ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 5, May 2016

Controversy Trend of Anomaly Detection for Discovering Emerging Topics

BHAWANA SARODE[1], PROF . P.L.RAMTEKE[2]

**[1] Department of Computer science & Information Technology HVPM's COET,
SGBAU, Amravati (MH),India .**

**[2] Department of Computer science & Information Technology Engineerin HVPM's COET,
SGBAU, Amravati (MH), India.**

ABSTRACT **-**Detecting and generating new concepts has attracted much attention in data mining era, nowadays. The emergence of new topics in news data is a big challenge. The problem can be extended as *"finding breaking news"*. Years ago the emergence of new stories were detected and followed up by domain experts. 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. 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.

1.INTRODUCTION

Online social networks (OSNs) (e.g., Facebook, Twitter) are now among the most popular sites on the Web. An OSN provides a powerful means of establishing social connections and sharing, organizing, and finding content[8]. For example, Facebook presently has over 500 million users. Unlike current file or video sharing systems (e.g., BitTorrent and YouTube), which are mainly organized around content, OSNs are organized around users. OSN users establish friendship relations with realworld friends or virtual friends, and post their profiles and content such as photos, videos, and notes to their personal pages[9]. 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[7]. 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 [1], [12]. 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[11]. 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[10]. 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[5]. The main aim of this project is to detect emerging topics as early as the keyword-based methods and evaluate whether the proposed approach can detect the emergence of the topics recognized and collected by people[17].

1.1 SYSTEM ARCHITECTURE DESIGN

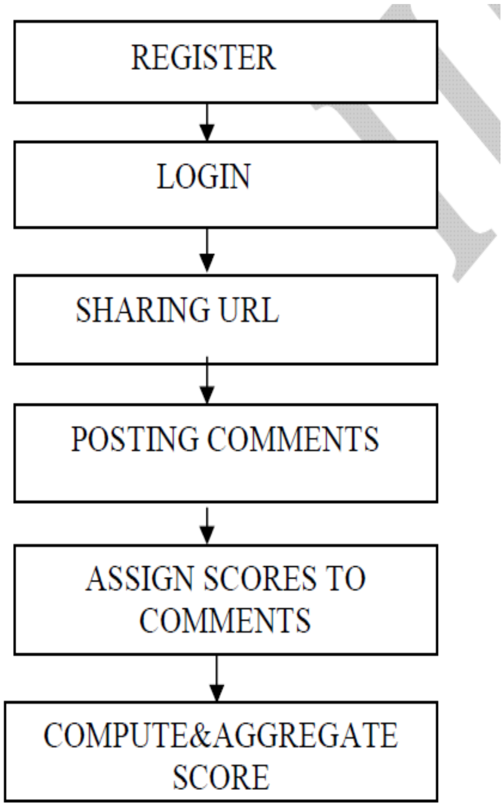

Fig.1 Emerging New Topics

2.DETECTION OF EMERGING TOPICS

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. Compared to conventional media, social media are able to capture the earliest, unedited voice of ordinary people[11]. 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. Mentions links to other users of the same social network in and post their profiles and content such as photos, videos, and notes to their personal pages. 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[6]. Compared to conventional media, social media are able to capture the earliest, unedited voice of ordinary people[16]. Therefore, the challenge is to detect the emergence of a topic as early as possible at a moderate number of false positives. We are the form of message-to, reply-to, retweet-of, or explicitly in the text[13]. 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. The main aim of this project is to detect emerging topics as early as the keyword-based methods and evaluate whether the proposed approach can detect the emergence of the topics recognized and collected by people[18].

3. PROBABILITY MODEL

Probability model used to capture the normal mentioning behavior of a user and describes how to train the model. Also characterized 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 taken into accounhere. 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, this system would like to avoid putting an artificial limit on the number of users mentioned in a post. Instead, 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[15].

4. Computing the link-anomaly score

This subsection describes how to compute the deviation of a user's behavior from the normal mentioning behaviour modeled 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, and compute the probability 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 predictive distribution of the number of mentions, and the predictive distribution of the mentionee. This section describes 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 (t) u. In order to measure the general trend of user behavior, aggregated anomaly scores obtained for posts x1 . . . xn using a discretization of window size $\tau > 0$[4].

5. BURST DETECTION METHOD

In addition to the change-point detection based on SDNML followed by DTO was described; also test the combination of our method with Kleinberg's burst detection method. More specifically, implemented a two-state version of Kleinberg's burst detection model. The reason to chose the two-state version was because in this experiment 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.

6. SCALABILITY OF THE PROPOSED ALGORITHM

The proposed link-anomaly based change-point detection is highly scalable. Every step described in the previous subsections requires only linear time against the length of the analyzed time period. Computation of the predictive distribution for the number of mentions can be computed in linear time against the number of mentions. Computation of the predictive distribution for the mention probability in and can be efficiently performed using a hash table[14]. Aggregation of the anomaly scores from different users takes linear time against the number of users, which could be a computational bottle neck but can be easily parallelized. SDNML- based change-point detection

requires two swipes over the analyzed time period. Kleinberg's burst detection method can be efficiently implemented with dynamic programming.

7. LITERATURE SURVEY

SDNML Change point analysis model 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. 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. 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. Dynamic Threshold Optimization need to convert the changepoint 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[19].

8.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 also studied the survey of outlier/Anomaly linking for discovering emerging topics in news datasets. We further analyzed different techniques used for topic classification, Outlier Detection, Concept Detection and Concept generations. 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[20]. We have proposed a probability model that captures both the number of mentions per post and the frequency of mentionee. 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.

9.REFERENCES

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

[2] J. Takeuchi and K. Yamanishi, "A Unifying Framework for Detecting Outliers and Change Points from Time Series," IEEE Trans. Knowledge Data Eng., vol. 18, no. 4, pp. 482-492, Apr. 2006.

[3] J. Rissanen, "Strong Optimality of the Normalized ML Models as Universal Codes and Information in Data," IEEE Trans. Information Theory, vol. 47, no. 5, pp. 1712-1717, July 2001.

[4] T. Roos and J. Rissanen, "On Sequentially Normalized Maximum Likelihood Models," Proc. Workshop Information Theoretic Methods in Science and Eng., 2008.

[5] J. Rissanen, T. Roos, and P. Myllyma¨ki, "Model Selection by Sequentially Normalized Least Squares," J. Multivariate Analysis, vol. 101, no. 4, pp. 839-849, 2010.

[6] C. Giurc_aneanu, 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.

[7] 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.

[8] Social media, web 2.0 and internet stats. http://thefuturebuzz.com/2009/01/12/social-media-web-20-
internet-numbers-stats/.

[9] Y. Huang, Z. Fu, D. Chiu, C. Lui, and C. Huang. Challenges, design and analysis of a large-scale P2P VoD system. In *Proc. SIGCOMM*, 2008. *TOMCCAP*, 2008.

[10] C.-P. Ho, S.-Y. Lee, and J.-Y. Yu. Cluster-based replication for P2P-based video-on-demand service. In *Proc. of ICEIE*, 2010.

[11] J. Wang, C. Huang, and J. Li. On ISP-friendly rate allocation for peer-assisted VoD. In *Proc. of MM*, 2008.

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

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

[14] 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' ,2011.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] H. Small, "Visualizing Science by Citation Mapping," J. Am. Soc.

10.AUTHORS PROFILE

	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,

	Prof. P.L.Ramteke[1] 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.