International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 3 1338 - 1343

Detecting Emerging Areas in Social Streams

K. Ramya[1]
PG Student,Dept of CSE
Bharath University
Chennai, India
toramya269@gmail.com

S. Brintha Rajakumari[2]
Assistant Professor,Dept of CSE
Bharath University
Chennai, India
brintha.ramesh@gmail.com

Dr.T.Nalini[3]
Professor,Dept of CSE
Bharath University
Chennai, India
nalinicha2002@gmail.com

Abstract- Detecting the emerging areas becomes interest by the fast development of social networks. As the information exchanged in social networks post include not only the text but also images, URLs and video therefore conventional-term-frequency-based approaches may not be appropriate in this context. Emergence of areas is focused by social aspects of these networks. To detect the emergence of new areas from the hundreds of users based on the responds in social network posts. A probability model is proposed for mentioning behavior of social networks by the number of mentions per post and the occurrence of users taking place in the mentions. The basic assumption is that a new emerging topic is something people feel like discussing, stating or forwarding the data further to their friends. In the proposed system the link anomaly model is combined with word based and text based approach.

Keywords- *Anomaly detection,Burst detection, Social networks, Sequentially discounted normalized maximum-likelihood coding Topic detection.*

******

I. INTRODUCTION

Communication over social networks such as Facebook and Twitter is gaining its value in our day today's life. As the social networks are grown, the message or information exchanged between users are not only texts, but also URLs, images, and videos, they are demanding testbeds for the study of data mining. They are involved in the discovery of emerging areas from social streams which are posted by hundreds of users [1]. Unedited voice of the normal or ordinary people is able to capture via social media. Hence, the challenge is to find the emergence of topics as early as possible at a moderate number of false positives. Here, by mentioning links to other users of the same social network in the form of reply, responds and an answer came back with. One post may contain a number of comments. Some users may include comments in their posts rarely; other users may mention their friends always. The basic guess is that a new areas, i.e. new emerging area is something users of social networks are feeling to discuss, comment or forward the information further to their friends. Conventional approaches for discovering the areas have mainly been concerned with the frequencies of terms or words.

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 textual information. 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 result of the proposed paper. Section VIgives 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.

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 groups is 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 was interesting to discover the emerging topics from the post posted by hundreds of users. A new (emerging) area is something people feel like discussing, commenting, or forwarding the information further to their friends. To find the emergence of areas in social networks streams are done via Link anomaly model. Probability model is used to capture both the number of mentions per post and the frequency of it. It does not rely only on the textual contents of 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 preprocessing to obtain (the information is often separated from the contents), and are available irrespective of the nature of the contents. Since the existing method does not rely on the textual contents of social network posts, it is strong to rephrasing and it can be applied to the case where topics are concerned with information other than texts, such as pictures, video, audio, etc. All the analysis was conducted offline, and link anomaly model does not instantly tell what the anomaly is.

 ISSN: 2321-8169
1338 - 1343

IV. PROPOSED METHOD

As the social networks are grown it is interesting to discover the emerging areas. To detect the emergence of new topics from the hundreds of users based on the responds in social network posts. A probability model is proposed for mentioning behavior of social networks by the number of mentions per post and the occurrence of users taking place in the mentions. The basic assumption is that a new emerging 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 mixture of word based approach with link anomaly model would do well to both from the performance of the mention mode and the intuitiveness of the word based approach.

