ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

DISCOVERING EMERGING TOPICS IN SOCIAL MEDIA POSTS BASED ON MENTION RELATIONSHIPS

Sandhiya.R,
PG Student,
Priyadarshini Engineering College, Vaniyambadi-635751,
Sandhiyacse06@gmail.com
Samundeeswari.M,
Associate Professor,
Priyadarshini Engineering College, Vaniyambadi-635751,
samusankar@gmail.com

ABSTRACT:

The emerging topic detection in social network is used to improve the information sharing in social networks. Conventional term-frequency-based approaches may not be appropriate in this context, because of the information exchanged are not only texts but also images, URLs, and videos.It focus on mention of users,the mention is based on anomaly detection model uses a probability model.This model is combined with SDNML change-point detection algorithm and burst detection model to pinpoint the emergence of a topic.Anomaly scores are aggregated from hundreds of users. The combination of mention-anomaly model with text-based approaches is used, because the mention & text based anomaly model would benefit both from the performance of the mention models and the intuitiveness of the word-based approach.

Index Terms—Topic detection, anomaly detection, social networks, sequentially discounted normalized maximum-likelihood coding, burst detection.

INTRODUCTION

All Rights Reserved © 2016 IJARBEST

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

Over the past few years the Internet has not only become the most important source of information, but also a key-player in event formation. The open community of publishing news and information made it an important indication for the pulse of the society.Social networks have become a very important source of information and recently a source of creating information. Blogs, Twitter and Facebook, have played a great role in near past and current events all over the world. For all this, it was very important to have a system that can extract these information without human intervention. The participation in social media, e.g., posting and/or reading, has gradually become a routine part of many peoples'lives.

The posts cover a wide range of topics,including daliy activities, event,opinions,comments,photographs,andlinks to web pages. The popularity of this form of communication has been driven by advances in mobile phone technology.Smartphone,which enable access to internet services, are becoming increasingly popular at an unprecedented rate. Social media applications for smartphones have also been developed and popularized. These client applications have features that exploitsmartphone'sancillary functions such as a global posting system (GPS) and acamera,which,for example,enable users to post still or video images,and determines their current location. The detection of topics can be done in many ways. But detecting the topics through social network is a big challenge in web. The detection of topics via micro blog sites sites such as twitter. The topic detection can be done easily if the shared information is in the form of text. The text anomaly based approach is suitable if the posted is image, video or url. Hence the link anomaly based approach is proposed to detect the new topics.

The four data sets I have analyzed above, the proposed link-anomaly-based methods compared favorably against the text-anomaly-based methods on "Youtube", "NASA", and "BBC" data sets.

On the other hand, the text anomaly- based methods were earlier to detect the topics on "Job hunting" data set. The proposed link-anomaly-based methods performed even better than the keyword-based methods on "NASA" and "BBC" data sets. The above results support that the emergence of new topic is reflected in the anomaly of the way people communicate to each other and also that such emergence shows up earlier and more reliably in the anomaly

All Rights Reserved © 2016 IJARBEST

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

of the mentioning behavior than the anomaly of the textual contents. This is probably because the textual words suffer from variations caused by rephrasing.

In this paper, we propose a probability model that can capture the normal mentioning behaviour of a uses and proposed change-point detection technique based on the Sequentially Discounting Normalized Maximum Likelihood (SDNML) coding [3]. This technique can detect a change in the statistical dependence structure in the time series of aggregated anomaly scores, and pin- point where the topic emergencies.

Topic tracking is the process of monitoring a stream of news stories to find those that track (or discuss) the same event as one specified by a user.

TASKS OF TDT:

Five tasks of TDT: 1)Story segmentation No need for it if we have already the data as documents.2)Topic detection Builds a set of clusters, each contains stories about the same topic. It assigns a story to one or more possible cluster.3)Topic tracking is the selection of a certain cluster specified by one or more example stories First story detection detects a story discusses unknown topic. It generates a new cluster for this topic.4)Link detection a kernel function which established if two stories are linked or not.

The detection of topics can be done in many ways. But detecting the topics through social network is a big challenge in web. The detection of topics via micro blog sites sites such as twitter. The topic detection can be done easily if the shared information is in the form of text. The text anomaly based approach is suitable if the posted is image, video or url. Hence the link anomaly based approach is proposed to detect the new topics.

