ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Using Link-Anomaly Detection to Discover Emerging Topics in Social Streams

Pattan Imran Khan
PG Scholar,
Department of CSE,
Sri Chundi Ranganayakulu Engineering College,
Chilakaluripet, Guntur, AP, India.

Karpurapu Sudhakara Babu
Assistant Professor,
Department of CSE,
Sri Chundi Ranganayakulu Engineering College,
Chilakaluripet, Guntur, AP, India.

Abstract:

Detection of emerging topics is now receiving renewed interest motivated by the rapid growth of social networks. Conventional term-frequency-based approaches may not be appropriate in this context, because the information exchanged in social¬network posts include not only text but also images, URLs, and videos. We focus on emergence of topics signaled by social aspects of theses networks. Specifically, we focus on mentions of users - links between users that are generated dynamically (intentionally or unintentionally) through replies, mentions, and retweets. We propose a probability model of the mentioning behaviour of a social network user, and propose to detect the emergence of a new topic from the anomalies measured through the model. Aggregating anomaly scores from hundreds of users, we show that we can detect emerging topics only based on the reply/mention relationships in social network posts. We demonstrate our technique in several real data sets we gathered from Twitter. The experiments show that the proposed mention-anomaly-based approaches can detect new topics at least as early as text-anomaly-based approaches, and in some cases much earlier when the topic is poorly identified by the textual contents in posts.

I.INTRODUCTION:

Communication over social networks, such as Facebook and Twitter, is increasing its importance in our daily life. Since the information exchanged over social networks are not only texts but also URLs, images, and videos, they are challenging test beds for the study of data mining. In particular, we are interested in the problem of detecting emerging topics from social streams, which can be used to create automated "breaking news", or discover hidden market needs or underground political movements. Compared to conventional media, social media are able to capture the earliest, unedited voice of ordinary people.

Therefore, the challenge is to detect the emergence of a topic as early as possible at a moderate number of false positives. A term-frequency-based method could undergo from the ambiguity caused by synonyms or homonyms. It may also require complex preprocessing (e.g., segmentation) depending on the target language and it cannot be applied when the contents of the messages are mostly no textualinformation. A probability model is proposed that can capture the normal mentioning behavior of the user, i.e. number of mentions per post and the occurring of the mentions. By the use of probability model, the novelty or possible impact of the post can measure and will aggregate the anomaly scores for it. Apply a recently proposed change point detection technique based on the sequential discounting normalized maximum-likelihood coding [3], [4], [5].In proposing system link anomaly model is combined with text based approaches and also with a word based approach to give a good performance of the mentioned model.The layout of the paper is as follows. In section II, address the above mentioned techniques and also give a brief on the literature being reviewed for the same. Section III, presents issues in current environment. Section IV, describes proposed system, method applied and there algorithm. Section V gives the conclusion and finally provides references and about authors.

II. RELATED WORK:

In this paper [1] user discovers the emerging topics from the social networks. As the information exchanged in the social networks post includes not only the text, but also images, URLs and video therefore conventional-term frequency- based approaches may not be appropriate in this context.Based on the responds from hundreds of users in social networks post is used to detect the emergence of new topics. In this paper probability model is proposed to capture a number of mentions per post and the frequency of users occurring in the mention. The disadvantage is all the analysis presented in the paper was conducted offline.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

In this paper [2] model selection in Gaussian linear regression use of the normalized maximum likelihood which poses troubling because the normalization coefficient is not finite. The methodology is comprehensive and discussed two particular cases, they are rhomboidal and the ellipsoidal constraints. By rigorous analysis eight NML based criteria are tested and yields a new NML based formulas. The disadvantage is normalized coefficient is not finite. In this paper [3] Autoregressive modeling yields high resolution power spectral density valuation, therefore it is widely used for stationary time series. The information theoretic criteria (ITC) have increased constantly for selecting the order of autoregressive (AR) models. The Author has modified the predictive density criterion (PDC) and sequentially normalized maximum likelihood (SNML) criterion to be compatible with the forgetting factor least squares algorithm.

In this paper [4] Author has thoroughly studied the predictive least squares (PLS) principle for model selection in perspective of regression model and autoregressive. The aim of the model selection is not used to pick the correct model, but it is used to minimize future prediction errors. SNLS is a best method with a very small margin. In [5] author has monitored the occurrence of topics in a stream of events. There are several algorithms, produces very different results to monitor the occurrence of topics. Kleinberg's burst model and Shasha's burst model are used to monitor. It works well for tracking topic bursts of MeSH terms in the bioscientific Literature; it can also be used for forecasting oncoming bursts and momentum based topic dynamics burst model have a significant advantage. The disadvantage is Hierarchical structure deserves greater attention on burst.