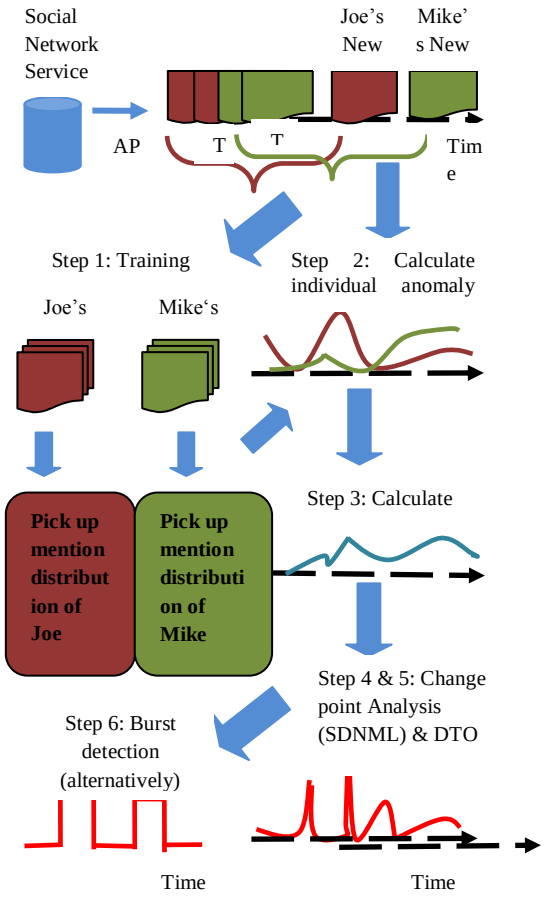

Fig. 1. Architecture Diagram

The overall architecture of the system is shown in Fig. 1. Each step in the flow is described. The data arrive from a social network service in a sequential manner through some API. For each new post, samples are used within the past time interval of length for the corresponding user for training the mention model. Assigned an anomaly score to each post based on the learned probability distribution. The score is then aggregated over users and further fed into SDNML-based change point analysis. And also describe Kleinberg's burst-detection technique, which can be used instead of the SDNML-based change-point analysis.

A. *Probability Model*

In this paper probability model is proposed to capture the normal mentioning actions of the user and to train the model. There are two types of infinity is taken into the account. The first is the number of users, where number of users cannot mention hundreds of further users in a post and would like to avoid placing an artificial limit.The second is number of post posted by users; Chinese Restaurant Process [7] is used to limit the number of mentionees. We have calculated predictive distribution with respect to the number of mentions and also for mentioning user. Link anomaly score is calculated to both predictive distribution of the number of mentions and as well as to the mentionees. Then combine the anomaly scores from hundreds of users. For each user anomaly score is calculated by current post of the user with past post of the same user. Aggregate is done for all the individual anomaly score. This aggregate is applied by change point detection through sequentially discouting normalized maximum-likelihood (SDNML)coding;dynamic threshold optimization (DTO) in addition Kleinberg's burst detection method.

B. *Algorithm Description*

1) *Change-Point Detection via SDNML coding:*

This algorithm is used to describe how to detect change points from the sequence of aggregated an anomaly scores. The change-point detection technique as follows: For suitability, the aggregated anomaly score is denoted as $x_j$instead of .

1. First-layer learning. Let $x^{j-1}:=\{x_1,\ldots..,x_{j-1}\}$ be the collection of aggregated anomaly scores from discrete time 1 to j-1. Sequentially learn the SDNML density function $p_{SDNML}(x_j|x^{j-1})(j=1,2,\ldots)$;

2. First –Layer scoring. Compute the intermediate change-point score by smoothing the log loss of the SDML density function with window size k as follows:

$$y_j = \frac{1}{K} \sum_{j'=j-k+1}^{j} (-\log P_{SDNML}(x_j \mid x^{j-1})).$$

3. Second-layer learning. Let $y^{j-1} := \{y1,\ldots.\acute{y}j\text{-}1\}$ be the collection of smoothed change-point score obtained as above. Consecutively learn the second layer SDNML density function $p_{SDNML}(y_j|y^{j-1})$ $(j=1,2\ldots)$;

4. Second-layer scoring. Compute the final change-point score by smoothing the log loss of the SDNML density function as follows:

$$\text{Score}(y_j) = \frac{1}{K} \sum_{j'=j-k+1}^{j} (-\log P_{SDNML}(y_j \mid y^{j-1})).$$

2) Dynamic Threshold Optimization (DTO):

It is a final step, should transform the change-point scores into binary alarms by threshold. As the distribution of change-point scores may change over time, it should be dynamically adjust the threshold to analyze a sequence over a long period of time. In this subsection, it is defined how to dynamically optimize the threshold using the method of dynamic threshold optimization proposed. In DTO, one-dimensional histogram is used for the representation of the score distribution. The Procedure follows:

Known: {$Score_j$ | j=1,2,…}: Scores, N_H: total number of cells, ρ: parameter for threshold, λ_H: estimation parameter, r_H: discounting parameter, M: data size .

