International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 17
November2016

Analysis of Twitter Trending Topics via LinkAnomaly Detection

Mr.P. Harikanth
M.Tech, Computer Science &Engineering
Department of CSE
Vaagdevi College Of Engineering, Warangal
Mr .Mahipal Reddy Pulyala
Assistant professor, Department of CSE
Vaagdevi College Of Engineering, Warangal

Abstract: Most of us be aware ofrising themes signaled by way of social disorderswith these sites. In exact, we pay attention tothe mentions of users-links in between end userswhich can be generated dynamically (intentionallyas good as unintentionally) by way of responds,retweets, additionally to mentions. The probabilitymodel of the mentioning behavior of a social enduser captures both the number of mentions per postas well as the frequency of mentionee. We thenaggregated the paradox rankings from numerous endusers and we show that we may realize thetrending themes easily utilising the reply/point outrelationships in twitter posts. In this paper, themixture of point out-anomaly model with termfrequency ways is proposed. We illustrate ourprocess on the datasets received through twitter.

Keywords-anomaly-detection, social network, change-point analysis, burst detection.

I. INTRODUCTION

Now-a-days social networking websites are fitting the primarycommunication media among participants and companies.Social network are internet-based software, social networkparticipants to assemble a public or semi-public profiles,articulate record of other users with whom they share aconnection, and traverse and view their record of connections andthese made by others in the process. Folks share theirindividual expertise, photographs, videos, URLs, suggestions on thesesites. People are living in contact with their family, buddies andcolleague. However leakage of individual data createssecurity predicament, cyber bullying, spreading the hatredmessages and so forth. Malicious users may intent many extreme disorderslike De-Anonymizationattack, neighbourhood attack, profilecloning, social phishing, junk mail attacks and lots of extra. Consequentlydevelopment of risk-free anomaly detection in socialnetworking websites is extremely essential [4]. In this feel,mention way like a language, this mention contain quantityof phrases equal to the quantity of person in a social media.We are fascinated about detecting tendencies themes from socialcommunity streams situated on citing behaviour of users. Ournormal assumption is that a new (rising) topic is somethingpersons believe like discussing about, commenting about, orforwarding the data extra to their friends [2]. Earlymethods for subject notice have as a rule been concernedwith the frequencies of words. In this procedure, first, the socialnetwork is shown in a graph, after which similarity amongst clients,this graph is split into smaller communities [7].

Afterwards, all the similar profiles to the actual profile aregathered, then strength of relationship is calculated, and not morestrength of relationship will likely be tested through mutual friendapproach. In this study, in order to assess proposed method,

allsteps are utilized on a dataset of facebook, Twitter, Google+,and in the end this work is when put next with two earlier works by way ofapplying them on the dataset. Together with chance modelcan capture the usual mentioning behaviour of a user, thislikelihood model consists of each the quantity of mentions perpublish and the frequency of users happens in the mentions. Thenprobability model is used to measure the paradox of futureuser behaviour [2]. Using this model, quantitatively measurethe novelty or possible have an impact on of a put up mirrored within theciting behaviour of the user. Here we mean through mentionslinks to different users of the equal social network within the form ofmessage-to, reply-to, retweet-of, or explicitly within the text. Onepost may contain a number of mentions. Some users couldcomprise mentions of their posts not often; different users is alsobringing up their friends at all times. Some clients (likecelebrities) could acquire mentions every minute; for others,being acknowledged probably a rare party. In this experience,point out is sort of a language with the number of words equal tothe number of users in a social network. We're interested indetecting rising themes from social community streams basedon monitoring the mentioning behaviour of users.

II. RELATED WORKS

A new (trending) topics is something people feellike discussing, commenting the information furtherto their friends. Conventional key-based approachesfor topic detection have mainly focus on frequenciesof (textual) words [4].A key-based approach could suffer from theambiguity caused by synonyms. It may also requirecomplicated pre-processing (e.g., segmentation). Itcannot be applied when the contents of the messagesare mostly no textual information. Another way,words formed by mentions are unique, require littlepre-processing to obtain.

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 todiscover topic trends and analyse their dynamics in real-time.For example, it is desired in the social stream area to grasp anew trend of topics in online user claims every day and totrack a new topic as soon as it emerges. A topic is here definedas a seminal event or activity detection and detect of topicshave been studied in the area of topic detection and tracking.In this context, the main task is to either a new document intoone of the known topics or to detect the none of the knowncategories. Alternative, temporal structure of topics that havebeen 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 useof word content of the documents, but not the social content ofthe documents. The social content i.e. link has been utilized inthe study of citation networks. However, citation networks areoften analyzed in a stationary setting. The novelty of thecurrent paper lies in focusing on the social content of thedocuments (posts) and in combining this with a change-pointanalysis [2].

