International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Link-Anomaly Detection in Twitter Streams

Shari P S

M. Tech Student, Department of Computer Science, Mount Zion College of Engineering Pathanamthitta

Abstract: ***Rapid growth of social network gives* emergence *to the detection of emerging topics. The information exchanged over social network post not only includes text but also images, URLs and videos therefore conventional frequency based appropriate in this context. By taking into consideration the links between users that are generated dynamically through replies ,mentions, and retweets are included. This paper highlights the analysis of a probability model that mention the behavior of a social network user. This model is used to detect the anomalies emerged . From hundreds of users anomaly scores are aggregated. In the proposed system it is only based on replay/mention relationship and is experiment zed with in real datasets gathered from twitter.***

Keywords: social network, anomaly detection, term-based approach, dynamic threshold optimization, topic detection and tracking.

1. Introduction

COMMUNICATION is an unavoidable aspect in our daily life. Communication progresses with the usage of social network sites such as Face book and Twitter. Here the information's exchanged are not only texts but also URLs, images and videos. When comparing to conventional approaches, social media can capture the earliest ,unedited voice of ordinary people. To detect the emergence of a topic as early as possible at a moderate number of false positives is a challengeable factor. The *mentions* are responsible for making a social media social. *Mentions* can also be treated as *links* in the form of message-to, replay-of or explicitly in text. Any number of mentions can be included in a single post. Mentions can be of different types. Some people may include mentions in their posts, some mentions their friends all the times, some may receive mentions all the time and some others depends on rare occasions. Detection of emerging topics depends upon the monitoring the mentioning behavior of users. A new topic can be identified by the discussion, comments and forwarding the information between friends and followers. Conventional approaches for topic detection concerned with the frequencies of (textual) words [1],[2]. In a term – based frequency –based approach ,synonyms or homonyms leads to ambiguity .

Fig 1 shows an example of the emergence of a topic through posts on social networks. Alice and John are friends of Bob therefore the first post contains mentions to Alice and John. Bob replies John in the second post but it is also visible to many friends of John they are not direct friendly to Bob.Dave, one of John's friends, forwards (called retweet in Twitter) the information further down to his own friends in the third post. From this figure we can understood that the messages mentioned between users are kept secret (textual information,) only links in the text is visible.

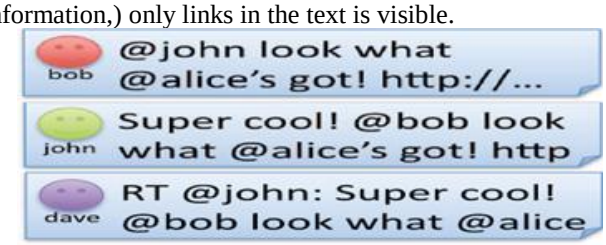

Figure 1: Demonstration of the emergence of a topic in social streams

Fig. 2. Shows the overall flow of the proposed method. This technique is used to detect a change in the statistical dependence structure in the time series of aggregated anomaly scores, and it is concentrated in collecting data sets collected from Twitter. Hence proved that mention-anomaly-based approaches can detect the emergence of a new topic in a faster way.

2. Related Work

Detection and tracking of topics have been studied extensively in the area of topic detection and tracking(TDT)[1]. The main task is to either classify a new document into one of known topics or to detect that belongs to none of the known categories. Temporal structure of topics has been modeled and analyzed through dynamic model selection[4],temporal text mining [5] and Factorial hidden Markov models[6] .

These studies makes use of textual contents of the documents. It does not focuses on social content of the documents. The basic concept behind current paper is based on social contents of documents (posts) along with change point analysis.

3. Proposed Method

From the overall flow of the proposed method we can analyses that each step in the flow consists of corresponding subsections. Through some API, the data driven from a social network service is in a sequential manner. For each new post we uses the samples of previous time interval with in a given training phase. For each probability distribution we assign an anomaly score for individual posts. The score is then aggregated over number of user's posts and then fed into a change point analysis technique. Change-point Detection is accomplished via SDNML Coding technique. By monitoring the compressibility nof a new piece of data SDNML detects a change in statistical dependence structure of a time series. Dynamic Threshold Optimization(DTO) is used to dynamically adjust the threshold to analyze a sequence over a long period of time.

