications Workshops (ICC), 9-13 June 2013
- [7] Oliver Brdiczka, Juan Liu, Bob Price, Jianqiang Shen, Akshay Patil, Richard Chow, Eugene Bart, Nicolas Ducheneaut, "Proactive Insider Threat Detection through Graph Learning and Psychological Context," IEEE CS Security and Privacy Workshops, 24-25 May 2012
- [8] Nisheeth Shrivastava, Anirban Majumder, Rajeev Rastogi, "Mining (Social) Network Graphs to Detect Random Link Attacks," IEEE 24<sup>th</sup> International conference on data engineering, pp.486-495, 7-12 April 2008.
- [9] Jiong Zhang, Mohammad Zulkernine, and Anwar Haque, "Random- Forests-Based Network Intrusion detection system" IEEE Transactions on Systems, Man, and Cybernetic, Vol. 38, No. 5, September 2008
- [10] Kush R. Varshney, "Bounded Confidence Opinion Dynamics in a Social Network of Bayesian Decision Makers," IEEE Journal of Selected Topics In Signal Processing, Vol. 8, No. 4, August 2014
- [11] Yang Li, Bin-Xing Fang, "A Lightweight Online Network Anomaly Detection Scheme Based on Data Mining Methods," International Conference on Network Protocols, pp.340-341, 16-19 Oct. 2007
- [12] Alexander Y. Liu and Dung N. Lam, "Using Consensus Clustering for Multi-view Anomaly Detection," IEEE CS Security and Privacy Work, pp.117-124, 24-25 May 2012.
- [13] Gabriel Weimann, "Terror on Facebook, Twitter, and Youtube," The Brown Journal of World Affairs, volume 16, issue 2, 2010
- [14] Mr. A. A. Sattikar, Dr. R. V. Kulkarni, "Natural Language Processing For Content Analysis in Social Networking," International Journal of Engineering Inventions, Volume 1, Issue 4, pp.06-09, September 2012 .