Asian Journal of Science and Applied Technology (AJSAT)
Vol.3.No.1 2015 pp 10-13
available at: www.goniv.com
Paper Received :08-03-2015
Paper Accepted:20-03-2015
Paper Reviewed by: 1. R. Venkatakrishnan 2. R. Marimuthu
Editor : Prof. P.Muthukumar

DETECT MENTIONS IN SOCIAL STREAM VIA MENTION ANOMALY MODEL

C. RAVICHANDRAN., S.MURUGESAN, M.E
M.E CSE (Student), Asst.Professor
Tagore Engineering College Chennai
sunmoon.ravi11@gmail.com, muruga13@gmail.com

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. Based on focus of the emergence of topics signaled by social aspects of these networks. Specifically, I focus on mentions of user links between users that are generated dynamically (intentionally or unintentionally) through replies, mentions, and retweets. I 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. I demonstrate my technique in several real data sets 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 over social networks, such as Facebook and Twitter, is gaining 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. Therefore, the challenge is to detect the emergence of a topic as early as possible at a moderate number of false positives. Another difference that makes social media social is the existence of 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 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. Our basic assumption is that a new emerging topic is something people feel like discussing, commenting, or forwarding the information further to their friends. On the other hand, the "words" formed by mentions are unique, require little preprocessing to obtain, and are available regardless of the nature of the contents. In this paper, we propose a probability model that can capture the normal mentioning behavior of a user, which consists of both the number of mentions per post and the frequency of users occurring in the mentions. Then this model is used to measure the anomaly of future user behavior. 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

II LITERATURE SURVEY

Detection and tracking of topics have been studied extensively 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. Subsequently, temporal structure of topics has been modeled and analyzed through dynamic model selection [4], temporal text mining [5], and factorial hidden 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 the time varying Poisson process with a hidden discrete process that controls the firing rate [2]. The social content (links) has been utilized in the study of citation networks [8]. However, citation networks are often analyzed in a stationary setting.

III PROPOSED WORK

Each step in the flow is described in the corresponding subsection. We assume that the data arrive from a social network service in a sequential manner through some API. For each new post, we use samples within the past time interval of length T for the corresponding user for training the mention model we propose below:

(Step 1): We assign an anomaly score to each post based on the learned probability distribution

(Step 2): The score is then aggregated over users

(Step 3): and further fed into SDNML-based change point analysis

(Steps 4 and 5): We also describe Kleinberg’s burst-detection method, which can be used instead of the SDNML-based change-point analysis.

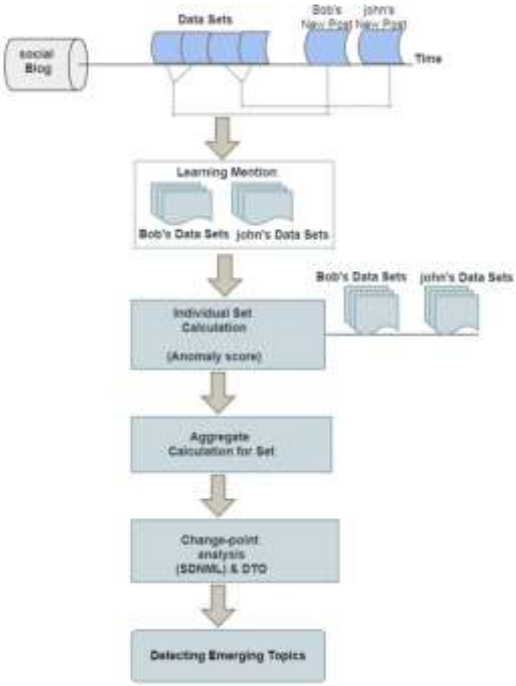

Fig 1. Overall flow of the proposed method

A. *Probability Model*

In this subsection, we describe the probability model that we used to capture the normal mentioning behavior of a user and how to train the model; see Step 1. 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 mentionees (users who are mentioned in the post). There are two types of infinity we 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, we would like to avoid putting an artificial limit on the number of users mentioned in a post. Instead, we 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. To avoid limiting the number of possible mentionees, we use Chinese Restaurant Process (CRP) based estimation; use CRP for infinite vocabulary.

B. Computing the Link-Anomaly Score

In this subsection, we describe how to compute the deviation of a user’s behavior from the normal mentioning behavior modeled in the previous subsection; see Step 2. 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 $T_u^{(t)}$, which is the collection of posts by user u in the time period (t –T, t). (we use T ¼ 30 days in this paper).

