www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

MINING OF RECENT TRENDS IN SOCIAL NETWORK USING LINK ANOMOLY DETECTION

Mrs CHITRA.B

PG-Scholar

Computer Science & Engineering,
P.S.V College of Engineering & Technology,
Krishnagiri ,-635 108, Tamilnadu, India
Mail id : chitrakalyani10051991@gmail.com
Mobile No : +91-9445470332

Prof.S. CHANDRA SEKARAN,
Computer Science & Engineering,
P.S.V College of Engineering & Technology,
Krishnagiri ,-635 108, Tamilnadu, India
Mail id :chandrudpi@gmail.com
Mobile no :9443057461

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 these networks. Specifically, we focus on mentions of user links between users that

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

are generated dynamically (intentionally or unintentionally) through replies, mentions, and retreats. 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.

CHAPTER 1

INTRODUCTION

The news data is growing tremendously in real-time so the new concepts are getting added to the web. The focus is towards the new topics which can be discovered by mapping some of the previously discussed or published data. Social media platforms have evolved far beyond passive facilitation of online social interactions. It is the need of an hour to analyze the information content in online social media (news articles, blogs, tweets etc.). It allows business to understand public opinion about policies and products. In most of these cases,

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

data points appear as a stream of high dimensional feature vectors. We revisit the problem of online learning of topics from social media content in real-world industrial deployment scenarios. On one hand, the statistics of incoming data points is adapted by the topics dynamically and on the other h and, early detection of new trends is important in many applications. Previous methods propose online nonnegative matrix factorizations framework. This framework is mainly used to capture the evolution and emergence of themes in unstructured text under a novel temporal regularization framework. An optimization algorithm is developed for this framework and also for streaming Twitter data. Emerging themes are rapidly captured by the previous system. Also previous system can track the existing topics over time while maintaining temporal consistency and can be explicitly configured to bind the amount of information being presented to the user.

Communication over social networks, such as Facebook 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. 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.

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

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, commenting, or forwarding the information further to their friends. Conventional approaches for topic detection have mainly been concerned with the frequencies of (textual) words [1], [2]. 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.

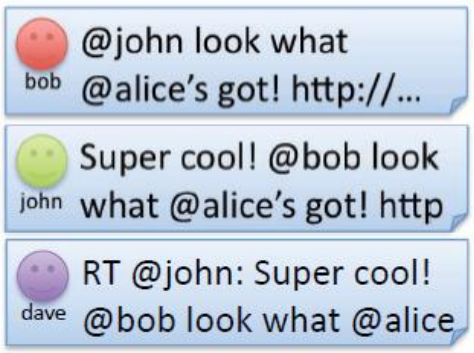

Fig.1.1. Emergence of a topic in social streams.

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

This model is 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 behaviour 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]. 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 emergence is; see Figure 2. The effectiveness of the proposed approach is demonstrated on four data sets we have collected from Twitter. We show that our mention-anomaly based approaches can detect the emergence of a new topic at least as fast as text-anomaly based counterparts. Furthermore, we show that in three out of four data sets, the proposed mention-anomaly based methods can detect the emergence of topics much earlier than the text-anomaly based methods, which can be explained by the keyword ambiguity we mentioned above.

CHAPTER 2

LITERATURE SURVEY

2.1. Topic Detection and Tracking Pilot Study Final Report

James Allan , Jaime Carbonell, George Doddington , Jonathan Yamron

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.

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

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 study is intended to explore techniques for detecting the appearance of new topics and for tracking the reappearance and evolution of them. During the first portion of this study, the notion of a "topic" was modified and sharpened to be an "event", meaning some unique thing that happens at some point in time. The notion of an event differs from a broader category of events both in spatial/temporal localization and in specificity. For example, the eruption of Mount Pinatubo on June 15th, 1991 is considered to be an event, whereas volcanic eruption in general is considered to be a class of events. Events might be unexpected, such as the eruption of a volcano, or expected, such as a political election.

2.2. Bursty and Hierarchical Structure in Streams

J. Kleinberg

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.

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

The goal of the work is to develop a formal approach for modelling 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; 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.

2.3. Real-time change-point detection using sequentially discounting normalized maximum likelihood coding

Y. Urabe, K. Yamanishi, R. Tomioka, and H. Iwai

The issue of real-time change-point detection in time series is concerned in this paper. 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.