In this paper [6] Normalization produces normalized maximum likelihood (NML) distribution. Sequential normalized maximum likelihood (SNML) is easier to compute and include a random process. SNLS is the best method, with the exception of the smallest sample sizes. AIC, BIC, PLS, SNLS methods is used to estimate the order of an AR Model.BIC is known to have a tendency to underestimate rather than overestimate the order. Similarly, it is not too surprising that AIC, which a priori favors more complex models than the other criteria, wins for the smallest sample size.The problem was considered relating to groups of data where each study within a group is a draw from a combination model. Yee Whye Tech, Michael I, Jordan, Matthew J, Beal, and David M. Blei [7] has represents hierarchical Dirichlet process in the term of the stick breaking process that gives random measures explicitly, a chinese restaurant process that is referred as "Chinese restaurant franchise" describes a representation of marginal's in terms of an urn model and representation of the process in terms of an formulation of three MCMC sampling schemes for posterior inference. In this method to the problem sharing clusters among multiple related groupsis a nonparametric Bayesian approach. Andreas Krause, Jure Leskovec, Carlos Guestrin [8] presented a unified model; it is traditionally viewed as two tasks: Data association and intensity tracking of multiple topics over time. To solve the problem, this approach combines an extension of the factorial Hidden Markov model for topic intensity tracking with exponential order statistics for implicit data association. This approach improves the accuracy of intensity tracking, classification, and also detects correct topic intensities even with 30% topic noise.

This paper [9] is concerned with the problem of detecting outliers and change points from time series. Unified frame worked was used to the deal the problem. The score for the data was calculated in the deviation from the learned model. Change point detection was used to reduce the issue of detecting outliers in that time series. The advantage of this approach is Change points from nonstationary are much more efficient than conventional methods. It would be challenging problem to design of an algorithm to detect variance decrease change point. In this paper [10] Temporal Text Mining (TTM) was concerned with discovering temporal patterns in text information together over time. Since most text information bears some time stamps. The advantage is the proposed technique is based on hidden Markov models for analyzing the life cycle of each theme. This process would first determine the globally attractive themes and then compute the strength of a theme in each time period. This agrees us to not only see the trends of strength variations of themes, but also compares the relative strengths of different themes over time. It is flat structure of themes and was not considered.

III. ISSUE IN CURRENT ENVIRONMENT:

By rapid growth of social networks, it wasinteresting to discover the emerging topics from the postposted by hundreds of users. A new (emerging) area issomething people feel like discussing, commenting, orforwarding the information further to their friends. To findthe emergence of areas in social networks streams are donevia Link anomaly model.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Probability model is used tocapture both the number of mentions per post and thefrequency of it. It does not rely only on the textual contentsof social network posts and it can also be applied to images,video, audio and so on.On the other hand, the "words"formed by mentions are unique, require little preprocessingto obtain (the information is often separated from thecontents), and are available irrespective of the nature of thecontents.

Since the existing method does not rely on the textual contents of social network posts, it is strong torephrasing and it can be applied to the case where topics areconcerned with information other than texts, such aspictures, video, audio, etc. All the analysis was conductedoffline, and link anomaly model does not instantly tell whatthe anomaly is.

IV. PROPOSED METHOD:

As the social networks are grown it is interesting todiscover the emerging areas. To detect the emergence ofnew topics from the hundreds of users based on the respondsin social network posts. A probability model is proposed formentioning behavior of social networks by the number ofmentions per post and the occurrence of users taking placein the mentions.

The basic assumption is that a newemerging topic is something people feel like chatting,stating or sending the information further to their friends.This paper shows that the proposed approach i.e. by mixtureof word based approach with link anomaly model would dowell to both from the performance of the mention mode andthe intuitiveness of the word based approach.Fig. 1. Architecture DiagramThe overall architecture of the system is shown in Fig.1. Each step in the flow is described.

The data arrive from asocial network service in a sequential manner through someAPI. For each new post, samples are used within the pasttime interval of length for the corresponding user fortraining the mention model. Assigned an anomaly score toeach post based on the learned probability distribution.

Thescore is then aggregated over users and further fed intoSDNML-based change point analysis. And also describeKleinberg's burst-detection technique, which can be usedinstead of the SDNML-based change-point analysis.