RELATED WORK:

Detection and tracking of topics have been studied exten- sively in the area of topic detection and tracking (TDT) [1]. In this context, the main task is to either classify a new document into one of the known topics (tracking) or to detect that it belongs to none of the known categories. Sub- sequently, temporal structure of topics have been modeled and analyzed through dynamic model selection [4], temporal text mining [5], and factorial hidden

All Rights Reserved © 2016 IJARBEST

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

Markov models [6]. Another line of research is concerned with formalizing the notion of "bursts" in a stream of documents.

In his seminal paper, Kleinberg modeled bursts using time varying Poisson process with a hidden discrete process that controls the firing rate [2]. Recently, He and Parker developed a physics inspired model of bursts based on the change in the momentum of topics [7].

All the above mentioned studies make use of textual content of the documents, but not the social content of the documents. The social content (links) have been utilized in the study of citation networks [8]. However, citation networks are often analyzed in a stationary setting. The novelty of the current paper lies in focusing on the social content of the documents (posts) and in combining this with a change-point analysis.

PROPOSED APPROACH:

The overall flow of the proposed method is shown in Figure 1. We assume that the data arrives from a social network service in a sequential manner through some API.

A.Probability Model

Describe the probability model that I use to capture the normal mentioning behavior of a user and how to train the model. I have to characterize a post in a social network stream by the number of mentions k it contains, and the set V of names (IDs) of the mentions (users who are mentioned in the post). There are two types of infinity I have to take into account here. The first is the number k of users mentioned in a post. Although, in practice a user cannot mention hundreds of other users in a post, I would like to avoid putting an artificial limit on the number of users mentioned in a post.

Instead, I will assume a geometric distribution and integrate out the parameter to avoid even an implicit limitation through the parameter. The second type of infinity is the number of users one can possibly mention.

Joint probability distribution

All Rights Reserved © 2016 IJARBEST

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

$$P(k, V \mid \theta, \{\pi_v\})) = P(k \mid \theta) \prod_{v' \in V} \pi_{v'}.$$

