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

Discovering Emerging Topics in Social Streams via Link-Anomaly Detection

C.Dedeepya Divya
P.G. Scholar (M. Tech),
Department of CSE,
Modugula Kalavathamma Institute of Technology for Women, Rajampet, Kadapa District.

T.Mallika Devi
Associate Professor,
Department of CSE,
Modugula Kalavathamma Institute of Technology for Women, Rajampet, Kadapa District.

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 user links between users that are generated dynamically (intentionally or unintentionally) through replies, mentions, and retweets. We propose a probability model of the mentioning behavior 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.

Index Terms:

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

I. INTRODUCTION:

Communication through social networks, such as Face book 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.There is another type of information that is intentionally or unintentionally exchanged over social networks: mentions. Here we mean by mentions links to other users of the same social network in the form of message-to, reply-to, retweet-of, or explicitly in the text. One post may contain a number of mentions. Some users may include mentions in their posts rarely; other users may be mentioning their friends all the time. Some users (like celebrities) may receive mentions every minute; for others, being mentioned might be a rare occasion. In this sense, mention is like a language with the number of words equal to the number of users in a social network. We are interested in detecting emerging topics from social network streams based on monitoring the mentioning behaviour of users. Our basic assumption is that a new (emerging) topic is something people feel like discussing about, commenting about, or forwarding the information further to their friends. Conventional approaches for topic detection have mainly been concerned with the frequencies of (textual) words .A term frequency based approach could suffer from the ambiguity caused by synonyms or homonyms. It may also require complicated preprocessing (e.g., segmentation) depending on the target language. Moreover, it cannot be applied when the contents of the messages are mostly non-textual information. On the other hands, the “words” formed by mentions are unique, requires little prepossessing to obtain (the information is often separated from the contents), and are available regardless of the nature of the contents. In this paper, we propose a probability model that can capture the normal mentioning behaviour of a user, which consists of both the number of mentions per post and the frequency of users occurring in the mentions. Then this models used to measure the anomaly of future user behaviour.Using the proposed probability model, we can quantitatively measure the novelty or possible impact of a post reflected in the mentioning behavior of the user. We aggregate the anomaly scores obtained in this way over hundreds of users and apply a recently proposed change-point detection technique based on the Sequentially Discounting Normalized Maximum Likelihood (SDNML) coding [3].

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

This technique can detect a change in the statistical dependence structure in the time series of aggregated anomaly scores, and pinpoint where the topic emergence is; see Figure 1. The effectiveness of the proposed approach is demonstrated on two data sets we have collected from Twitter.

II. RELATED WORK:

Detection and tracking of topics have been studied extensively in the area of topic detection and tracking (TDT) 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. Subsequently, temporal structure of topics have been modeled and analyzed through dynamic model selection, temporal text mining, and factorial hidden Markov models .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 physics inspired model of bursts based on the change in the momentum of topics 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.

EXISTING SYSTEM:

A new (emerging) topic is something people feel like discussing, commenting, or forwarding the information further to their friends. Conventional approaches for topic detection have mainly been concerned with the frequencies of (textual) words.

DISADVANTAGES OF EXISTING SYSTEM:

A term-frequency-based approach could suffer from the ambiguity caused by synonyms or homonyms. It may also require complicated pre-processing (e.g., segmentation) depending on the target language. Moreover, it cannot be applied when the contents of the messages are mostly nontextual information. On the other hand, the "words" formed by mentions are unique, require little pre-processing to obtain (the information is often separated from the contents), and are available regardless of the nature of the contents.

III. PROPOSED SYSTEM:

In this paper, we have proposed a new approach to detect the emergence of topics in a social network stream. The basic idea of our approach is to focus on the social aspect of the posts reflected in the mentioning behaviour of users instead of the textual contents. We have proposed a probability model that captures both the number of mentions per post and the frequency of mentionee.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. For each new post we use samples within the past T time interval for the corresponding user for training the mention model we propose below. We assign anomaly score to each post based on the learned probability distribution. The score is then aggregated over users and further fed into a change point analysis.

A. Probability Model:

We 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 users mentioned in the post. Formally, we consider the following joint probability distribution.

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

Here the joint distribution consists of two parts: the probability of the number of mentions k and the probability of Each mention given the number of mentions. The probability of the number of mentions P(kj_) is defined as a geometric distribution with parameter _ as follows:

$$P(k|\theta) = (1-\theta)^k \theta.$$

On the other hand, the probability of mentioning users in Vis defined as independent, identical multinomial distribution with parameters

$$\pi_v \ (\textstyle\sum_v \pi_v = 1).$$

Suppose that we are given n training examples T =f(k1; V1), : : :, (kn; Vn)g from which we would like to learn the predictive distribution

$$P(k, V|\mathcal{T}) = P(k|\mathcal{T}) \prod_{v \in V} P(v|\mathcal{T}).$$

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

First we compute the predictive distribution with respect to the the number of mentions P(kjT). This can be obtained By assuming a beta distribution as a prior and integrating out the parameter _. The density function of the beta prior Distribution is written as follows:

$$p(\theta|\alpha,\beta) = \frac{(1-\theta)^{\beta-1}\theta^{\alpha-1}}{B(\alpha,\beta)},$$

