RESEARCH PAPER | Engineering | Volume : 5 | Issue : 4 | April 2015 | ISSN - 2249-555X

Survey on Detection of Emerging Areas in Social Streams

KEYWORDS: Topic detection, anomaly detection, social networks, sequentially discounted normalized maximum-likelihood coding, burst detection.

K.Ramya	S.BrinthaRajakumari	Dr.T.Nalini
PG Student, Dept of CSE, Bharath University, Chennai.	Assistant Professor, Dept of CSE, Bharath University, Chennai. brintha.ramesh@gmail.com	Professor, Dept of CSE, Bharath University, Chennai. nalinicha2002@gmail.com

ABSTRACT *Discovering of emerging topics becomes interested in the fast development of 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. Emergence of topics is focused by the social aspects of these networks. In this survey paper, compares various research parameters for detection of emerging topics and the techniques used in it. The study papers was effective to understand the techniques and gives idea to propose a probability model for the mentioning behavior of social networks by the number of mentions per post and the occurrence of users taking place in the social network.*

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 topics 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.

Another dissimilarity that makes social media social is the being of mentions. 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. Finding emerging topics from social network streams based on reply or respond from the users is shown involvement. The basic assumption is that a new to topics, i.e. new emerging topic is something users of social networks are feeling to discuss, comment or forward the information further to their friends. Conventional approaches for discovering the topics have mainly been anxious with the frequencies of terms or words.

A term-frequency-based approach 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 and apply a recently proposed change point detection technique based on the sequential discounting normalized maximum-likelihood coding [3], [4], [5]. The data set is applied in a proposed system to find the emerging of topics. 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 a comparative study of the various research works explored in the previous section. Section IV, describes about future work. Section V gives the conclusion in and lastly provides references.

RELATED WORK

In this paper [1] user discovers the emerging topics from the social networks. The rapid growth of social networks is motivated and become interesting for the detection of emerging topics. 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. Here it is focused on URLs, image, videos, word, and text with mentions of user's relations between users that are generated dynamically through reply, responds and an answer came back with. Here the 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 generalized 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 estimation, therefore it is

Volume : 5 | Issue : 4 | April 2015 | ISSN - 2249-555X

widely used for stationary time series. The information theoretic criteria (ITC) have increased constantly for selecting the order of autoregressive (AR) models. The information theoretic criteria is not straightforward to employ them in combination with the forgetting factor least – squares algorithm. 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. The ITC is transformed to become compatible with the forgetting factor least squares algorithm. The theoretical analysis is difficult which will be complicated is one of the disadvantages of this paper.

In this paper [4] Author has thoroughly studied the predictive least squares (PLS) principle for model selection in perspective of regression model and autoregressive. Also introduced a new standard based on sequentially minimized squared deviation and proved that the standard has a probabilistic interpretation within a given class of distribution so called stochastic complexity. 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. The resulting model is usually not random process and the normalizing integral or sum is hard to compute in many cases. Sequential normalized maximum likelihood (SNML) is easier to compute and include a random process. Sequentially normalized least squares (SNLS) model is interesting and asymptotically optimal. SNLS is clearly 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. The different group having mixture models necessarily share mixture components. 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 approach to the problem sharing clusters among multiple related groups is a nonparametric Bayesian approach. Dirichlet process has two parameters they are scaling parameter and base probability measure.

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. In the majority of previous work the outlier detection and change point detection have not been related explicitly. 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 while accuracy is comparable and also able to detect sudden changes of variances in data distribution while the conventional ones are not. The disadvantages are 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. In this paper, discovering and summarizing the evolutionary patterns of themes in a text stream. 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 allows us to not only see the trends of strength variations of themes, but also compares the relative strengths of different themes over time. The disadvantages are flat structure of themes was not considered.

COMPARATIVE STUDY:

In this section analyzed the various research works on several parameters and presented their comparison in the table below.

TABLE 1.COMPARISON OF VARIOUS RESEARCH WORKS

S.No	Title	Author	Issue	Method Used	Tools	Advantage & Disadvantage
1.	Discovering Emerging Topics in Social Streams via Link-Anomaly Detection	Toshimitsu Takahashi, Ryota Tomioka and Kenji Yamanishi	Conventional term frequency based approaches may not be appropriate in this context.	Change point detection via SDNML coding, Dynamic Threshold Optimimization and Kleinbergs Burst Detection Method	Mentions	**Advantages:** 1) Capture the normal mentioning behavior of a user. 2) Quantitatively measure the novelty. **Disadvantages:** The proposed link anomaly model does not immediately tell what the anomaly is.