2.4. Tracking dynamics of topic trends using a finite mixture model

S. Morinaga and K. Yamanishi

In a wide range of business areas dealing with text data streams, including CRM, knowledge management, and Web monitoring services, it is an important issue to discover topic trends and analyze their dynamics in real-time. Specifically we consider the following three tasks in topic trend analysis: 1)*Topic Structure Identification*; identifying what kinds of main topics exist and how important they are, 2)*Topic Emergence Detection*; detecting the emergence of a new topic and recognizing how it grows, 3)*Topic Characterization*; identifying the characteristics for each of main topics. For real topic analysis systems, we may require that these three tasks be performed in an on-line fashion rather than in a retrospective way, and be dealt with in a single framework. This paper proposes a new topic analysis framework which satisfies this requirement from a unifying viewpoint that a topic structure is modeled using a finite mixture model and that any change of a topic trend is tracked by learning the finite mixture model dynamically. In this framework we propose the

usage of a time-stamp based discounting learning algorithm in order to realize real-time topic structure identification. This enables tracking the topic structure adaptively by forgetting out-of-date statistics. Further we apply the theory of dynamic model selection to detecting changes of main components in the finite mixture model in order to realize topic emergence detection. We demonstrate the effectiveness of our framework using real data collected at a help desk to show that we are able to track dynamics of topic trends in a timely fashion.

2.5. Discovering evolutionary theme patterns from text: an exploration of temporal text mining

Q. Mei and C. Zhai

Temporal Text Mining (TTM) is concerned with discovering temporal patterns in text information collected over time. Since most text information bears some time stamps, TTM has many applications in multiple domains, such as summarizing events in news articles and revealing research trends in scientific literature. In this paper, we study a particular TTM task – discovering and summarizing the evolutionary patterns of themes in a text stream. We define this new text mining problem and present general probabilistic methods for solving this problem through (1) discovering latent themes from text; (2) constructing an evolution graph of themes; and (3) analyzing life cycles of themes. Evaluation of the proposed methods on two different domains (i.e., news articles and literature) shows that the proposed methods can discover interesting evolutionary theme patterns effectively.

2.6. Data Association for Topic Intensity Tracking

Andreas Krause, Jure Leskovec, Carlos Guestrin

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

A unified model of what was traditionally viewed as two separate tasks: data association and intensity tracking of multiple topics over time is presented. In the data association part, the task is to assign a topic (a class) to each data point, and the intensity tracking part models the bursts and changes in intensities of topics over time. Our approach to this problem combines an extension of Factorial Hidden Markov models for topic intensity tracking with exponential order statistics for implicit data association. Experiments on text and email datasets show that the interplay of classification and topic intensity tracking improves the accuracy of both classification and intensity tracking. Even a little noise in topic assignments can mislead the traditional algorithms. However, our approach detects correct topic intensities even with 30% topic noise.

2.7. Topic Dynamics: An Alternative Model of 'Bursts' in Streams of Topics

Dan He, D. Stott Parker

For some time there has been increasing interest in the problem of monitoring the occurrence of topics in a stream of events, such as a stream of news articles. This has led to different models of bursts in these streams, i.e., periods of elevated occurrence of events. Today there are several burst definitions and detection algorithms, and their differences can produce very different results in topic streams. These definitions also share a fundamental problem: they define bursts in terms of an arrival rate. This approach is limiting; other stream dimensions can matter. We reconsider the idea of bursts from the standpoint of a simple kind of physics. Instead of focusing on arrival rates, we reconstruct bursts as a dynamic phenomenon, using kinetics concepts from physics - mass and velocity - and derive momentum, acceleration, and force from these. We refer to the result as topic dynamics,

permitting a hierarchical, expressive model of bursts as intervals of increasing momentum. As a sample application, we present a topic dynamics model for the large PubMed/MEDLINE database of biomedical publications, using the MeSH (Medical Subject Heading) topic hierarchy. We show our model is able to detect bursts for MeSH terms accurately as well as efficiently.