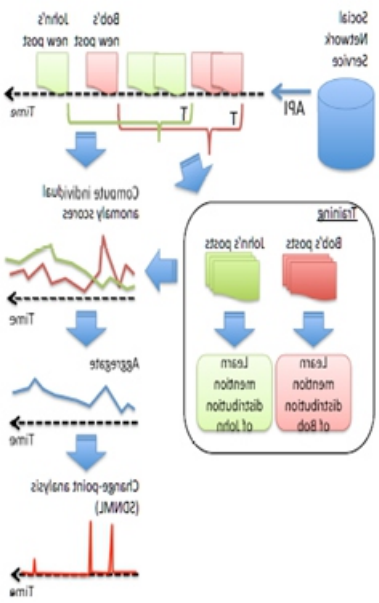


Figure 1. Overall flow of the proposed method.

Where _ and _ are parameters of the beta distribution and B(_; _) is the beta function. By the Bayes rule, the predictive distribution can be obtained as follows:

$$P(k|\mathcal{T},\alpha,\beta) = \frac{P(k,k_1,\ldots,k_n|\alpha,\beta)}{P(k_1,\ldots,k_n|\alpha,\beta)} = \frac{\int_0^1 (1-\theta)^{\sum_{i=1}^n k_i+k+\beta-1}\theta^{n+1+\alpha-1}\mathrm{d}\theta}{\int_0^1 (1-\theta)^{\sum_{i=1}^n k_i+\beta-1}\theta^{n+\alpha-1}\mathrm{d}\theta}.$$

Both the integrals on the numerator and denominator can be obtained in closed forms as beta functions and the predictive distribution can be rewritten as follows:

$$P(k|\mathcal{T},\alpha,\beta) = \frac{B(n+1+\alpha,\sum_{i=1}^n k_i+k+\beta)}{B(n+\alpha,\sum_{i=1}^n k_i+\beta)}.$$

Using the relation between beta function and gamma function we can further simplify the expression as follows:

$$P(k|\mathcal{T},\alpha,\beta) = \frac{n+\alpha}{m+k+\beta}\prod_{j=0}^{k}\frac{m+\beta+j}{n+m+\alpha+\beta+j},$$

Next, we derive the predictive distribution P(vjT) of mentioning user v. The maximum likelihood (ML) estimator is given as P(vjT) = mv=m, where m is the number of total mentions and mv is the number of mentions to user v in the data set T . The ML estimator, however, cannot handle users that did not appear in the training set T ; it would assign probability zero to all these users, which would appear infinitely anomalous in our framework. Instead we use the Chinese Restaurant Process (CRP; see [9]) based estimation. The CRP based estimator assigns probability to each user v that is proportional to the number of mentions mv in the training set T ; in addition, it keeps probability proportional to for mentioning someone who was not mentioned in the training set T . Accordingly the probability of known users is given as follows:

$$P(v|\mathcal{T}) = \frac{m_v}{m+\gamma} \quad (\text{for } v\text{: } m_v \geq 1).$$

On the other hand, the probability of mentioning a new user is given as follows:

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

B. Computing the link-anomaly score

In order to compute the anomaly score of a new post x = (t; u; k; V) by user u at time t containing k mentions to users V , we compute the probability (3) with the training set T (t) u , which is the collection of posts by user u in the time period [t T; t] (we use T = 30 days in this paper). Accordingly the link-anomaly score is defined as follows:

$$s(\boldsymbol{x}) = -\log P(k|\mathcal{T}_u^{(t)}) - \sum_{v\in V}\log P(v|\mathcal{T}_u^{(t)}). \quad (\text{illegible})$$

The two terms in the above equation can be computed via the predictive distribution of the number of mentions (4), and the predictive distribution of the mentionee (5)–(6), respectively.

C. Combining Anomaly Scores from Different Users:

The anomaly score in (7) is computed for each user depending on the current post of user u and his/her past behaviour T (t) u . In order to measure the general trend of user behaviour, we propose to aggregate the anomaly scores obtained for posts x1; : : : ; xn using a discretization of window size T_ > 0 as follows:

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

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

D. Change-point detection via Sequentially Discounting Normalized Maximum Likelihood Coding:

Given an aggregated measure of anomaly (8), we apply a change-point detection technique based on the SDNML coding [3]. This technique detects a change in the statistical dependence structure of a time series by monitoring theCompressibility of the new piece of data. The SDNML proposed in [3] is an approximation for normalized maximum likelihood (NML) code length that can be computed sequentially and employs discounting in the learning of the AR models; see also . Algorithmically, the change point detection procedure can be outlined as follows. For convenience, we denote the aggregate anomaly score as xj instead of s′j .

