

International Journal of Computer Science and Mobile Computing



A Monthly Journal of Computer Science and Information Technology

ISSN 2320-088X

IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 5, May 2016, pg.416 – 426

Exploiting Emerging Topics in Social Networking Sites for Textual and Graphical Stream

Bhawana Sarode¹, Prof. P.L.Ramteke²

^{1,2} Department of Computer Science & Information Technology HVPMS COET, SGBAU, Amravati (MH), India
bhawana.sarode12@gmail.com¹, pl_ramteke@rediff.com²

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 users—links between users that are generated dynamically (intentionally or unintentionally) through replies, mentions, and retweet. 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.

1. INTRODUCTION

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 users—links between users that are generated dynamically (intentionally or unintentionally) through replies, mentions, and re-tweet. 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.

2. LITERATURE SURVEY

S.Saranya et. al. identifies various concepts involved in social networks for finding the emerging topics. They focused on the various methods that can be applied for detecting the anomaly. The methods used are Hidden Markov Model, UMass Approach, CMU Approach, Change Finder method and Finite Mixture Model. These methods involve texts, videos, audios, URLs and mentions which are shared in the social networks. Kullback-Leibler divergence measure is used here to discover coherent themes and topics over time. [1]

[2]Link Anomaly detection is one of the most important topics in social network. Many of the social networks such as Facebook, Google+, LinkedIn, or twitter require an effective and efficient framework to identify deviated data. Anomaly detection methods are typically implemented in social stream mode, and thus cannot be easily extended to large-scale problems without sacrificing computation where the user's link is generated dynamically (replies, mentions, and re-tweets). A new approach model i.e. probability model, this model to the capture normal linking behavior of a social network users, and propose to detect the trending topic from

the social networks through the probability model. We collect anomaly score from the different user. And aggregated score feed to change-point analysis or change-point detection, or with burst detection, finally show that to detect trending topics only based on the reply/mention in social network posts. Our technique to collect number of real data from real time twitter account.

[3]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 purpose of the Topic Detection and Tracking (TDT) Pilot Study is to advance and accurately measure the state of the art in TDT and to assess the technical challenges to be overcome. At the beginning of this study, the general TDT task domain was explored and key technical challenges were clarified. This document defines these tasks, the performance measures to be used to assess technical capabilities and research progress, and presents the results of a cooperative investigation of the state of the art.

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; 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. [4]

[5] Text streams often contain latent temporal theme structures which reflect how different themes influence each other and evolve over time. Discovering such evolutionary theme patterns can not only reveal the hidden topic structures, but also facilitate navigation and digestion of information based on meaningful thematic threads. In this paper, we propose general probabilistic approaches to discover evolutionary theme patterns from text streams in a completely unsupervised way. To discover the evolutionary theme graph, our method would first generate word clusters (i.e., themes) for each time period and then use the Kullback-Leibler divergence measure to discover coherent themes over time. Such an evolution graph can reveal how themes change over time and how one theme in one time period has influenced other themes in later periods. We also propose a method based on hidden Markov models for analyzing the life cycle of each theme. This method would first discover the globally interesting themes and then compute the strength of a theme in each time period. This allows us to not only see the trends of strength variations of themes, but also compare the relative strengths of different themes over time.

We evaluated our methods using two different data sets. One is a stream of 50 days' news articles about the tsunami disaster that happened recently in Asia, and the other is the abstracts of the KDD conference proceedings from 1999 to 2004. In both cases, the proposed methods can generate meaningful temporal theme structures and allow us to summarize and analyze the text data from temporal perspective. Our methods are generally applicable to any textstream data and thus have many potential applications in temporal text mining.

There are several interesting directions to further extend this work. First, we have only considered flat structure of themes; it would be interesting to explore hierarchical theme clustering, which can give us a picture of theme evolutions at different resolutions. Second, we

can develop a temporal theme mining system based on the proposed methods to help a user navigate the stream information space based on evolutionary structures of themes. Such a system can be very useful for managing all kinds of text stream data. Finally, temporal text mining (TTM) represents a promising new direction in text mining that has not yet been well-explored. In addition to evolutionary theme patterns, there are many other interesting patterns such as associations of themes across multiple streams that are interesting to study.