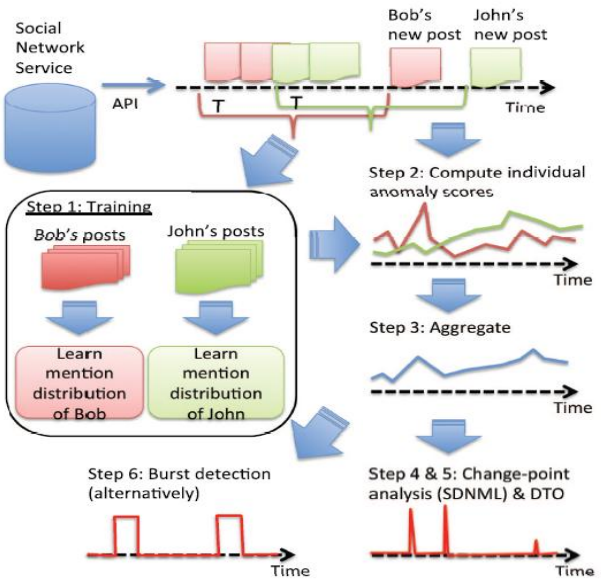

Figure 2: Block diagram for overall flow of the proposed method

4. Scope of the Project

In a social network stream ,a new approach is proposed to detect malicious URLs in emerging topics. The basic idea is to focus the social aspects of posts reflected in mentioning behavior of users instead of textual contents. This method relay on the textile contents of social network posts. Probability model to capture the behavior of social network user is proposed. It can be accomplished through large data set collected and labeled from twitter.

Probability model will capture the normal mentioning behavior of the user, which consists of both the number of mentions per post and frequency of users occurring in the mentions. Then the hundreds of users will accomplish in this method. In order to identify change point detection technique, sequentially discounting normalized maximum-likelihood (SDNML) coding is adopted. This technique can detect a change in statistical dependence structure in the time series of aggregated anomaly score, scores and pinpoint where the topic emergencies. The effectiveness of proposes approach is demonstrated by the four data sets collected from twitter. Thus proved the mentioned anomaly-based approaches can detect the emergence of a new topic as fast as text anomaly based counter parts.

5. Module Description

There are mainly five modules in the proposed system. They are stream listener module, tweet filter module, tweet packing module, window processing mopdule and keyword ranker module.

5.1Stream Listener Module

This module receives streaming data in the form of Twitter messages, which can come directly from the Twitter API or some other source. This data is parsed and encapsulated. After the encapsulation of each message it is en-queued in memory for the next module in the pipeline. It should be noted that message encapsulation is prone to delays caused by the Internet bandwidth connection and it leads to data loss and hence affects Twitter's information delivery rate.

5.2 Twitter Filter Module

This module discards messages which are not written in languages accepted by our system. We perform language classification using a Naive Bayes classifier. This module also standardizes tweets according to the following rules:

a) Marks separation : Replacing special characters and removal of accents, apostrophes.
b) Data Standardization : Replacement of special characters and conversion of upper and lower case

5.3 Tweet Packing Module

Filtered and standardized tweets in queue Q1, are grouped into a common set determined by creation timestamp**.** This set of tweets, represent an individual time-window. It is important to note that the arrival of tweets maintains a chronological order. In the case that an old or delayed tweet appears, it is included in the current window. Each of these windows is sent to the following stage for processing.

5.4 Window Processing Module

Each keyword, composed of a single or adjacent word n-grams, is mapped into a hash table data structure. This structure manages keywords in addition to the information of its two adjacent windows and their rates. Consider as n-grams the n ordered correlative words.

O (1) ,is the best time complexity for most cases and O(n) is the worst case complexity when collisions occur. This process is detailed in the algorithm. This data structure controls the complexity of the algorithm with optimal insertions and search O(1).