2.8. Visualizing science by citation mapping

Henry Small

Science mapping is discussed in the general context of information visualization. Attempts to construct maps of science using citation data are reviewed, focusing on the use of co-citation clusters. New work is reported on a dataset of about 36,000 documents using simplified methods for ordination, and nesting maps hierarchically. An overall map of the dataset shows the multidisciplinary breadth of the document sample, and submaps allow drilling down to the document level. An effort to visualize these data using advanced virtual reality software is described, and the creation of document pathways through the map is seen as a realization of Bush's (1945) associative trails.

2.9. Naive (Bayes) at forty: The independence assumption in information retrieval

David D. Lewis

The naive Bayes classifier, currently experiencing a renaissance in machine learning, has long been a core technique in information retrieval. We review some of the variations of naive Bayes models used for text retrieval and classification, focusing on the distributional assumptions made about word occurrences in documents. 1 Introduction The naive Bayes

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

classifier, long a favorite punching bag of new classification techniques, has recently emerged as a focus of research itself in machine learning. Machine learning researchers tend to be aware of the large pattern recognition literature on naive Bayes, but may be less aware of an equally large information retrieval (IR) literature dating back almost forty years. In fact, naive Bayes methods, along with prototype formation methods, accounted for most applications of supervised learning to information retrieval until quite recently.

2.10. A unifying framework for detecting outliers and change points from time series

J. Takeuchi and K. Yamanishi

We are concerned with the issue of detecting outliers and change points from time series. In the area of data mining, there have been increased interest in these issues since outlier detection is related to fraud detection, rare event discovery, etc., while change-point detection is related to event/trend change detection, activity monitoring, etc. Although, in most previous work, outlier detection and change point detection have not been related explicitly, this paper presents a unifying framework for dealing with both of them. In this framework, a probabilistic model of time series is incrementally learned using an online discounting learning algorithm, which can track a drifting data source adaptively by forgetting out-of-date statistics gradually. A score for any given data is calculated in terms of its deviation from the learned model, with a higher score indicating a high possibility of being an outlier. By taking an average of the scores over a window of a fixed length and sliding the window, we may obtain a new time series consisting of moving-averaged scores. Change point detection is then reduced to the issue of detecting outliers in that time series. We

compare the performance of our framework with those of conventional methods to demonstrate its validity through simulation and experimental applications to incidents detection in network security.

CHAPTER 3

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 preprocessing (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 preprocessing to obtain (the information is often separated from the contents), and are available regardless of the nature of the contents.

CHAPTER 4

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

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

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.

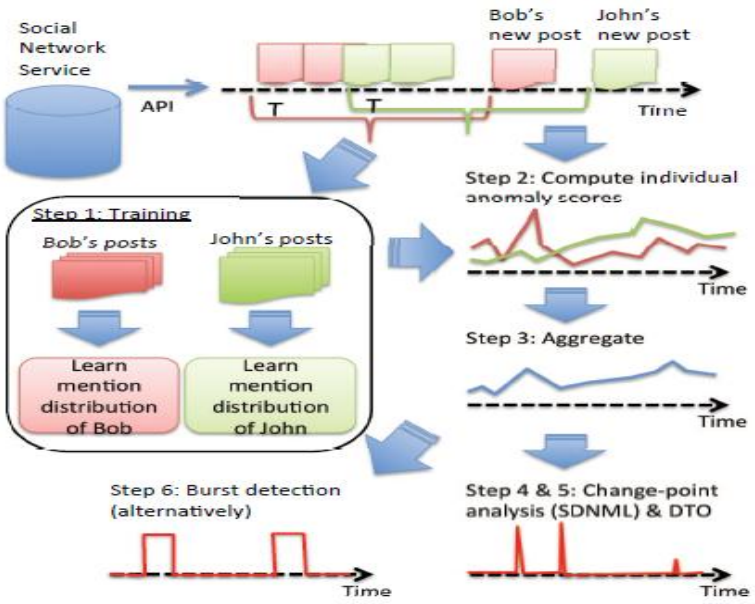

Fig.4.1. Proposed System

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. Recently, He and Parker developed a 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. 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.

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.