Initialization: Let $q_1^{(1)}(h)$ (a weighted satisfactory statistics) be a uniform distribution.For j=1,…, M-1do

Threshold Optimization: Let l be the least index such that $\Sigma_{h-1}^{l} q^{(j)}(h) \geq 1-\rho$. The threshold at time j is given as

$$\eta(j) = a + \left(\frac{b-a}{N_H - 2}\right)(l+1)$$

Alarm output: Raise an alarm if $Score_j \geq \eta(j)$.

Histogram Update:

$$q_1^{(j+1)}(h) = \begin{cases} (1-r_H)q^{(j)}(h) + r_H & \text{if Score}_j \text{ falls into the hth cell,} \\ (1-r_H)q^{(j)}(h) & \text{otherwise.} \end{cases}$$

$$q^{(j+1)}(h) = (q_1^{(j+1)}(h) + \lambda_H)/(\sum_h q_1^{(j+1)}(h) + N_H \lambda_H)$$

end for

3) Kleinberg's Burst-Detection Method:

In addition to the change-point detection based on SDNML followed by DTO, it also tested the mixture of our method with Kleinberg's burst-detection method. More specifically, it is implemented a two-state version of Kleinberg's burst detection model. The reason for choosing the two-state version is because in this experiment hierarchical structure is not expected. The burst-detection process is based on a probabilistic automaton model with two states, burst state and nonburst state. Some events (e.g., arrival of posts) are assumed to happen according to a time-varying poisson processes whose rate parameter depends on the current state. The burst-detection method estimates the state transition sequence i_t £ {nonburst; burst} (t=1; . . . ; n) that maximizes the likelihood

$$P_{sw}^{b}(1-p_{sw})^{n-b} \prod_{t=1}^{n} f_{exp}(x_t;\alpha_{it}),$$

Where $p_{sw}$is a given state transition probability, b is the number of state transitions in the sequence it (t = 1, . . . , n), $f_{exp}(x;\alpha)$ is the probability density function of the exponential distribution with rate parameter α, and x_t is the t^{th}interevent interval. The optimal sequence can be capably obtained by dynamic programming. To obtain the event times and their intervals, it is defined an event as a point in time when the aggregated link anomaly score exceeds a threshold θ_{burst}.

C. ComponentIllustration

The Detection of Emerging Topics in Social Streams via Combination of Word Based Approach with Link Anomaly Model is implemented in three different module and they are:

- 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 and implement its functionalities such as posts, comments; online users, profiles and all the information will be stored in the repository. Here data has to take into account, that means the post and comments, mentions, retweets that is required for the process from the social networking stream that are used as input. Then collect the post of each and everyone who are registered with the social network stream and also deal with the comments that are posted against each post.

2) Data Collection and Training:

In this component, categorize each individual's posts first and then find the mentioning distribution. Next have to find how many numbers of mentions in a post as well as comments and also the frequency with which each user is mentioned. There are two types of infinities have to take into account here. The first is the number k of users mentioned in a post. The second type of infinity is the number of users one can possibly mention. Calculate the predictive distribution with the number of mentions. Then calculate the predictive distribution with number of users.

3) Link Anomaly Score Calculation:

In the third component, the deviation of a user's behavior from the normal mentioning behavior modeled in the previous section is computed here. To compute the

anomaly score of a new post x of user u at time t containing k mentions to users V, next step is to compute the probability with the training set T on u, which is the collection of posts by user u in the time period. The two terms are used that can be computed via the predictive distribution of the number of mentions and the predictive distribution of the mentionee respectively. The anomaly score is computed for each user depending on the current post of user u and his/her post. To measure the general trend of user behavior, and propose to aggregate the anomaly scores obtained for posts x1; . . . ; xn . Then have to find how to detect change points from the sequence of aggregated anomaly scores and plot the results in a graph. Lastly find the emerging topics in the social streams.

V. RESULT

This section shows the screen shots of the result. The results for proposed system i.e. emerging areas in the social network are obtained by SDNML-based change point analysis and DTO and with Kleinberg's burst-detection model.