C. Change-Point Detection via SDNML Coding

In this subsection, we describe how to detect change points from the sequence of aggregated anomaly scores; see Step 4. Given an aggregated measure of anomaly, we apply a change-point detection technique. This technique is an extension of Change Finder that detects a change in the statistical dependence structure of a time series by monitoring the compressibility of a new piece of data. 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.

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

Algorithmically, the change-point detection procedure can be outlined as follows: For convenience, we denote the aggregated anomaly score as x_j instead of s0j.

1. First-layer learning. Let x $^{(j-1)}$ = {x_1., xj$_{-1}$} be the collection of aggregated anomaly scores from discrete time 1 to j _ 1.

2. First-layer scoring. Compute the intermediate change-point score by smoothing the log loss of the SDNML density function with window size.

3 .Second-layer learning. Let y^{j-1}= {y_1…, yj_{-1}} be the collection of smoothed change-point score obtained as above.
4. Second-layer scoring. Compute the final change-point score by smoothing the log loss of the SDNML density function.

D. Dynamic Threshold Optimization (DTO)
As a final step in our method, we need to convert the change-point scores into binary alarms by thresholding. Since the distribution of change-point scores may change over time, we need to dynamically adjust the threshold to analyze a sequence over a long period of time. In this subsection, we describe how to dynamically optimize the threshold using the method of dynamic threshold optimization; see Step 5. In DTO, we use a one-dimensional histogram for the representation of the score distribution. We learn it in a sequential and discounting way.

E. Kleinberg's Burst-Detection Method
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 method. 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 are assumed to happen according to a time-varying Poisson processes whose rate parameter depends on the current state.

IV. DISCUSSION

Within the four data sets we have analysed above, the proposed link-anomaly-based methods compared favourably 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 our hypothesis 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 of the mentioning behaviour than the anomaly of the textual contents.

In particular, in the case of "NASA" data set, people had been mentioning "arsenic"-eating organism earlier than NASA's official release but only rarely. Thus, the keyword frequency- based methods could not detect the keyword as an emerging topic, although the keyword "arsenic" appeared earlier than the official release. For "BBC" data set, the proposed link-anomaly-based burst model detects two bursty areas. Interestingly, the link-anomaly-based change-point analysis only finds the first area, whereas the text-anomaly-based methods and the keyword-frequency-based methods only find the second area. In our approach, the alarm was raised if the change-point score exceeded a dynamically optimized threshold based on the significance level parameter.

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. We have combined the proposed mention model with the SDNML change-point detection algorithm and Kleinberg's burst-detection model 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.

We have applied the proposed approach to four real data sets that we have collected from Twitter. The four data sets included a wide-spread discussion about a controversial topic ("Job hunting" data set), a quick propagation of news about a video leaked on Youtube ("Youtube" data set), a rumor about the upcoming press conference by NASA("NASA" data set),and an angry response to a foreign TV show ("BBC" data set). In all the data sets, our proposed approach showed promising performance.

In three out of four data sets, the detection by the proposed link-anomalybased methods was earlier than the text-anomaly-based counterparts. Furthermore, for "NASA" and "BBC" data sets, in which the keyword that defines the topic is more ambiguous than the first two data sets, the proposed link-anomaly-based approaches have detected the emergence of the topics even earlier than the keyword-based approaches that use hand-chosen keywords.

REFERENCES

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

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

[3] Y. Urabe, K. Yamanishi, R. Tomioka, and H. Iwai, "Real-Time Change-Point Detection Using Sequentially Discounting Normalized Maximum Likelihood Coding," Proc. 15th Pacific-Asia

Conf. Advances in Knowledge Discovery and Data Mining (PAKDD' 11), 2011.

[4] *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.*

[5] *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.*

[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] *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.*

[8] *H. Small, "Visualizing Science by Citation Mapping," J. Am. Soc. Information Science, vol. 50, no. 9, pp. 799-813, 1999.*

[9] *D. Aldous, "Exchangeability and Related Topics," Saint-Flour XIII—1983, pp. 1-198, Springer, 1985.*

[10] *Y. Teh, M. Jordan, M. Beal, and D. Blei, "Hierarchical Dirichlet Processes," J. Am. Statistical Assoc., vol. 101, no. 476, pp. 1566-1581, 2006.*

[11] *K. Yamanishi and J. Takeuchi, "A Unifying Framework for Detecting Outliers and Change Points from Non-Stationary Time Series Data," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2002.*