5.5 Keyword Ranker Module

Bursty keywords are included implicitly into the hash table. Therefore, we extract bursty keywords by discarding those that do not classify as having a positive relevance variation. We discard non-bursty keywords using the criteria described below.This to helps prevent the size of the hash table from growing out of control:

1) It is the first occurrence of the keyword. We must wait until the next window to check it again.
2) We observe a negative variation in frequency rates between adjacent windows.
3) Low arrival rate: Many words do not appear frequently. We discard these words if the average arrival rate is lower than 1.0 (keywords per time-window). The remaining keywords are sorted in descending order according to their Relevance Variation Rate. In top positions keywords with the highest variation rate or burstiness are ranked .

6. Background Study

Social network deals with communication paradigms. Communication concept consists of audio and video format. Consentrating on sites such as Face Book and Twitter not only messages but also links(mentions) appears. Conventional approach deals with text- based paradigm. Here links are not considered. Proposed system considers links along with messages and rank them based on probability model.

7. Result Analysis

This paper implemented in ASP.Net programming language. The main advantage of this paper is identified; analyze posts and links with in twitter streams. Following screenshots represents the output of the work carried out on the project.

Figure 3: Registration Form

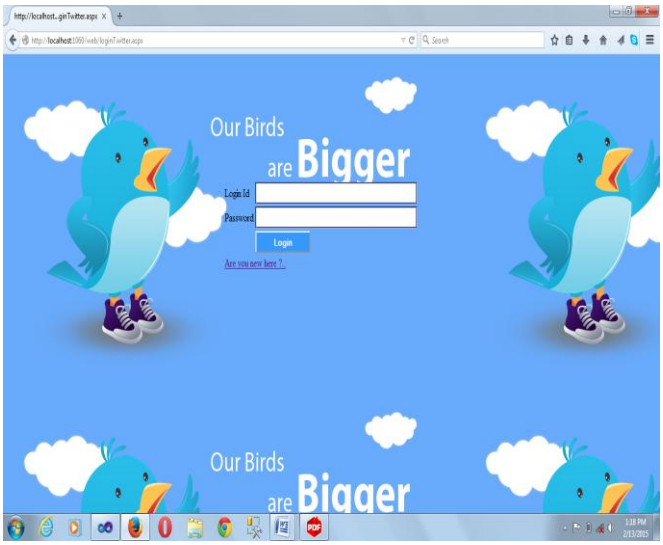

Figure 4: Login Page

Figure5: User Home Page

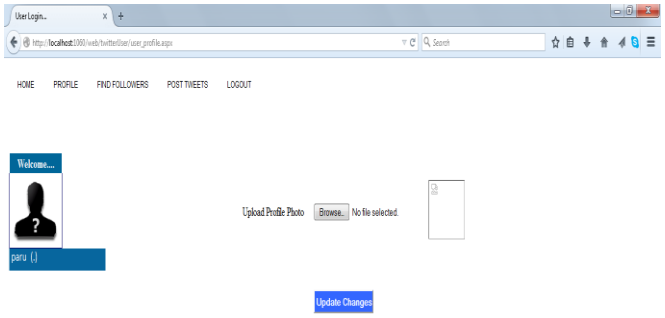

Figure 6: User Profile

Figure 9: Posting tweets

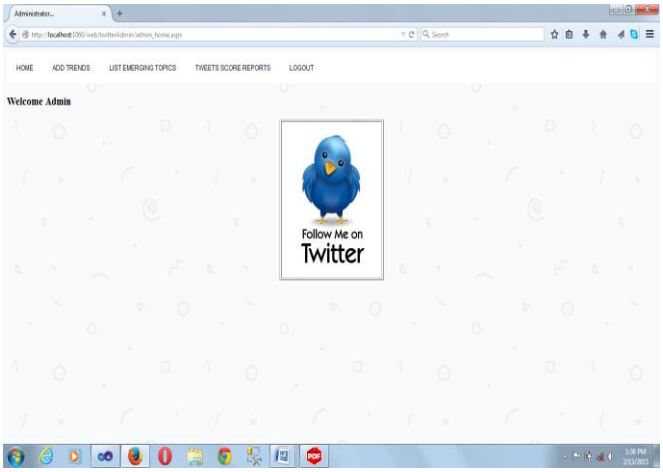

Figure 10: Admin Home Page