3. SYSTEM ANALYSIS

3.1 Existing System

The current system identifies various concepts involved in social networks for finding the emerging topics. It focus on the various methods that can be applied for detecting the anomaly. The methods used are Hidden Markov Model, UMass Approach, CMU Approach, Change Finder method and Finite Mixture Model. These methods involve texts which are shared in the social networks. Kullback-Leibler divergence measure is used here to discover coherent themes and topics over time

The current model i.e. probability model, this model is capture normal linking behavior of a social network users, and predict to detect the trending topic from the social networks through the probability model. It collect anomaly score from the different user. And aggregated score feed to change-point analysis or change-point detection, or with burst detection, finally show that to detect trending topics only based on the reply/mention in social network posts.

3.2 Existing System Architecture

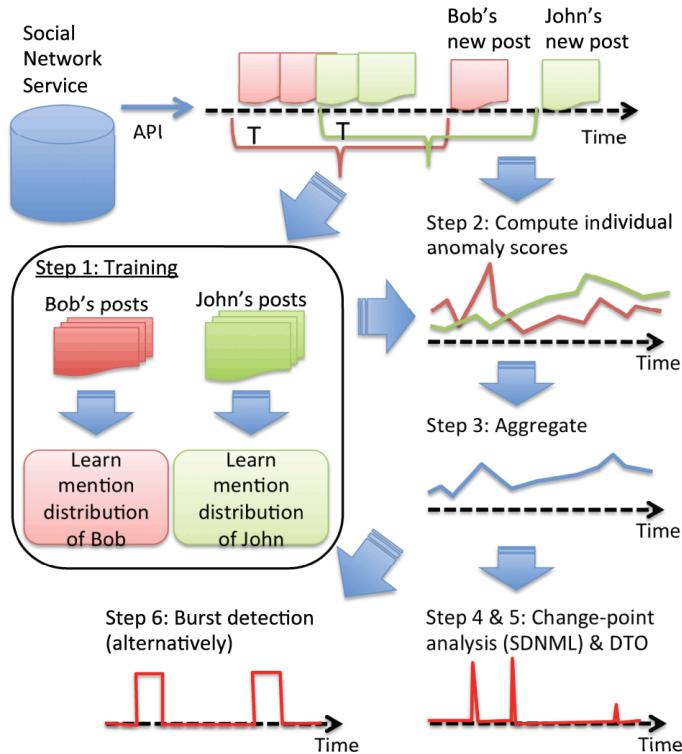


Fig: Existing System Architecture

3.3 Existing Algorithms

- 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 in Fig. 2. 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)..

- 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 in Fig. 2.

- Combining Anomaly Scores from Different Users

In this subsection, we describe how to combine the anomaly scores from different users; see Step 3 in Fig. 2.

- Check Point

In this subsection, we describe how to detect change points from the sequence of aggregated anomaly scores; see Step 4 in Fig. 2.

- 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 proposed in [19]; see Step 5 in Fig. 2.

4. PROPOSED ALGORITHMS

Hits Algorithm:

Hyperlink-Induced Topic Search (HITS; also known as **hubs and authorities**) is a link analysis algorithm that rates links. The idea behind Hubs and Authorities stemmed from a particular insight into the creation of web pages when the Internet was originally forming; that is, certain web pages, known as hubs, served as large directories that were not actually authoritative in the information that they held, but were used as compilations of a broad catalog of information that led users direct to other authoritative pages. In other words, a good hub represented a page that pointed to many other pages, and a good authority represented a page that was linked by many different hubs.

```
1 G := set of pages
2 for each page p in G do
3   p.auth = 1 // p.auth is the authority score of the page p
4   p.hub = 1 // p.hub is the hub score of the page p
```

```
5 functionHubsAndAuthorities ( $G$ )
6   for step from 1 to k do // run the algorithm for k steps
7     norm = 0
8     for each page  $p$  in  $G$ do // update all authority values first
9        $p.auth = 0$ 
10      for each page  $q$  in  $p.incomingNeighbors$ do //  $p.incomingNeighbors$ 
is the set of pages that link to  $p$ 
11         $p.auth += q.hub$ 
12      norm += square( $p.auth$ ) // calculate the sum of the squared auth
values to normalise
13      norm = sqrt(norm)
14      for each page  $p$  in  $G$ do // update the auth scores
15         $p.auth = p.auth / norm$  // normalise the auth values
16      norm = 0
17      for each page  $p$  in  $G$ do // then update all hub values
18         $p.hub = 0$ 
19        for each page  $r$  in  $p.outgoingNeighbors$ do //  $p.outgoingNeighbors$ 
is the set of pages that  $p$  links to
20           $p.hub += r.auth$ 
21        norm += square( $p.hub$ ) // calculate the sum of the squared hub
values to normalise
22        norm = sqrt(norm)
23        for each page  $p$  in  $G$ do // then update all hub values
24           $p.hub = p.hub / norm$  // normalise the hub values
```

5. IMPLEMENTATION

- Probability Model:

Fetching the post or comment.

- Computing the Link-Anomaly Score:

Scoring that post or comment.

- Combining Anomaly Scores from Different Users

Scoring for each user.

- Dynamic Threshold Optimization (DTO):

Finding post according to visiting percentage

These all procedure is for text and graphical base post.

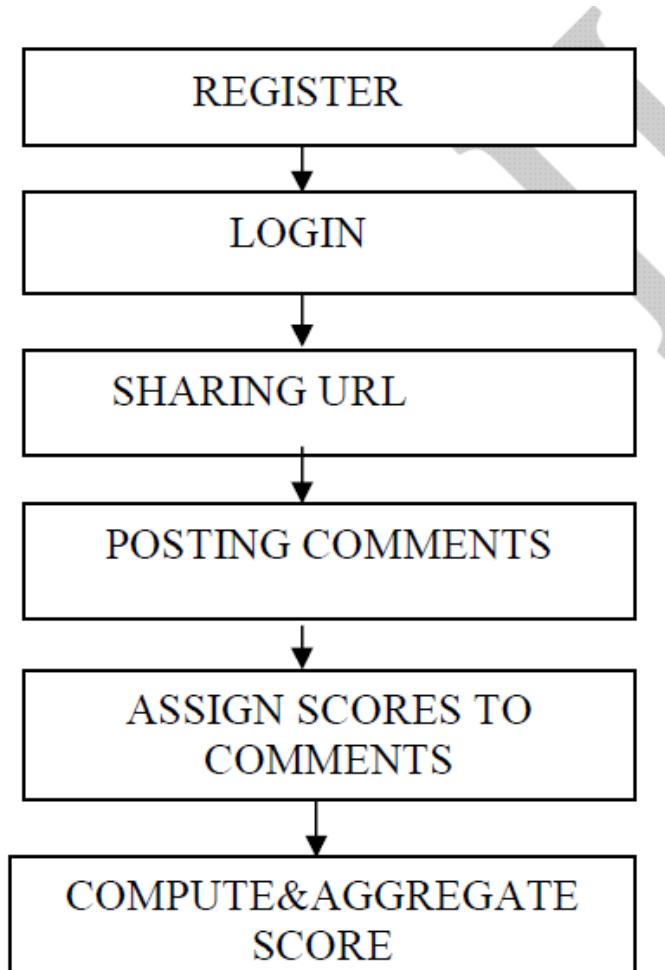


Fig: Emerging new topic

6. CONCLUSION

In this seminar, 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 mentioned. 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.

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," _ Ecole d' _ Ete' de Probabilites de 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

AUTHORS PROFILE

	<p>Bhawana Sarode received the B.E.(I.T) and M.E. degrees in Computer science and information technology from HVPM COET Amravati respectively During 2014-2016,</p>
---	---

	<p>Prof. P.L.Ramteke is Associate Professor & Head of Department of Information Technology. He has completed Bachelor & Master Degree of Engineering in Computer Science & Engineering from SGB Amravati University Amravati.</p>
--	--