International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue 14
November 2017

Burstyn Topic Detection from Twitter Data Using Topicsketch

VELDI NIKHITHA **Dr. RAVISANKAR MALLADI**

PG Scholar, Department of CSE, Vaagdevi College of Engineering, Bollikunta, Warangal, Telangana, Mail id: veldinikhitha93@gmail.com

Professor, Department of CSE, Vaagdevi College of Engineering, Bollikunta, Warangal, Telangana, Mail id:drmrsankar@gmail.com

Abstract—Twitter has become one of the largest platforms for users around the world to share anything happening around them with friends and beyond. A bursty topic in Twitter is one that triggers a surge of relevant tweets within a short time, which often reflects important events of mass interest. How to leverage Twitter for early detection of bursty topics has therefore become an important research problem with immense practical value. Despite the wealth of research work on topic modeling and analysis in Twitter, it remains a huge challenge to detect bursty topics in real-time. As existing methods can hardly scale to handle the task with the tweet stream in real-time, we propose in this paper TopicSketch, a novel sketch-based topic model together with a set of techniques to achieve real-time detection. We evaluate our solution on a tweet stream with over 30 million tweets. Our experiment results show both efficiency and effectiveness of our approach. Especially it is also demonstrated that TopicSketch can potentially handle hundreds of millions tweets per day which is close to the total number of daily tweets in Twitter and present bursty event in finer-granularity.

I. INTRODUCTION

With 200 million active users and over 400 million tweets per day as in a recent report 1 , Twitter has become one of the largest information portals which provides an easy, quick and reliable platform for ordinary users to share anything happening around them with friends and other followers. In particular, it has been observed that, in life-critical disasters of societal scale, Twitter is the most important and timely source from which people find out and track the breaking news before any mainstream media picks up on them and rebroadcast the footage. For

example, in the March 11, 2011 Japan earthquake and subsequent tsunami, the volume of tweets sent spiked to more than 5,000 per second when people post news about the situation along with uploads of mobile videos they had recorded 2 . We call such events which trigger a surge of a large number of relevant tweets "bursty topics". year-old girl from Singapore named Adelyn (not her real name) caused a massive uproar online after she was unhappy with her mom's incessant nagging and resorted to physical abuse by slapping her mom twice, and boasted about her actions on Facebook with vulgarities. Within hours, it soon went viral on the Internet, trending worldwide on Twitter and was one of the top Twitter trends in Singapore. For many bursty events like this, users would like to be alerted as early as it starts to grow viral to keep track. However, it was only after almost a whole day that the first news media report on the incident came out. In general, the sheer scale of Twitter has made it impossible for traditional news media, or any other manual effort, to capture most of such bursty topics in real-time even though their reporting crew can pick up a subset of the trending ones. This gap raises a question of immense practical value: Can we leverage Twitter for automated real-time bursty topic detection on a societal scale? Unfortunately, this real-time task has not been solved by the existing work on Twitter topic analysis. First of all, Twitter's own trending topic list does not help much as it reports mostly those all-time popular topics, instead of the bursty ones that are of our interest in this work. Secondly, most prior research works study the topics in Twitter in a retrospective off-line manner, e.g., performing topic modeling, analysis and tracking for all tweets generated in a certain time period]. While these findings have offered interesting insight into the topics, it is our belief that the greatest values of Twitter bursty topic detection has yet to be brought out, which is to detect the bursty topics just in time as they are taking place. This real-time task is prohibitively challenging for existing algorithms because of the high computational complexity inherent in the topic models as well as the ways in which the topics are usually learnt, e.g., Gibbs Sampling [11] or variational inference [3]. The key research challenge that makes this problem difficult is how to solve the following two problems in real-time: (I) How to efficiently maintain proper statistics to trigger detection; and (II) How to model bursty topics without the chance to examine the entire set of relevant tweets as in

traditional topic modeling. While some work such as [24] indeed detects events in real-time, it requires pre-defined keywords for the topics. We propose a new detection framework called TopicSketch. To our best knowledge, this is the first work to perform real-time bursty topic detection in Twitter without pre-defined topical keywords. It can be observed from Figure 1 that TopicSketch is able to detect this bursty topic soon after the very first tweet about this incident was generated, just when it started to grow viral and much earlier than the first news media report. We summarize our contributions as follows. First, we proposed a two-stage integrated solution TopicSketch. In the first stage, we proposed a novel data sketch which efficiently maintains at a low computational cost the acceleration of three quantities: the total number of all tweets, the occurrence of each word and the occurrence of each word pair. These accelerations provide as early as possible the indicators of a potential surge of tweet popularity. They are also designed such that the bursty topic inference would be triggered and achieved based on them. The fact that we can update these statistics efficiently and invoke the more computationally expensive topic inference part only when necessary at a later stage makes it possible to achieve real-time detection in a data stream of Twitter scale. In the second stage, we proposed a sketch-based topic model to infer both the bursty topics and their acceleration based on the statistics maintained in the data sketch. Secondly, we proposed dimension reduction techniques based on hashing to achieve scalability and, at the same time, maintain topic quality with proved error bounds. Finally, we evaluated TopicSketch on a tweet stream containing over 30 million tweets and demonstrated both the effectiveness and efficiency of our approach. It has been shown that TopicSketch is able to potentially handle over 300 million tweets per day which is almost the total number of tweets generated daily in Twitter. We also presented case studies on interesting bursty topic examples which illustrate some desirable features of our approach, e.g., finergranularity event description.