Volume : 5 | Issue : 4 | April 2015 | ISSN - 2249-555X

2.	Variable selection in linear regression: Several approaches based on normalized maximum likelihood	Ciprian DoruGiurcaneanu, Seyed Alireza Razavi, Antti Liski	The use of the normalized maximum likelihood for model selection in Gausian linear regression poses troubles because the normalization coefficient is not finite	Normalize maximum Likelihood	SC, MML, CME, BIC and AIC	**Advantage:** Introduced a general methodology from this rhomboidal constraint yields a new NML-based formula. Disadvantage: When the sample size is large the result will be modest and CME poses troubles for some model.
3.	AR order selection in the case when the model parameters are estimated by forgetting factor least-squares algorithms	Ciprian DoruGiurcaneanu, Seyed Alireza Razavi	Information theoretic criteria is not straightforward to employ them in combination with the forgetting factor least-squares algorithm.	Sequentially normalized least squares.	PLS, SRM, PDC, SNML, BIC and AIC.	**Advantage:** The predictive densities criterion (PDC) and the sequentially normalized maximum likelihood (SNML) criterion is compatible. Disadvantage: The investigation is not done for case of variable forgetting factor, which is known to account better for the non-stationary of the signal.
4.	Model selection by sequentially normalized least squares	Jorma Rissanen, Teemu Roos, Petri Myllymäki	Concerned with deriving a model selection criterion for a class of normal models.	Sequentially normalized least squares	Parameters	**Advantage:** 1) Minimize future prediction errors. 2) Evaluated efficiently & exactly Disadvantage: No adjustable hyper parameters.
5.	Topic Dynamics: An Alternative Model of 'Bursts' in Streams of Topics	Dan He, D.Stott Parker	Defined burst in terms of an arrival rate. This approach is limiting other stream dimensions.	Kleinberg's Burst model and Shasha's burst model.	Momentum.	**Advantage:** 1) Used for forecasting oncoming bursts. 2) Momentum based topic dynamics burst model have significant advantage. **Disadvantage:** Hierarchical structure deserves greater attention on burst
6.	On Sequentially Normalized Maximum Likelihood Models	Teemu Roos, Jorma Rissanen	The result is not a random process and hard to compute the sum.	Sequentially normalized maximum likelihood(SNML) & Sequentially normalized least squares (SNLS)	String of letters	**Advantage:** 1) Sequential normalized maximum likelihood (SNML) is easy to compute. 2) SNLS is a best method. **Disadvantage:** 1) BIC is known to have a tendency to underestimate rather than overestimate the order.
7.	Hierarchical Dirichlet Processes	Y.Teh, M.Jordan, M.beal, D.Blei	Group of data is considered, which is observed within the group from mixture model.	Hidden Markov model	Mixture Componets	**Advantage:** Measures explicitly. **Disadvantage:** Clustering problems can be approached within probabilistic framework during finite mixture model.
8.	Data Association for Topic Intensity Tracking	A.Krause, J.Leskovec and C.Guestrin	Data association and intensity tracking viewed as a separate task	Factorial Hidden Markov models	Documents	**Advantage:** 1)Improves the accuracy of intensity tracking & classification. 2) Detects correct topic intensities even with 30% topic noise. 3) Simultaneously address data association and intensity tracking. **Disadvantage:** very flexibility

9.	A Unifying Framework for Detecting Outliers and Change Points from Time Series	Jun-ichi akeuchi, Kenji Yamanishi	Outlier detection and change point detection have not been related explicitly.	Change point detection.	Time series	**Advantage:** Change points from nonstationary are much more efficient than conventional methods. **Disadvantage:** Challenging problem to design of an algorithm to detect variance decrease change point.
10.	Discovering Evolutionary Theme Patterns from Text An Exploration of Temporal Text Mining	Qiaozhu Mei, ChengXiang hai	Discovering evolutionary theme patterns not only reveal the hidden topic structures, but also navigation and digestion of information based on significant thematic threads.	Temporal Text Mining	Theme	**Advantage:** 1) The technique is based on hidden Markov models for analyzing the life cycle of each theme. 2) This process would first determine the globally attractive themes and then compute the strength of a theme in each time period. **Disadvantage:** Flat structure of themes was not considered

FUTURE WORK

The case study was very useful to understand the techniques. It is well understood that how the techniques are used to detect the emerging topics from social networks.

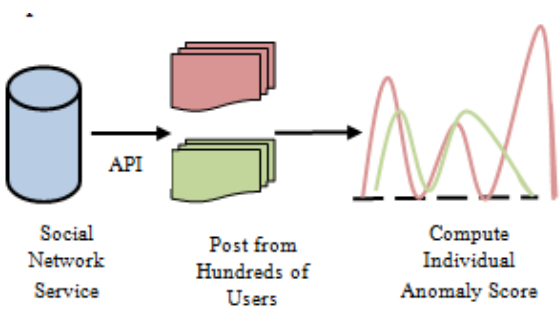

Fig. 1. Emergence of topics in social streams.

The data arrives for the social network service in sequential order through API. The post which posted by different users can be URLs, image videos, word, and text. The reply, responds to the post from hundreds of users on social network computes individually and techniques such as SDNML based change point analysis and Kleinberg's burst-detection method are applied to it. The Figure 1 shows that the post are arriving from social network service through some API and computes the anomaly scores for each post.

CONCLUSION

In this paper, literature survey on discovering of emerging topic was useful to understand the technique and how the techniques are developed to find the emerging topics from social streams. Detecting of emerging topics becomes attracted by the fast development of social networks. 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. As the information exchanged in the social networks post includes not only the text but also images, URLs and video and so on. Case study helped to propose a probability model that captures both the number of mentions per post and the frequency of mention. The SDNML change-point detection algorithm and Kleinberg's burst-detection model is used to pinpoint the emergence of a topic.

REFERENCE

[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 Dirichlet Processes ", 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, ChengXiang Zhai, "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. |