III. THE PROPOSED APPROACH

In this paper, we propose the probabilitymodel which is able to capture average mentioningbehavior of an end user, which consist of both thequantity of mentions for every post and thefrequency of users happening in the mentions. Andthen, we are able to use this model to calculate theanomaly of future user

behavior.In Fig.1, the entire flow of the proposedmethod is verified. Within the followingsubsections, we've explained each single step ofthe total float of the proposed method. Right here, weconsider that in a sequential manner, the data comesfrom a social network provider via some API. Forevery single new put up, we make use of samplesfrom the prior time interval associated with size Tin regards to the corresponding user for coachingthe point out model. We determine an anomalyranking for every publish utilizing the discoveredprobability distribution. This rating can be thenaggregated round users and extra providedKleinberg's burst detection method along with TF-IDF procedure.

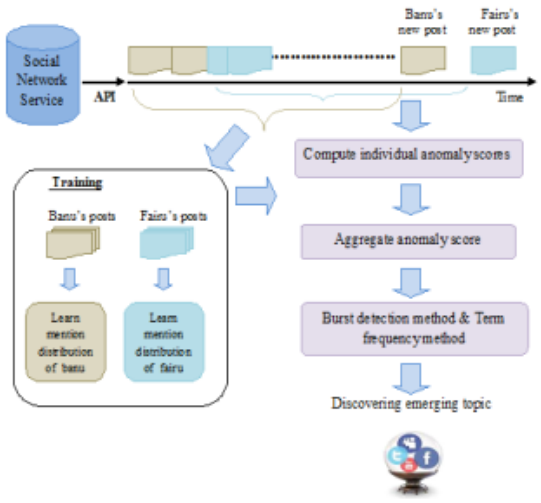

Fig.1. Overall flow of the proposed method

A single post x within a Twitter is characterizedby means of the number of mentions k it has, aswell as the set V of names (IDs) of the mentionees.The deviation of the user behavior is calculatedbased on the anomaly score. The anomaly score iscalculated for each post in the social data. Forcalculating the anomaly the joint probabilitydistribution and geometric distribution is used.

A. Burst detection method

Next, we use a burst-detection process toassess the frequency of a subject in a giveninterval. The procedure has been proposed in [3]. Thisprocess detects whether the interval of the arrivalmessages is denser than that in a typical conditionthrough evaluation with other file streams.The burst-detection method defines aprobabilistic automaton (A) consisting of twostates:

(1) at any time when A is in state q_0, messagesarrive at a slower cost.

(2) whenever A is in stateq_1, messages arrive at a turbo cost. ‘

The interval (T)is defined because the interval between the appearance of thefirst message and that of the final message, n + 1. Ifthe arrival time is random, a hole time x betweenthe messages i and that i + 1 follows an exponentialdistribution.

According to Poisson distribution, instate q_0, the probability of arrival of the nextmessage at interval x is $f_0(x) = \alpha_0 e^{-\alpha_0 x}$, where $\alpha_0 = n/T$. In state q_1, the gap time is shorter than in stateq0. Consequently, the probability of interval xbetween any two consecutive message is $f1(x) = \alpha_1 e^{-\alpha_1 x}$, where $\alpha_1 > \alpha_0$.In addition, we determine a given number of nmessages with a particular arrival time as innerarrival gaps $x = (x_1, x_2 \ldots\ldots, x_n)$, where $x_i > 0$.Similarly, we set the conditional probability of astate sequence $q = (q_{i1}, q_{i2} \ldots\ldots, q_{in})$. Each statesequence q derives a density function f oversequences of gaps, which is represented by thefollowing formula.

$$f_q(x_1 \ldots\ldots, x_n) = \prod_{t=1}^{n} f_{i_t}(x_t)$$

B. Tf-idf method

Tf-idf is short for term frequency-inversedocument frequency, and the tf-idf weight is astatistical measure helpful to estimate how essentiala word is to a document in a collection or corpus

tf=idf(w,t,d)= tf(w,t)*idf(t,d)…… (1)

$$idf = (t,d) = log \frac{T}{|\{t \epsilon d : w \epsilon t\}|} \quad \text{……..(2)}$$

Equation of tf-idf for word w is defined in (1).In (1) t represents one tweet and d stands for adocument, which would be the corpus of tweets inour work. tf(w,t) is the term-frequency, idf(t,d) isdefined in(2), in which T denotes the total numberof tweets in document d.

C. Clustering procedure for proposedmethod

1) Select topic ti Є TOPIC, where i=1…..n

2) For each topic ti, sort the link-messages,images and videos in ascending order oftheir score.

