SYSTEM TO USE TEXT STREAM MINING TECHNIQUES FOR IDENTIFICATION OF ABNORMAL BEHAVIOR

[1]Prachi Bobade,[2]Geetanjali Boke,[3]Prajakta Kshirsagar,[4]Pranita Musande
[1]Ashwini Khairkar
UG Student Supervisor
Department Of Information Technology Engineering
Bharati Vidyapeeth's College Of Engineering For Women's Pune

Abstract :

A blog that is smaller than a traditional blog and contains very short entries is called micro blog. RING proposed a graph analytic approach in which elastic search and spark is used to design graph. Temporal evolution tracking used in this system recovers the evolution process of anomalous event to trace its origin. RING is among the first to discover emerging anomalies correlations in a streaming fashion, RING works on the real time data that work on the minutes to months' data. Social networking sites promote timely and active discussions and comments towards products, market as well as public events and have attracted lot of attentions from organizations. So each and every twit or post is important. It is necessary to keep eye on each post to find out the anomalies or any abnormal things to avoid unnecessary fights and help users to concentrate on current trend only. We are using full text indexing engine and graph processing system because they carry out heavy computation workload during anomaly detection and monitoring. After detecting anomaly, user gets notified about its presence with message request to remove or delete it.

***Key words*— RING, Twitter text stream, blog and microblogs, emerging, anomalies.**

I. INTRODUCTION

Social media are interactive Web-2.0 based applications like Facebook and Twitter etc. Microblogs platforms are extremely popular in the big data era due to real-time nature and viral diffusion of information. Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, and updating and information privacy. Social media can have positive and negative impacts. Social media can help to improve individuals' sense of connectedness with real or online communities and social media can be an effective communication (or marketing) tool for corporations, entrepreneurs, nonprofit organizations, including advocacy groups and political parties and governments. A popular component and feature of Twitter is retweeting. Twitter allows other people to keep up with important events; stay connected with their peers, and can contribute in various ways throughout social media. Retweeting is beneficial strategy, which notifies individuals on Twitter about popular trends, posts, and events. On the basis of these popular trends some abnormal things (anomalies) are happen so in this project we are introducing RING System, to identify the anomalies over text streams.

Problem Definition

Early detection of emerging anomalies using RING before they go viral to find out the correlated keywords with anomalies, ranking events according to their importance and popularity, and to find out the noisy data.

Objectives

- To find emerging topic within short time
- Real-time detection of anomalies
- Notify user about the anomaly
- Analyze every comment to detect anomaly
- Remove anomalies to avoid unnecessary fights and discussions
- Give scope to the positive discussions only

System Architecture

The prime focus of this system is to analyze the post to their tweets on social network. We are applying our system on social media like Twitter, Twitter have some current trends with his hash tags and current topics. And we will use those current topics and hashtags to detect trending topics which will have anomalies. Using this real-time diffusion of information is going on. So, in that so many abnormal discussions on microblogs, this thing are trending on the social network and be able to monitor their evolution and find related anomalies. So, in this system we proposed a system RING (real-time emerging anomaly monitoring system over microblogs text streams).

