Detecting the Emerging topic from Social Network

Pallavi Chakinala
(M.Tech student) Department of Computer science and Engineering Auraba’s Technological and Research Institute
Email id:Pallavi.chakinala@gmail.com

Mrs. B. Malathi
(Associate Professor) Deportment of CSE and IT Auraba’s Technological and Research Institute
Email id:malathi.bio10@gmail.com

Abstract:
Social network is a place where people exchange and share information related to the current events all over the world. 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. A term-frequency-based approach could suffer from the ambiguity caused by synonyms or homonyms. 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.

It may also require complicated pre-processing (e.g., segmentation) depending on the target language. We propose a probability model of the mentioning behaviour of a social network user, and propose to detect the emergence of a new topic from the anomalies measured through the model. We demonstrate our technique in several real data sets we gathered from social networks. We have combined the proposed mention model with the SDNML change point detection algorithm and Kleinbreg’s burst detection model to pinpoint the emergence of a topic. We have to do is deliver to people the best and freshest most relevant information possible.

Key Words:-Social network; Anomaly Detection; Topic 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 testbeds 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. Compared to conventional media, social media are able to capture the earliest, unedited voice of ordinary people. 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 (like celebrities) may receive mentions every minute; for others, being mentioned might be a rare occasion. In this sense, a 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 behavior of users. Our basic assumption is that 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. 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 hand, the “words” formed by mentions are unique, require little preprocessing to obtain (the information is often separated from the contents), and are available regardless of the nature of the contents.

**Purpose:**
The queried levels are transmuted into polyhedral with high security (in the RASP perturbation) in the RASP-perturbed data space, that can be processed with the support of indexing structures in the perturbed space. The RASP kNN query service (kNN-R) uses the RASP range query service to process kNN queries. The key components in the RASP framework include (1) the definition and properties of RASP perturbation; (2) the construction of the privacy-preserving range query services; (3) the construction of privacy-preserving kNN query services; and (4) an analysis of the attacks on the RASP-protected data and queries. In summary, the proposed approach has a number of unique contributions.

- The RASP perturbation is a unique combination of OPE, dimensionality expansion, random noise injection, and random projection, which provides strong confidentiality guarantee.
- The proposed service constructions are able to minimize the in-house processing.
- The proposed service constructions are able to minimize the in-house processing workload because of the low perturbation cost and high precision query results. This is an important feature enabling practical cloud-based solutions.

We have assessed our scheme with real and synthetic datasets. And the obtained results show its unique advantages on all aspects of the CPEL criteria.

**III LITERATURE SURVEY**

1 “Detection and Tracking Pilot Study,”

Topic Detection and Tracking (TDT) is a DARPA-sponsored initiative to investigate the state of the art in finding and following new events in a stream of broadcast news stories. The TDT problem consists of three major tasks: (1) segmenting a stream of data, especially recognized speech, into distinct stories; (2) identifying those news stories that are the first to discuss a new event occurring in the news; and (3) given a small number of sample news stories about an event, finding all following stories in the stream. The TDT Pilot Study ran from September 1996 through October 1997. The primary participants were DARPA, Carnegie Mellon University, Dragon Systems, and the University of Massachusetts at Amherst. This report summarizes the findings of the pilot study. The TDT work continues in a new project involving larger...
training and test corpora, more active participants, and a more broadly defined notion of "topic" than was used in the pilot study.

2. Bursty and Hierarchical Structure in Streams

A fundamental problem in text data mining is to extract meaningful structure from document streams that arrive continuously over time. E-mail and news articles are two natural examples of such streams, each characterized by topics that appear, grow in intensity for a period of time, and then fade away. The published literature in a particular research field can be seen to exhibit similar phenomena over a much longer time scale. Underlying much of the text mining work in this area is the following intuitive premise --- that the appearance of a topic in a document stream is signaled by a "burst of activity," with certain features rising sharply in frequency as the topic emerges. The goal of the present work is to develop a formal approach for modeling such "bursts," in such a way that they can be robustly and efficiently identified, and can provide an organizational framework for analyzing the underlying content. The approach is based on modeling the stream using an infinite-state automaton, in which bursts appear naturally as state transitions; in some ways, it can be viewed as drawing an analogy with models from queueing theory for bursty network traffic. The resulting algorithms are highly efficient, and yield a nested representation of the set of bursts that imposes a hierarchical structure on the overall stream. Experiments with e-mail and research paper archives suggest that the resulting structures have a natural meaning in terms of the content that gave rise to them.