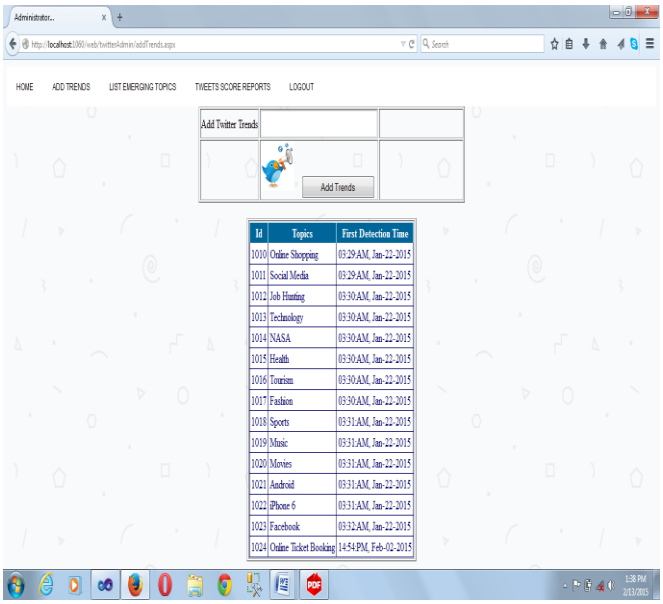

Figure 11: Adding Trends

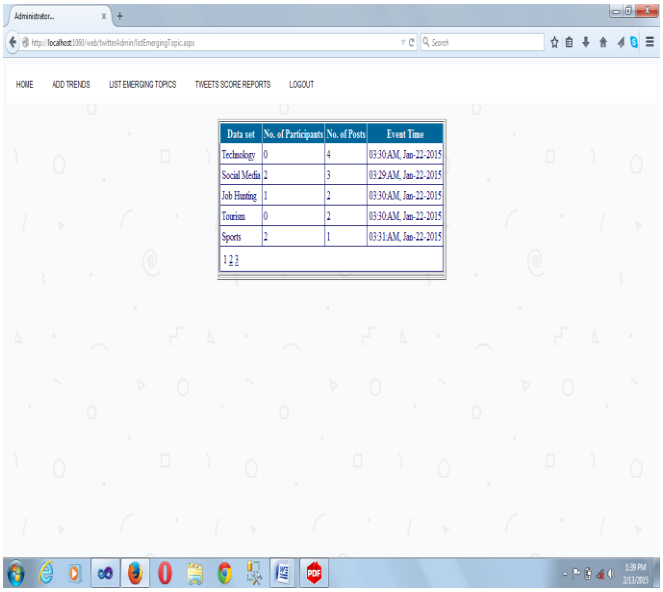

Figure 12: Listing emerging topics

These screen shot shows all the real time example based on the concept of our proposed system.

8. Conclusion

Link – Anomaly Detection is a new approach for real time detection of malicious URLs in emerging topics.. The basic idea of this approach is to focus on the social aspect of the posts reflected in the mentioning behavior of users in twitter instead of the textual contents. It is used to detect malicious tweets quickly and massively. Conventional suspicious URL detection systems are ineffective on the conditional redirections.

A probability model is proposed to capture both the number of mentions per post and the frequency of mentionee. Links between users that are generated dynamically through replies, mentions, and re tweets are taken into consideration.This system can effectively handle the conditional redirections. The new features of suspicious URLs are discovered. In the future work it can be adapted to other services like Face Book and LinkedIn.

References

[1] J. Allan et al., "Topic Detecgtion and Tracking Pilot Study: Final Report,"*Proc DARPA Broadcast News Transcription and Understanding Workshop*,1998.

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

[3] 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(PAKKDD' 11),* 2011.

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

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

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

Author Profile

Shari P S received the Btech degrees in ComputerScience Engineering from Caarmel Engineering College ,Perunad in 2011. Currently doing Mtech degree in Computer Science Engineering under Mahatma Gandhi University.