E. Dynamic Threshold Optimization (DTO):

We make an alarm if the change-point score exceeds a threshold, which was determined adaptively using the method of dynamic threshold optimization (DTO) [13]. In DTO, we use a 1-dimensional histogram for the representation of the score distribution. We learn it in a sequential and discounting way. Then, for a specified value _, to determine the threshold to be the largest score value such that the tail probability beyond the value does not exceed _. We call _ a threshold parameter. The details of DTO are summarized in Algorithm 1.

ADVANTAGES OF PROPOSED SYSTEM:

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. The proposed link-anomaly-based methods performed even better than the keyword-based methods on “NASA” and “BBC” data sets.

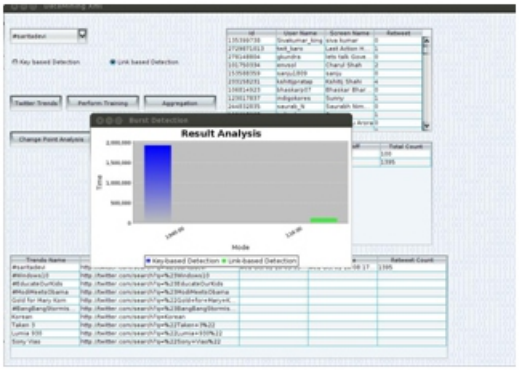
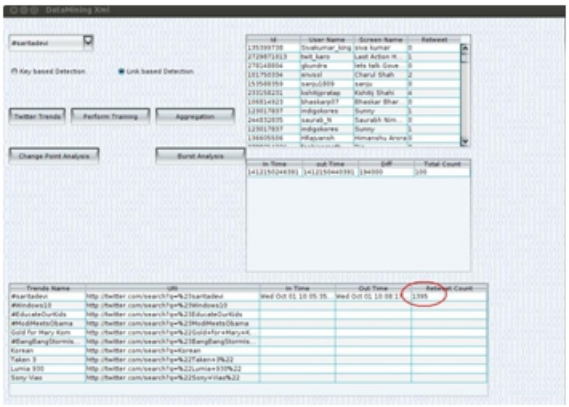


Figure 2. Result analysis between link based detection and key based detection



Figurer3. In link based detection Aggregation counts retweet counts along with start time and end time.

V. CONCLUSION:

In this paper, we have proposed a new approach to detect the emergence of topics in a social network stream. The basic idea of our approach is to focus on the social aspect of the posts reflected in the mentioning behavior of users instead of the textual contents. We have proposed a probability model that captures both the number of mentions per post and the frequency of mentionee. Furthermore, we have combined the proposed mention model with recently proposed SDNML change-point detection algorithm [3] to pin-point the emergence of a topic. We have applied the proposed approach to two real data sets we have collected from Twitter. In all the data sets our proposed approach showed promising performance; the detection by the proposed approach was as early as term frequency based approaches in the hindsight of the keywords that best describes the topic that we have manually chosen afterwards. Furthermore, for “BBC” data set, in which the keyword that defines the topic is more ambiguous than the other data set, the proposed link-anomaly based

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

approach has detected the emergence of the topics much earlier than the keyword-based approach.

REFERENCES:

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, Y. Yang et al., "Topic detection and tracking pilot study: Final report," in Proceedings of the DARPA broadcast news transcription and understanding workshop, 1998.

[2] J. Kleinberg, "Bursty and hierarchical structure in streams,"Data Min. Knowl. Disc., vol. 7, no. 4, pp. 373–397, 2003.

[3] Y. Urabe, K. Yamanishi, R. Tomioka, and H. Iwai, "Realtime change-point detection using sequentially discounting normalized maximum likelihood coding," in Proceedings. Of the 15th PAKDD, 2011.

[4] S. Morinaga and K. Yamanishi, "Tracking dynamics of topic trends using a finite mixture model," in Proceedings of the 10th ACM SIGKDD, 2004, pp. 811–816.

[5] Q. Mei and C. Zhai, "Discovering evolutionary theme patterns from text: an exploration of temporal text mining," in Proceedings of the 11th ACM SIGKDD, 2005, pp. 198–207.

[6] A. Krause, J. Leskovec, and C. Guestrin, "Data association for topic intensity tracking," in Proceedings of the 23rd ICML, 2006, pp. 497–504.

[7] D. He and D. S. Parker, "Topic dynamics: an alternative model of bursts in streams of topics," in Proceedings of the 16th ACM SIGKDD, 2010, pp. 443–452.

[8] H. Small, "Visualizing science by citation mapping," Journal of the American society for Information Science, vol. 50, no. 9, pp. 799–813, 1999.

[9] D. Aldous, "Exchangeability and related topics," in E´cole d'E´te´ de Probabilite´s de Saint-Flour XIII—1983. Springer, 1985, pp. 1–198.

[10] J. Takeuchi and K. Yamanishi, "A unifying framework for detecting outliers and change points from time series," IEEE T. Knowl. Data En., vol. 18, no. 44, pp. 482–492, 2006.

[11] J. Rissanen, "Strong optimality of the normalized ML models as universal codes and information in data," IEEE T. Inform. Theory, vol. 47, no. 5, pp. 1712–1717, 2002.