We are concerned with the issue of real-time change-point detection in time series. This technology has recently received vast attentions in the area of data mining since it can be applied to a wide variety of important risk management issues such as the detection of failures of computer devices from computer performance data, the detection of masqueraders/malicious executables from computer access logs, etc. In this paper we propose a new method of real-time change point detection employing the sequentially discounting normalized maximum likelihood coding (SDNML). Here the SDNML is a method for sequential data compression of a sequence, which we newly develop in this paper. It attains the least code length for the sequence and the effect of past data is gradually discounted as time goes on, hence the data compression can be done adaptively to non-stationary data sources. In our method, the SDNML is used to learn the mechanism of a time series, then a change-point score at each time is measured in terms of the SDNML code-length. We empirically demonstrate the significant superiority of our method over existing methods, such as the predictive-coding method and the hypothesis testing method, in terms of detection accuracy and computational efficiency for artificial data sets. We further apply our method into real security issues called malware detection. We empirically demonstrate that our method is able to detect unseen security incidents at significantly early stages.


Model selection by means of the predictive least squares (PLS) principle has been thoroughly studied in the context of regression model selection and autoregressive (AR) model order estimation. We introduce a new criterion based on sequentially minimized squared deviations, which are smaller than both the usual least squares and the squared prediction errors used in PLS. We also prove that our criterion has a probabilistic interpretation as a model which is asymptotically optimal within the given class of distributions by reaching the lower bound on the logarithmic prediction errors, given by the so called stochastic complexity, and approximated by BIC. This holds when the regressor (design) matrix is non-random or determined by the observed data as in AR models. The advantages of the criterion include the fact that it can be evaluated efficiently and exactly, without asymptotic approximations, and importantly, there are no adjustable hyper-parameters, which makes it applicable to both small and large amounts of data.

5. "Dynamic Syslog Mining for Network Failure Monitoring,"

Syslog monitoring technologies have recently received vast attentions in the areas of network management and network monitoring. They are used to address a wide
range of important issues including network failure symptom detection and event correlation discovery. Syslog are intrinsically dynamic in the sense that they form a time series and that their behavior may change over time. This paper proposes a new methodology of dynamic syslog mining in order to detect failure symptoms with higher confidence and to discover sequential alarm patterns among computer devices. The key ideas of dynamic syslog mining are 1) to represent syslog behavior using a mixture of Hidden Markov Models, 2) to adaptively learn the model using an online discounting learning algorithm in combination with dynamic selection of the optimal number of mixture components, and 3) to give anomaly scores using universal test statistics with a dynamically optimized threshold. Using real syslog data we demonstrate the validity of our methodology in the scenarios of failure symptom detection, emerging pattern identification, and correlation discovery.

IV IMPLEMENTATION

Modules
1. Training
2. Identify individual Anomaly Score
3. Aggregate
4. Change Point Analysis and DTO
5. Burst Detection

Training:

In this section, we describe the probability model that we used to capture the normal mentioning behavior of a user and how to train the 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 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; who use CRP for infinite vocabulary.

Aggregate:

In this subsection, we describe how to combine the anomaly scores from different users. The anomaly score is computed for each user depending on the current post of user $u$ and his/her past behavior $T^u$. To measure the general trend of user behavior, we propose to aggregate the anomaly scores obtained for posts $x_1;...,x_n$ using a discretization of window size $\lambda>0$.

Identify individual 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.

Change Point Analysis and DTO:

This technique is an extension of Change Finder proposed, 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. 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 is an approximation to the NML code length that can be computed in a sequential manner. The SDNML proposed further employs discounting in the learning of the AR models.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 proposed.In DTO, we use a one-dimensional histogram for the representation of the score distribution. We learn it in a sequential and discounting way.
Burst Detection:

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 (e.g., arrival of posts) are assumed to happen according to a time-varying Poisson processes whose rate parameter depends on the current state.

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 [3] and Kleinberg’s burst-detection model [2] 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-anomaly based 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. All the analysis presented in this paper was conducted offline, but the framework itself can be applied online. We are planning to scale up the proposed approach to handle social streams in real time. It would also be interesting to combine the proposed link-anomaly model with text-based approaches, because the proposed link-anomaly model does not immediately tell what the anomaly is. Combination of the word-based approach with the link-anomaly model would benefit both from the performance of the mention model and the intuitiveness of the word-based approach.

REFERENCES