3) Score= Anomaly score + score generatedby retweets, mentions and forwards, forthe categories of messages, link-messages,images and videos.

4) We combine these score to generate onesingle list called EMERGING TOPIC list.

The proposed method may be useful to the boththe cases where topics are concerned withinformation like texts and the non-textualinformation like images, video, and so on.

IV. CONCLUSION

We have endorsed the latest method todiscover the emerging themes in a twitter. Thechance model does no longer depend on the textualcontents of twitter posts; it usually is valuable to useto the scenario anyplace themes are involved withexpertise as opposed to texts, like graphics, video,etc. The TF-IDF procedure is used when thecontents of the posts includes understanding liketexts. In my opinion point out-anomaly model(likelihood model) and time period-frequency-basedmethods isn't going to immediately notify whatprecisely the paradox is. By means of the blend ofpoint out-anomaly model with the TF-IDF methodwill get data each from the totalefficiency of the probability model and theintuitiveness of the text-frequency-based method.

REFERENCES

[1] J. Allan et al., “Topic Detection and TrackingPilot Study: Final Report,” Proc. DARPABroadcast News Transcription and UnderstandingWorkshop, 1998.

[2] Toshimitsu Takahashi, Ryota Tomioka, and Kenji Yamanishi,“Discovering Emerging Topics in Social Streams via Link-AnomalyDetection,” IEEE Transactions on Knowledge and Data Engineering, Vol.26,No. 1, January 2013.

[3] Shenghua Bao, Shengliang Xu, Li Zhang, Rong Yan, Zhong Su, DingyiHan, and Yong Yu, “Mining Social Emotions from Affective Text,” IEEETransactions on Knowledge and Data Engineering Vol. 24, No. 9, September2012.

[3] J. Kleinberg, “Bursty and Hierarchical Structurein Streams,” Data Mining Knowledge Discovery,vol. 7, no. 4, pp. 373-397, 2003

[4] D. Boyd and N. B. Ellison, “Social Network Sites: Definition, History, andScholarship,” Journal Computer-Mediated Communication, vol.13, no. 1-2,Nov. 20072002.

[5] T. Takahashi, R. Tomioka, and K. Yamanishi,"Discovering emerging topics in social streams vialink anomaly detection," arXiv:1110.2899v1[stat.ML], Tech. Rep., 2011.

[6] Amrutha Bennya, Mintu Philipb, "Keywordbased Tweet Extraction and Detection of RelatedTopics", International conference on informationand communication technologies, 2014.

[7] Oliver Brdiczka, Juan Liu, Bob Price, Jianqiang Shen, Akshay Patil,Richard Chow, Eugene Bart, Nicolas Ducheneaut, "Proactive Insider ThreatDetection through Graph Learning and Psychological Context," IEEE CSSecurity and Privacy Workshops,24-25 May 2012.

[8] Nisheeth Shrivastava, Anirban Majumder, Rajeev Rastogi, "Mining(Social) Network Graphs to Detect Random Link Attacks," IEEE 24thInternational conference on data engineering, pp.486-495,7-12 April 2008

[9] Jiong Zhang, Mohammad Zulkernine, and Anwar Haque, "RandomForests-Based Network Intrusion detection system" IEEE Transactions onSystems, Man, and Cybernetic, Vol. 38, No. 5, September 2008

[10] Kush R. Varshney, "Bounded Confidence Opinion Dynamics in a SocialNetwork of Bayesian Decision Makers," IEEE Journal of Selected Topics InSignal Processing, Vol. 8, No. 4, August 2014

[11] Yang Li, Bin-Xing Fang, "A Lightweight Online Network AnomalyDetection Scheme Based on Data Mining Methods," International Conferenceon Network Protocols, pp.340-341, 16-19 Oct. 2007

[12] Alexander Y. Liu and Dung N. Lam, "Using Consensus Clustering forMulti-view Anomaly Detection," IEEE CS Security and Privacy Work,pp.117-124, 24-25 May 2012.

Author's Profile

Mahipal Reddy Pulyala working as Assistant Professor, Department of CSE in Vaagdevi college Of Engineering , He received B.Sc (M.P.E) Degree from Vaagdevi Degree & PG College (K.U). He Completed M.C.A From Vaagdevi College Of Engineering(J.N.T.U). He Completed Mtech (C.S.E) ,From Vaagdevi College Of Engineering, (J.N.T.U) UGC Autonomous, Bollikunta, Warangal, Telengana State, India.

Mr . P. HARIKANTH pursuing M.Tech degree in Computer Science & Engineering in CSE Department in Vagdevi college Of Enginnering , UGC Autonomous, Bollikunta, Warangal, Telengana State, India.