ISSN 2395-1621

Identification of Anomalous User Behavior in Twitter

Anjali Changale, Anjali Ghutke, SonaliTakke, ShitalPeddawad

P.E.S Modern College of Engineering, Pune, India

,

ABSTRACT

The internet is one of the biggest blessing to man given by technology. Modern era people can't imagine the life without internet; everyone wants to connect with each other through internet via social media at all the time. A major part of modern world people use an online communication system, such as email and social media sites (e.g. Twitter, Facebook and LinkedIn) for entertainment and business. They generate lots of digital data for various users' activity; introduce a system to understand user communication behavior. In online communication system users' who have anomalous behavior is the potential threat to the society. Here we are proposing a system for identification of anomalous user behavior in Twitter. We are identifying the impact of user in twitter through their tweets, profile, followers and posting to know what they think. We propose a visual analysis system for detecting,summarizing and interpreting via the unsupervised learning model, visualizes the behavior of suspicious users.

Keywords: Anomalous behavior, unsupervised learning model.

ARTICLE INFO

Article History

Received: 27th May 2017

Received in revised form :

27th May 2017

Accepted: 31th May 2017

Published online :

1th June 2017

I. INTRODUCTION

The size of a database in twitter has increased rapidly day-to- day, Due to which the anomalies are also increasing. Anomaly detection is a problem of finding patterns in data that do not confirm to expected behavior. These nonconforming patterns are often called as anomalies. The real-life or interesting relevance of anomalies is a key feature of anomaly detection. We identify users' communication behavior by considering following features:

Behavior features: This features identifying user's role based on their posting and re-posting behaviors. For Example, it determines how many tweets are made by user in each 5 days.

Context features: This feature categories based on topical keywords. Contextual analysis is nothing but sentiment analysis. It finds number of positive and negative tweets.

Interaction features: This feature describe user based on their communication pattern, how users communicate with others. In this project, interaction feature is implemented by using retweet count of each tweet.

User profile features: In that we check user profile information. e.g.,Default profile,Default profile image, Geo Enabled, User verified,etc.

Temporal Features. In this category, posting, replying, receiving, interval frequency entropy,
measure the regularity of certain types of user behaviors.

Entropy count means the number of words in all tweets are determined.

I. SYSTEM

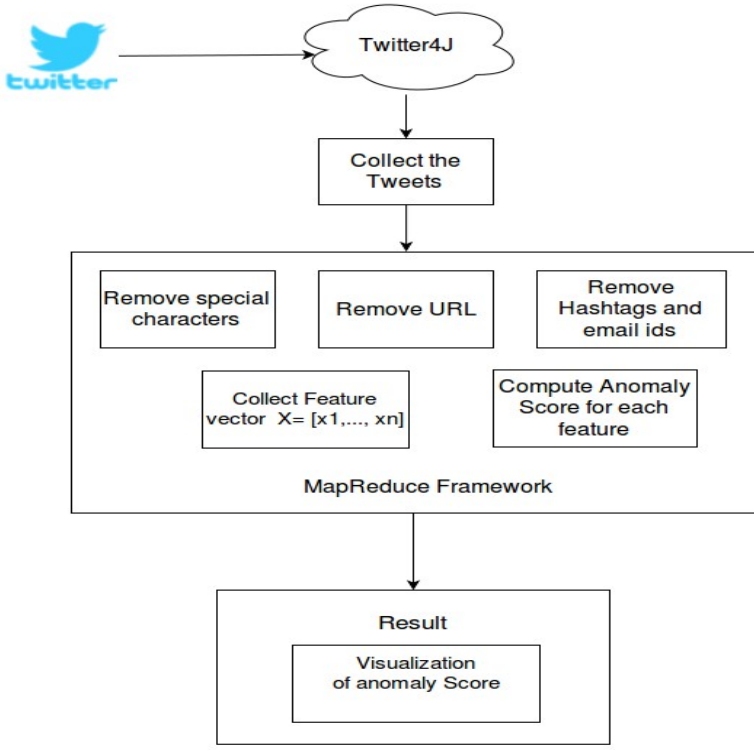
We are designing a novel visual analyze system by unsupervised learning model and visualization technique to detect anomalous user behavior.

Different social media sites have different uses, strengths and advantages. Twitter could be called a 'real-time social networking' site, a place for sharing information as it happens and for connecting with others in real-time, often resulting in lasting friendships and contacts. A lot of people communicate via social networking like twitter. In the unsupervised anomaly detection, we are given an input as a set of user data where it's unlabeled data (or noisy data). The goal is to identifying anomalous user behavior. We train noisy data and apply a traditional anomalous detection algorithm over the data.

In unsupervised machine learning anomalous detection has many advantages over supervised machine learning anomaly detection. The main advantage of unsupervised machine learning that they don't need a labeled data. The unsupervised anomaly detection algorithm does not need train dataset. In addition, unsupervised anomaly detection algorithm can use to analyze historical data.