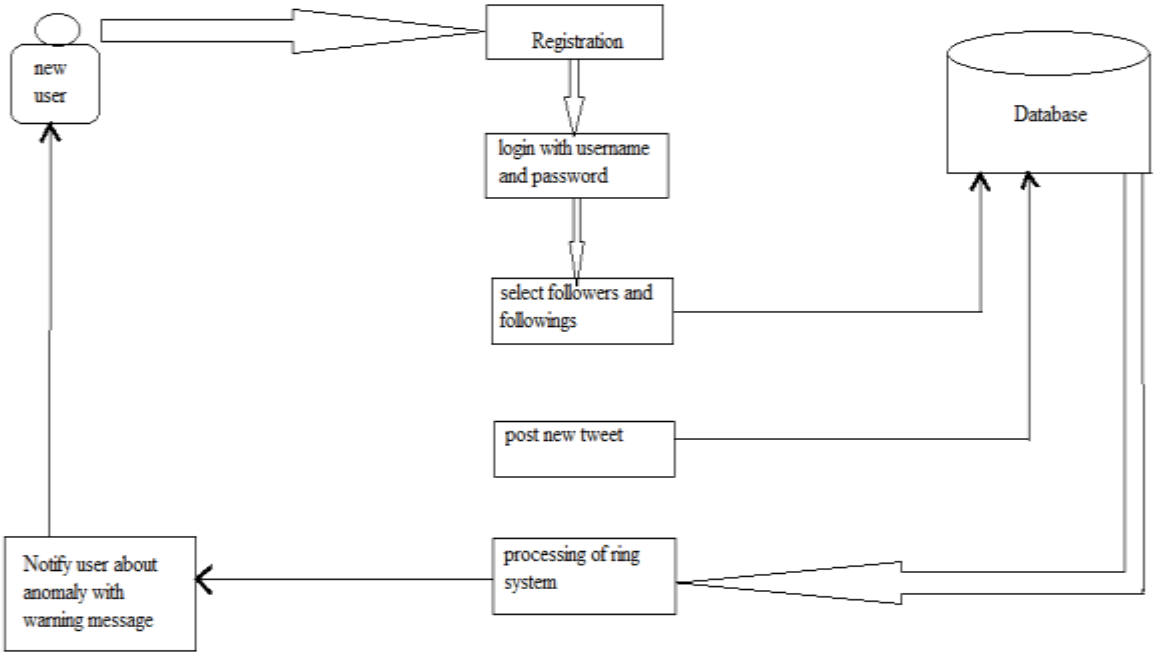

Fig 1: System Architecture

Ring System Architecture

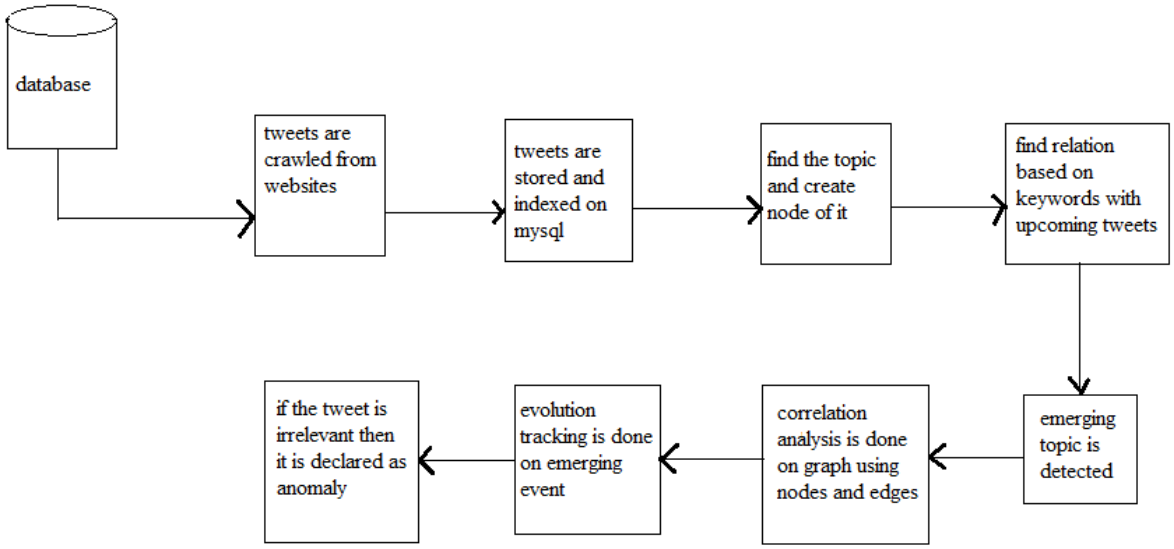

Fig 2: System RING Architecture

Related Works

[1] Earthquake shakes twitter users: real-time event detection by social sensors

Year: 2010

Author Name:

- Takeshi Sakaki
- Makoto Okazaki
- Yutaka Matsuo

Description:

Real time interaction of event such as earthquake is investigated in this system. Probabilistic spatiotemporal model is used to find the center and trajectory of event location. For location estimation each twitter user is considered as sensor and each tweet is assumed as sensor information. Support vector machine (SVM) is used to classify the tweets into positive and negative class and it is further used as training set. Particle filter algorithm and Sequential Importance Sampling (SIS) algorithm are used.

Limitations:

Only disastrous event is detected through this system. If the required numbers of tweets are not coming then event will not be detected promptly. This system fails to handle multiple events at the same time.

[2]Scalable Distributed Event Detection for Twitter.

Year: 2013

Author Name:

- Richard McCreadie
- Craig Macdonald
- IadhOunis
- Miles Osborne and SasaPetrovic

Description:

Novel lexical key partitioning strategy is used to distribute the computational costs of processing single document across multiple machines. Locality Sensitive Hashing (LSH) algorithm is used to approximate the distance to the closest document quickly. Buffering and aggregation of messages is done at the representation phase. Storm topology is used to process high-volume data streams in which each bolt maintains input and output message queues. Bolts in storm topology maintain clusters of documents based upon the closest document found.

Limitations:

Accuracy of the system gets affected if the LSH algorithm used in this system fails to identify the location of the document.