Mentioning new user

$$P(\{v : m_v = 0\} \mid T) = \frac{\gamma}{m + \gamma}.$$

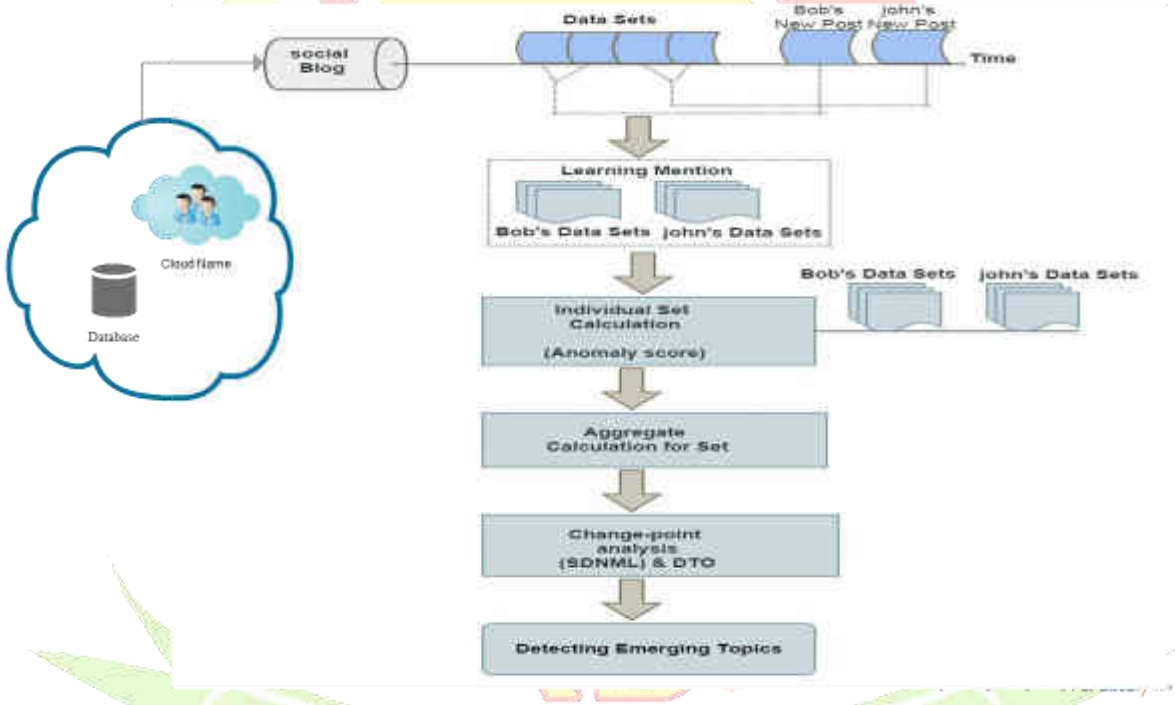

Figure 1:Overall method of proposed approach

B.Computing the link-anomaly score

Describe how to compute the deviation of a user's behavior from the normal mentioning behavior modeled In order to compute the anomaly score of a new post $\boldsymbol{x}$ = (t, u, k, V) by user u at time t containing k mentions to users V , I compute the probability with the

All Rights Reserved © 2016 IJARBEST

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

training set $T(t)\ u$, which is the collection of posts by user u in the time period $[t-T, t]$ (use $T = 30$ days in this paper).

Accordingly the link-anomaly score is defined as

$$s(x) = \log\left(P(k \mid T_u^{(t)}) \prod_{v \in V} P(v \mid T_u^{(t)})\right)$$
$$= \log P(k \mid T_u^{(t)}) \sum_{v \in V} \log P(v \mid T_u^{(t)}).$$

The two terms in the above equation can be computed via the predictive distribution of the number of mentions, and the predictive distribution of the mentionee.

C.Combining Anomaly Scores from Different Users

Describes how to combine the anomaly scores from different users; The anomaly score in is computed for each user depending on the current post of user u and his/her past behavior $T(t)\ u$. In order to measure the general trend of user behavior, I propose to aggregate the anomaly scores obtained for posts $\boldsymbol{x}1...\boldsymbol{x}n$ using a discretization of window size $\tau > 0$ as follows:

$$s'_j = \frac{1}{\tau} \sum_{t_i \in [\tau(j-1), \tau j]} s(x_i),$$

D.SDNML based change point detection model

Given an aggregated measure of anomaly , we apply a change-point detection technique. This technique is an extension of Change Finder proposed in that detects a change in the statistical dependence structure of a time series by monitoring the compressibility of a new piece of data. Urabe et al. proposed to use a sequential version of normalized maximum-likelihood (NML) coding called SDNML coding as a coding criterion instead of the plug-in predictive distribution used in. Specifically, a change point is detected through two layers of scoring processes. The first layer detects outliers and the second layer detects change-points .In each layer, predictive loss based on the SDNML coding distribution for an autoregressive (AR) model is used as a criterion for scoring. Although the NML code length is known to be optimal , it is often hard to compute. The SNML proposed in is an approximation to the NML

All Rights Reserved © 2016 IJARBEST

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

code length that can be computed in a sequential manner. The SDNML proposed in further employs discounting in the learning of the AR models.

The SDNML density p_{SDNML} in the change-point detection algorithm described in is obtained by applying the SNML proposed by Roos et al. to the class of AR model with a discounted ML estimation, which makes the SDNML-based change-point detection algorithm more flexible than an SNML based one. We define the pth-order AR model as follows:

$$p(x_t \mid x_{t-k}^{t-1} : \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \left(x_t - \sum_{i=1}^{p} a^{(i)} x_{t-i} \right)^2 \right),$$

E.Dynamic Threshold Optimization

Dynamic Threshold Optimization (DTO) adaptively "compresses" the decision space (DS) in a global search and optimization problem by bounding the objective function from below. This approach is different from "shrinking" DS by reducing bounds on the decision variables. DTO is applied to Schwefel's in 2 and 30 dimensions with good results. DTO is universally applicable, and the author believes it may be a novel approach to global search and optimization.