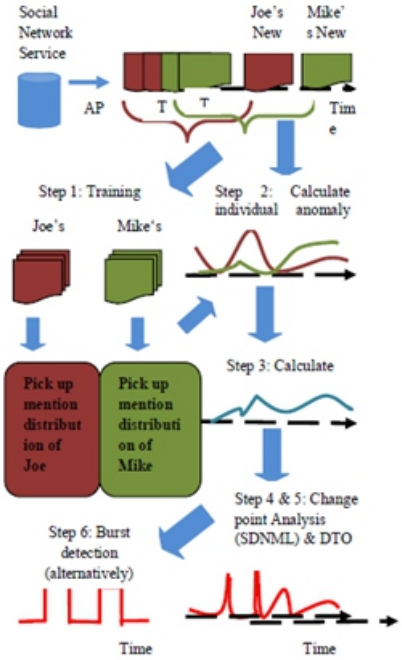

Fig. 1. Architecture Diagram

A. Probability Model:

In this paper probability model is proposed to capture thenormal mentioning actions of the user and to train themodel. There are two types of infinity is taken into theaccount. The first is the number of users, where number ofusers cannot mention hundreds of further users in a post and would like to avoid placing an artificial limit.The second isnumber of post posted by users; Chinese Restaurant Process[7] is used to limit the number of mentionees. We havecalculated predictive distribution with respect to the numberof mentions and also for mentioning user. Link anomalyscore is calculated to both predictive distribution of thenumber of mentions and as well as to the mentionees. Thencombine the anomaly scores from hundreds of users. Foreach user anomaly score is calculated by current post of theuser with past post of the same user. Aggregate is done forall the individual anomaly score. This aggregate is appliedby change point detection through sequentially discoutingnormalized maximum-likelihood (SDNML)coding;dynamicthreshold optimization (DTO) in addition Kleinberg's burstdetection method.

B. ComponentIllustration:

The Detection of Emerging Topics in SocialStreams via Combination of Word Based Approach withLink Anomaly Model is implemented in three differentmodule and they are:

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Social Network Service implementation
Data Collection And Training
Link Anomaly score calculation

1) Social Network service implementation:

Have to create the social network service first andimplement its functionalities such as posts, comments;online users, profiles and all the information will be storedin the repository. Here data has to take into account, thatmeans the post and comments, mentions, retweets that isrequired for the process from the social networking streamthat are used as input. Then collect the post of each andeveryone who are registered with the social network streamand also deal with the comments that are posted against eachpost.

2) Data Collection and Training:

In this component, categorize each individual's postsfirst and then find the mentioning distribution. Next have tofind how many numbers of mentions in a post as well ascomments and also the frequency with which each user ismentioned. There are two types of infinities have to takeinto account here. The first is the number k of usersmentioned in a post. The second type of infinity is thenumber of users one can possibly mention. Calculate thepredictive distribution with the number of mentions. Thencalculate the predictive distribution with number of users.

3) Link Anomaly Score Calculation:

In the third component, the deviation of a user'sbehavior from the normal mentioning behavior modeled inthe previous section is computed here. To compute the anomaly score of a new post x of user u at time t containingk mentions to users V, next step is to compute theprobability with the training set T on u, which is thecollection of posts by user u in the time period. The twoterms are used that can be computed via the predictivedistribution of the number of mentions and the predictivedistribution of the mentionee respectively.

The anomalyscore is computed for each user depending on the currentpost of user u and his/her post. To measure the general trendof user behavior, and propose to aggregate the anomalyscores obtained for posts x1; . . . ; xn . Then have to findhow to detect change points from the sequence ofaggregated anomaly scores and plot the results in a graph.Lastly find the emerging topics in the social streams..

V. CONCLUSION:

In this paper, we have proposed a probabilitymethod to identify of emerging areas which becomesinterest by the fast development of social networks. As theinformation exchanged in social networks post include notonly the text but also images, URLs and video and so onThe fundamental idea of the approach is to focus on thesocial aspect of the posts reflected in the mentioningbehavior of users instead of the textual contents.Probabilitymodel capture the number of mentions per post and theoccurrence of users taking place in the mentionsbycombining the link anomaly model with text and word basedapproach. The change-point detection algorithm viaSDNML and Kleinberg's burst-detection model are used topinpoint the emergence of a topic. This paper gives betterperformance and instantly tells what anomaly is.

REFERENCES:

[1]Toshimitsu Takahashi, RyotaTomioka and Kenji Yamanishi,"Discovering Emerging Topics in Social Streams via Link-Anomaly Detection", IEEE Trans on Knowledge and DataEngineering, Vol 26, No.1, Jan 2014.

[2]CiprianDoruGiurcaneanu,SeyedAlirezaRazavi, Antti Liski,"Variable selection in linear regression: Several approachesbased on normalized maximum likelihood", Signal Processing,Vol. 91, pp.1671-1692, 2011.

[3]CiprianDoruGiurcaneanu, SeyedAlirezaRazavi, "AR orderselection in the case when the model parameters are estimatedby forgetting factor least-squares algorithms", SignalProcessing, Vol. 90, no.2, pp.451-466, 2010.

[4]JormaRissanen, TeemuRoos, Petri Myllymäki, "Modelselection by sequentially normalized least squares ",J. Multivariate Analysis, Vol.101, No.4,pp.839-349, 2010.

[5]Dan He, and D. Stott Parker, "Topic Dynamics: An AlternativeModel of 'Bursts' in Streams of Topics", Proc. 16th ACMSIGKDD Int'l Conf. Knowledge Discovery and Data Mining,pp. 443-452, 2010.

[6]TeemuRoos and JormaRissanen, "On Sequentially NormalizedMaximum Likelihood Models ", Proc. Workshop InformationTheoretic Methods in Science and Eng.,2008.