[3] Emerging Topic Detection for Organizations from Micro blog

Year: 2013

Author Name:

- Yan chen
- HadiAmiri
- Zhoujun Li
- Tat-Seng Chua

Description:

To classify relevant and irrelevant tweets from microblogs binary SVM classifier is used. Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) are the probabilistic methods used to build topic models of static text. Streaming based

Fixed Keyword Crawler and Known Accounts Crawler are used. Incremental clustering framework is used to detect new topics and temporal features to help in promptly detecting hot emerging topics.

Limitations:

This system is unable to detect hot or emerging topic in real time. Crawlers used for streaming data have limitations for finding relevant data with the stored keyword.

[4] Topic Sketch: Real-time Bursty Topic Detection from Twitter

Year : 2012

Author Name:

- Wei Xie
- Feida Zhu
- Jing Jiang
- Ee-Peng Lim
- Ke Wang

Description:

This system introduces detection of Real-Time bursty topic. Only topics which triggers large amount of tweets within short time are handled by this method. Topic Sketch is a Novel Sketch –Based topic model which is use to achieve real time event detection. Topic Sketch is based on the two main techniques.

1. Sketch based topic model.
2. Hashing Based Dimension Reduction Technique.

Limitations:

Only real Time bursty events can detect in this system. It takes more time for detection of bursty topic.

[5] Sub-Event Detection from Tweets

Year: 2017

Author Name:

- James Allan
- Ron Papka

- Victor Lavrenko

Description:

This System looks at existing event detection methods that detect events from social media data. We then look into community evolution in graphs to help solve the problem of sub-event detection. Everyone assume that each tweet represents a single topic, but actually a single topic is made up of many keywords from multiple tweets. Each topic represents an event which contains the words that describe the event. The sub e-vent detection approach can be extended to include the events that disappear and reappear after some time

Limitations:

Static graphs used for the detection of related events between two communities cannot be drawn for dynamic or unstable communities.

Tools Used:

- *Hardware Requirements*

  1-2 Machines (PC'S or Laptops)

- *Software Requirements*

I. Operating System - Windows 7/8
II. Application Server - Tomcat 5.0/6.X
III. Front End - HTML, JSP
IV. Scripts - JavaScript.
V. Server side Script - Java Server Pages.
VI. Database - MySQL 5.0
VII. Database Connectivity - JDBC

Technique Used:

Algorithm

1. Naive Bayes:

Naive Bayes is applied to differentiate news, ads and wisdom words among detected trending events, where the latter two classes are major types of spams on Weibo. The classifier is trained with manually labeled data depends on features of content, users and temporal information.

Formula:

$$P(c|x) = P(x|c)P(c)/P(x)$$

- $P(c|x)$ is the posterior probability of *class* (*target*) given *predictor* (*attribute*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

2. Sub event Hierarchy construction:

- Input: Tweets
- Output: Get anomaly list.
- 1. User posts any tweets related to their trending topic.
- 2. Find the anomaly things in current trends
- 3. And then generate the graph of that anomaly things
- 4. Get the list of those anomaly things.

3. Graph model:

It take input as a tweets and then find emerging topic using keyword create node of graph find relations of incoming trends based on keywords and then consider it as anomaly.

Future Scope:

This system can be used for social sites like Facebook, instagram also. People who spread rumors or do negative type of comments about any sensational topics can be recognized and for some extents they can be controlled by the admin.

Conclusion

RING system integrates our efforts on both emerging event monitoring research and system research. After detection of anomalies present in the tweets, it is conveyed to the user. User get warning message for the remove or edit your tweet, and if it is not done then admin can delete the tweet. After admin action user gets the mail also.

Acknowledgement

We take this opportunity to thank our project guide **Prof. A.D.Khairkar** and Head of the Department **Prof.Dr.D.A.Godse** for their valuable guidance and for providing all the necessary facilities, which were indispensable in the completion of this project report. We are also thankful to all the staff members of the Department of InformationTechnology of Bharati Vidyapeeth's College of Engineering for Women, Pune for their valuable time, support, comments, suggestions and persuasion. We would also like to thank then institute for providing the required facilities, Internet access and important books.

PRACHI BOBADE

GEETANJALI BOKE

PRAJAKTA KSHIRSAGAR
PRANITA MUSANDE

References

[1] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake shakes twitter users: real-time event detection by social sensors,” in WWW, 2010.

[2] R. McCreadie, C. Macdonald, I. Ounis, M. Osborne, and S. Petrovic, “Scalable distributed event detection for twitter,” in IEEE BigData, 2013.

[3] Y. Chen, H. Amiri, Z. Li, and T.-S.Chua, “Emerging topic detection for organizations from micro blogs,” in SIGIR, 2013.

[4] W. Xie, F. Zhu, J. Jiang, E.-P. Lim, and K.Wang, “Topicsketch: Real-time bursty topic detection from twitter,” in ICDM, 2013

[5]SatyaKatragadda,RyanBenton,VijayRaghavan,"Sub-event detection from twitts"in IEEE,2017