CHAPTER 5

SYSTEM SPECIFICATION

5.1. Hardware Requirement:

Processor	:	Pentium IV 2.4 GHz
Hard Disk	:	250 GB
Monitor	:	15 VGA Colour

RAM : 2 MB

5.2. Software Requirement:

Operating System	:	Windows XP/7
Coding Language	:	JAVA/J2EE
IDE	:	Netbeans 7.4
Database	:	MySQL

CHAPTER 6

SYSTEM DESIGN

6.1. System Architecture:

In this system, 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 hundreds of users and apply a recently proposed change point detection technique based on the sequentially discounting normalized maximum-likelihood (SDNML) coding. 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 in this system. The effectiveness of the proposed approach is demonstrated on four data sets we have collected from Twitter. We show that our mention-

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

anomaly-based approaches can detect the emergence of a new topic at least as fast as text-anomaly based counterparts.

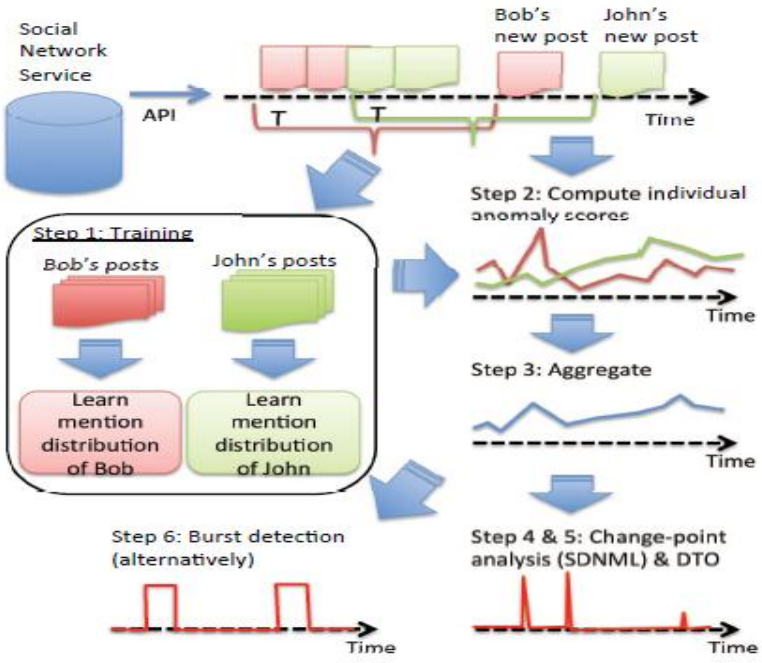

Fig.6.1. System Architecture

Furthermore, we show that in three out of four data sets, the proposed mention-anomaly based methods can detect the emergence of topics much earlier than the text-anomaly-based methods, which can be explained by the keyword ambiguity we mentioned above. 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-

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

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 nontextual 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. 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

6.2. Data Flow Diagram:

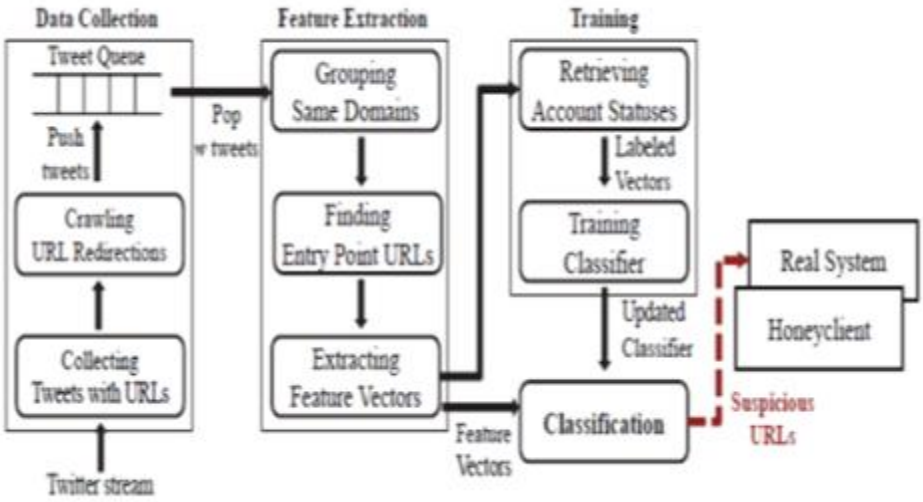

Fig.6.2. Data Flow Diagram

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

The anomalies are detected from each news class. That anomaly is mapped to the news which has same set of tokens but is not identified as anomalies. The TF-IDF values of all tokens are calculated in the mapped news which contains anomaly as well as the News class which has the same set of words. It is a numerical evaluation of how important a word is to a document in a corpus. It calculates possibility value of the number of times a word appears in the document. After that taking mean of all tokens in the mapped news. The mean is taken as threshold value. If token tf-idf value is greater than threshold then it is considered as most probable token and these tokens are ranked for further processing.