Fig. 2. Change point detection (SDNML) & DTO - the number of mentions per post and the occurrence of users taking place in the mentions.

Fig. 3. Burst detection method - the number of mentions per post and the occurrence of users taking place in the mentions.

Fig.4. Emerging topics in social streams.

VI. CONCLUSION

In this paper, we have proposed a probability method to identify of emerging areas which becomes interest by the fast development of social networks. As the information exchanged in social networks post include not only the text but also images, URLs and video and so on. The fundamental idea of the approach is to focus on the social aspect of the posts reflected in the mentioning behavior of users instead of the textual contents.Probability model capture the number of mentions per post and the occurrence of users taking place in the mentionsby combining the link anomaly model with text and word based approach. The change-point detection algorithm via SDNML and Kleinberg's burst-detection model are used to pinpoint the emergence of a topic. This paper gives better performance and instantly tells what anomaly is.

REFERENCES

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

[2] Ciprian DoruGiurcaneanu,Seyed Alireza Razavi, Antti Liski, "*Variable selection in linear regression: Several approaches based on normalized maximum likelihood"*, Signal Processing, Vol. 91, pp.1671-1692, 2011.

[3] Ciprian DoruGiurcaneanu, SeyedAlirezaRazavi, "*AR order selection in the case when the model parameters are estimated by forgetting factor least-squares algorithms*", Signal Processing, Vol. 90, no.2, pp.451-466, 2010.

[4] Jorma Rissanen, Teemu Roos, Petri Myllymäki, "*Model selection 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 Alternative Model of 'Bursts' in Streams of Topics*", Proc. 16th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 443-452, 2010.

[6] Teemu Roos and Jorma Rissanen, "*On Sequentially Normalized Maximum Likelihood Models* ", Proc. Workshop Information Theoretic Methods in Science and Eng.,2008 .

[7] Y.Teh, M.Jordan, M.beal and D.Blei, "*Hierarchical DirichletProcesses* ", J.Am. Statistical Assoc., vol.101, no.476, pp.1566-1581, 2006.
[8] A.Krause, j.Leskovec and C.Guestrin, "*Data Association for Topic Intensity Tracking*", Proc, 23rd Int'l Conf. Machine Learning(ICML'06), pp.497-504, 2006.
[9] Jun-ichi Takeuchi and Kenji Yamanishi, "*A Unifying Framework for Detecting Outliers and Change Points from Time Series* ", IEEE Transactions On Knowledge And Data Engineering, Vol. 18, No. 4, April 2006.
[10] Qiaozhu Mei, ChengXiangZhai, "*Discovering Evolutionary Theme Patterns from Text An Exploration of Temporal Text Mining*", Proc. 11th ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining, pp. 198-207, 2005.

BIOGRAPHY

K.Ramyapursing M.Tech Computer Science and Engineering at Bharath University and received the B.E degree in Computer Science &Engineering from Jerusalem College of Engineering, Chennai in 2008. She worked as Executive in Matrix Business Services India Pvt Limited, Chennai from 2008 to 2009. She participated in workshops on Networking, Android and R programming language, how to build a software and She has presented the paper inNational conference on Recent Trends in Engineering NCRTE' 15.

S. BrinthaRajakumari received MCA in 2001 from the Madurai Kamarajuniversity and ME from Sathyabama University in the year 2011. She is working as an Assistant professor in the Department of CSE at Bharath University for last seven years. She has more than 12 years of experience in academic and research in areas of interest being Database, Data Warehousing and Mining and Cloud Computing. She has published over 10 research papers in international and national journals of repute. She is pursuing Ph.D in the field of Database Management System.

Dr. T.Nalini received Ph.D and M.Tech from the Bharath University in 2004, 2007 respectively. Now she is working as a professor in the Department of CSE at Bharath University. She has published more than 80 research papers in international journals. She has presented the paper in 45 national conferences and 33 international conferences, and received Radha Krishnan gold medal Award for outstanding individual achievement in 2014. She is a member of many professional bodies like ISTE, CSI, IEEE and IAENG.