DTO is conceptually quite simple. Objective function $f(x)$ is multimodal with many local maxima and a single global maximum, and the problem is to locate that maximum (coordinates and value). DTO bounds $f(x)$ from below using a series of successively increasing "thresholds," in effect compressing DS in the direction of the dependent variable (from "below") instead of, as is sometimes done, shrinking DS by reducing the independent variable's domain (from the "sides"). Locating the global maximum is easier in the compressed DS because unwanted local maxima are progressively filtered out as the "floor" (threshold) rises. Because DTO is a general geometric technique, it is algorithm-independent so that it can be used with any global search and optimization algorithm. Although DTO is

All Rights Reserved © 2016 IJARBEST

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

described in the context of maximization, it can be applied to minimization with obvious modifications because $\max f(x)$ $\min f(x) f(x) = -\min f(x)$

DTO appears to be an effective technique for adaptively changing the topology of the decision space in a multidimensional search and optimization problem. DTO should be useful with any search and optimization algorithm. Bounding DS from below removes local maxima, and as the threshold or "floor" is increased, more and more local maxima are eliminated. In the limit, DS collapses to a plane whose value ("height") corresponds to the value of the global maximum. In that case, DS contains no information as to the global maximum's location, but the maximum's value is known precisely. In order to preserve location information, the DTO threshold should not be set too high, thereby retaining enough structure for efficient DS exploration.

There are many unanswered questions concerning how DTO should be implemented. For example, there almost certainly are better ways to set the threshold than the simple linear scheme used here. Thresholds that are progressively closer together probably will work better. Another question arises in connection with what optimization algorithm should be used. Even though DTO is algorithm-independent, it may work best when different algorithms are combined to take advantage of their different strengths and weaknesses. For example, CFO, which is inherently deterministic, often converges very quickly to the vicinity of a global maximum (good exploitation). But it's very determinism inhibits exploration in decision spaces with "sparse" structure (mostly planar, few local maxima).

F.Detecting Emerging topics

In addition to the change-point detection based on SDNML followed by DTO described in previous sections, we also test the combination of our method with Kleinberg's burst-detection model. More specifically, we implemented a two-state version of Kleinberg's burst detection model. The reason we chose the two-state version was because in this experiment we expect no hierarchical structure. The burst-detection method is based on a probabilistic automaton model with two states, burst state and non burst 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 that maximizes the likelihood

All Rights Reserved © 2016 IJARBEST

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

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

where p_{sw} is a given state transition probability, b is the number of state transitions in the sequence is the probability density function of the exponential distribution with rate parameter inter event interval. The optimal sequence can be efficiently obtained by dynamic programming.

CONCLUSION:

I have proposed a new approach to detect the emergence of topics in a social network stream. The basic idea of this approach is to focus on the social aspect of the posts reflected in the mentioning behavior of users instead of the textual contents. I have proposed a probability model that captures both the number of mentions per post and the frequency of mentionee. I have combined the proposed mention model with the SDNML change-point detection algorithm to pinpoint the emergence of a topic. Since the proposed method does not rely on the textual contents of social network posts, it is robust to rephrasing and it can be applied to the case where topics are concerned with information other than texts, such as images, video, audio, and so on.

REFERNCES

1. J. Allan et al., "Topic Detection and Tracking Pilot Study: Final Report," *Proc. DARPA Broadcast News Transcription and Understanding Workshop,* 1998.

2. D. Aldous, "Exchangeability and Related Topics," _ *Ecole d' _ Ete´ de Probabilite´s de Saint-Flour XIII—1983,* pp. 1-198, Springer, 1985.

3. C. Giurc_aneanu, S. Razavi, and A. Liski, "Variable Selection in Linear Regression: Several Approaches Based on Normalized Maximum Likelihood," *Signal Processing,* vol. 91, pp. 16711692, 2011.

4. C. Giurc_aneanu and S. Razavi, "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.

All Rights Reserved © 2016 IJARBEST

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

5. D. He and D.S. 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. 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.

7. J. Kleinberg, "Bursty and Hierarchical Structure in Streams," *Data Mining Knowledge Discovery,* vol. 7, no. 4, pp. 373-397, 2003.

8. D. Lewis, "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval," *Proc. 10th European Conf. Machine Learning (ECML' 98),* pp. 4-15, 1998.

9. S. Morinaga and K. Yamanishi, "Tracking Dynamics of Topic Trends Using a Finite Mixture Model," *Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,* pp. 811-816, 2004.

10. Q. Mei and C. 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.

All Rights Reserved © 2016 IJARBEST