CHAPTER 7

SYSTEM IMPLEMENTATION

7.1. Modules Description:

7.1.1. Twitter Trends:

- ✓ First we design the system using Key-word based detection and Link-based detection. We collected real time data sets from Twitter.
- ✓ Each data set is associated with a list of posts in a service.
- ✓ Also it's a collaborative service where people can tag Twitter posts that are related to each other and organize a list of posts that belong to a certain topic.
- ✓ Our goal is to evaluate whether the proposed approach can detect the emergence of the topics recognized and collected by people. For each list, we extracted a list of Twitter users that appeared in the list, and collected Twitter posts from those users.

www.ijrream.in Impact Factor: 2.9463
INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING AND MANAGEMENT (IJRREM)
Tamilnadu-636121, India
Indexed by
Scribd impact Factor: 4.7317, Academia Impact Factor: 1.1610
ISSN NO (online) : Application No : 17320 RNI –Application No 2017103794

7.1.2. 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.

7.1.3. Aggregate:

- ✓ In this module, 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 Ttu.
- ✓ To measure the general trend of user behavior, we propose to aggregate the anomaly scores obtained for posts x1;...;x xn using a discretization of window size $\lambda>0$.

- ✓ Also, we assign an anomaly score to each post based on the learned probability distribution

7.1.4. Change Point Analysis:

- ✓ 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.
- ✓ 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.

7.1.5. 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.

CHAPTER 8

CONCLUSION

Recently it is found that the discovering of news topics is challenging task and has much importance in data mining fields. In this paper, a new approach is used to detect the emergence of topics in a social network stream. The basic idea is to focus on anomaly detection in news class. Anomalies are detected and then mapped to the news class. After mapping these anomalies, a new concept is generated. Further it has application in forensic analysis to determine the new stories around a topic. So it will always be research field for future researchers.

CHAPTER 9

FUTURE ENHANCEMENT

In our data-driven approach, the input will have all the metadata about the search keywords includes user_name, retweet_count, hash_tags and etc., In earlier approach trends are detected based on the timely active which indicate the trending topics. Instead predicting trends based on the time alone we are using the metadata of the topic.

Trends are decided by the same topic actively used by people in the social network, Here instead indirectly capturing based on the time it will more precise to consider the retweet count of the each tag. For Example many of your friends are friends with person X, it is likely that you are also friend with X. So if Person X retweets for a topic it is directly proportional to active friends for chance of retweet if he/she likes it.

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, "Real-time 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 Probabilit´es de Saint-Flour XIII—1983*. Springer, 1985, pp. 1–198.

[10]. Y. Teh, M. Jordan, M. Beal, and D. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101,no. 476, pp. 1566–1581, 2006.

[11]. D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in *Proceedings of ECML-98, 10th European Conference on Machine Learning*. Springer, 1998, pp. 4–15.

[12]. K. Yamanishi and J. Takeuchi, "A unifying framework for detecting outliers and change points from non-stationary time seriesdata," in *Proceedings of the 8th ACM SIGKDD*, 2002.

[13]. 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.

[14]. 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.

[15]. T. Roos and J. Rissanen, "On sequentially normalized maximum likelihood models," in *Workshop on information theoretic methods in science and engineering*, 2008.

[16]. J. Rissanen, T. Roos, and P. Myllym¨aki, "Model selection by sequentially normalized least squares," *Journal of Multivariate Analysis*, vol. 101, no. 4, pp. 839–849, 2010.

[17]. C. Giurc˘*a*neanu, S. Razavi, and A. Liski, "Variable selection in linear regression: Several approaches based on normalized maximum likelihood," *Signal Processing*, vol. 91, pp. 1671–1692, 2011.

[18]. 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.

[19]. K. Yamanishi and Y. Maruyama, "Dynamic syslog mining for network failure monitoring," *Proceeding of the 11th ACM SIGKDD*, p. 499, 2005.