The major advantage of our system, it is flexibility. We can apply this system for different kind of data (e.g., email, Facebook, LinkedIn). We worked on TLOF algorithms for detecting anomalies. This algorithm is very efficient and can deal with high dimensional data.

II. SYSTEM ARCHITECTURE



III. BRIEF DESCRIPTION OF SYSTEM:

1. Data collection

Input : user Screen name, OAuthConsumerKey, OAuth Consumer Secret key, OAuthAccessToken, OAuthAccessTokenSecret.
Output : list of tweets
Processing :

1) Get Connection from twitter account by using api keys.
2) Extract one by one twitter pages by using pagination.
3) Fetching tweets from twitter pages.
4) Analyze the tweets and take only English tweets.

2. Stopword Removal

Input: list of tweet
Output : list of preprocessed tweet
Processing :

1) Remove white spaces from tweet
2) Remove webpages link from tweet
3) Remove special character from tweet
4) Tweet data split by using spaces and matched stop word remove from tweet

5) Tweet contains hashtag and email. In this process we remove hashtag and emails

3. Feature Extraction :

Input : list of tweet with given time
Output : list of feature
Processing :

1) Take sentiment score with given time.
2) Take retweet count with given time
3) Take tweet count with given time
4) Tweet split into word by using string split method and get the entropy count of tweet with given time.

4. TLOF

Input: feature vector X = [x1,x2,...,xT] , xt is a feature vector describing theUser's behaviors observed at time t ∈ [1, 2 , … T]
Output: - anomaly score s(X)

Algorithm :
TLOF :-Time Adaptive Local Outlier Factor
1) First find the K-nearest neighbors of each point in feature vector xt.
2) For certain points, calculate the reach-distance
3) Then we calculate the local reachability density of each point
4) Calculate LOF Scores of X
5) Z1 : Difference between current LOF and Average LOF
6) Z2 : probability of X, using mean and standard deviation of distribution
7) Finally, from Z1 and Z2 we get anomaly score

5. Plot Graph

Input: Anomaly Score
Output: Generate graph
Processing :

1) Each feature has its own anomaly score which is stored in session.
2) Gather anomaly score from session and store that data into JavaScript array.
3) Java script plugin take input as array and generate graph.

IV. RESULT :

Twitter User Name	Behavioral score	Content score	Interactio score
MelissaBachman	0.467496	0.2634445	-0.04978
narendramodi	0.2431492	0.823200	0.794413
kamaalrkhan	0.79723399	-0.59390816	0.547742

V. CONCLUSION :

In this paper, we design a system for detecting anomalous user behavior on twitter and assign anomaly score to that user. Finally show these anomaly score by using graphical representation. The data collection process will introduce us to the Java Twitter4J API. This project will help us to gain knowledge about installation and configuration of the Hadoop and Hadoop cluster.

VI. REFERENCES:

[1] Nan Cao, Conglei Shi, Sabrina Lin, Jie Lu, Yu-Ru Lin, Ching-Yung Lin ,TargetVue: Visual Analysis of Anomalous User Behaviors inOnline Communication Systems, IEEE Transactions on Visualization and Computer Graphics, Vol.22, No.1, January 2016.
[2] M. M. Breunig, H.-P.Kriegel, R. T. Ng, and J. Sander.Lof: identifying density-based local outliers. In ACM sigmod record, volume 29, pages 93–104. ACM, 2000.media. Journal of Visualization,2016.
[3] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. ACM Computing Surveys, 41(3):15, 2009.
[4]Hadoop: The Definitive Guide Book by John White.
[5] M. Pennacchiotti and A.-M.Popescu.A machine learning approach to twitter user classification.Proceedings of the AAAI International Confer-ence on Weblogs and Social Media, 11:281–288, 2011.
[6]Reza Hassanzadeh. Anomaly detection in online social networks: Using data mining Techniques,2014
[7]Leonid Kalinichenk,IvanShanin,IliaTaraban, Methods for Anomaly Detection: a Survey.

[7].Intelligent Residential Security Alarm and Remote Control System Based On Single Chip Computer 978-1-4244-1718-6/08/ 2008 IEEE

[8].A Dynamic Programming Algorithm for Leveraging Probabilistic Detection of Energy Theft in Smart Home 10.1109/TETC.2015.2484841, IEEE Transactions

[9].Vibration analysis for fouling detection using hammer impact test and finite element Simulation. 1-4244-1541-1/08/ C 2008 IEEE.

[10]. 10.Continuous and Damped Vibration Detection Based on Fiber Diversity Detection Sensor by Rayleigh Back-scattering. Journal of light wave technology, vol.26, no. 7, April 1, 2008.

[11] https://en.wikipedia.org/wiki/Haar-like_features.