II. RELATED WORK

While this work is the first to achieve real-time bursty event detection in Twitter without pre-defined keywords, related work can be grouped into three categories. Offline. In this category, it is assumed that there is a retrospective view of the data in its entirety. There has been a stream of research studies to learn topics offline from a text

corpus, from the standard topic models such as PLSA [14] and LDA [3], to a number of temporal topic models such as Since all these models learn topics off-line, they are not able to detect at an early stage the new bursty topics that are previously unseen and just started to grow viral. When it comes to finding bursts from data stream in particular, [18] proposed a state machine to model the data stream, in which bursts appear as state transitions. [16] proposed another solution based on a time-varying Poisson process model. Instead of focusing on arrival rates, [12] reconstructed bursts as a dynamic phenomenon using acceleration and force to detect bursts. Other off-line bursty topic modeling works include most noticeably. While MemeTracker [19] is an influential piece of work which gives an interesting characterisation of news cycle, it is not designed to capture bursty topics on the fly in Twitterlike setting as it is hard to decide what the meme of tweets are. Online. In this category, certain data structure is built based on some inherent granularity defined on the data stream. Detection is made by using the data structure of all data arriving before the detection point but none after. Some works make effort on the online learning of topics [2], [6], [13], while others focus on Topic Detection and Tracking (TDT) such as [1] and [5]. Yet these solutions do not scale to the overwhelming data volume like that of Twitter. In particular, [22] makes use of locality-sensitive hashing (LSH) to reduce time cost. However, even with LSH, the computational cost is huge to calculate, for each arriving tweet, the distances between this tweet and all previous tweets colliding with this tweet in LSH. Twevent [20] is the state-of-the-art system detecting events from tweet stream. The design of Twevent takes an inherent time window of fixed size (e.g., one day) to find bursty segments of tweets, falling short of the full dynamicity essential to the real-time detection task. Real-time. In this category, time is crucial, so much so that no fixed time window for detection should be assumed. While [24] does detect events in real-time, it needs predefined keywords for the topic, making it inapplicable to general bursty topic detection where no prior knowledge of the topic keywords is available. Besides, there are also works on finding frequent items from large data stream with small memory such as CountMin Sketch [8] among others including Our TopicSketch deals with a very different and harder problem of bursty topics.

III. SOLUTION OVERVIEW

A. Problem Formulation We first formulate our real-time Twitter bursty topic detection problem. In defining a bursty topic, we evaluate two criteria: (I) There has to be a sudden surge of the topic's popularity which is measured by the total number of relevant tweets. Those all-time popular topics therefore would not count; (II) The topic must be reasonably popular. This would filter away the large number of trivial topics which, despite the spikes in their popularity, are considered as noises because the total number of relevant tweets is neglectable. Denote D(t) as the set of all tweets generated in the tweet stream up to a given timestamp t. Each tweet d ∈ D(t) is represented as a bag of words denoted as a vector $\{d(i)\}1\leq i\leq N$ where d(i) is the number of appearance of word i in d and N is the size of the vocabulary. Each tweet d is associated with the timestamp of its generation denoted as td. We use |d| to denote the number of words in tweet d.

IV CONCLUSION

In this paper, we proposed TopicSketch a framework for real-time detection of bursty topics from Twitter. Due to the huge volume of tweet stream, existing topic models can hardly scale to data of such sizes for real-time topic modeling tasks. We developed a novel concept of "Sketch", which provides a "snapshot" of the current tweet stream and can be updated efficiently. Once burst detection is triggered, bursty topics can be inferred from the sketch.

REFERENCES

[1] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In SIGIR, pages 37–45, 1998.

[2] L. AlSumait, D. Barbara, and C. Domeniconi. On-line lda: adaptive topic ´ models for mining text streams with applications to topic detection and tracking. In ICDM, 2008.

[3] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. the Journal of machine Learning research, 3:993–1022, 2003.

[4] D. M. Blei and J. D. Lafferty. Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning, pages 113–120, 2006.

[5] T. Brants, F. Chen, and A. Farahat. A system for new event detection. In SIGIR, pages 330–337, 2003.

[6] K. R. Canini, L. Shi, and T. L. Griffiths. Online inference of topics with latent dirichlet allocation. In Proceedings of the International Conference on Artificial

Intelligence and Statistics, volume 5, pages 65–72, 2009.

[7] G. Cormode and S. Muthukrishnan. What's hot and what's not: tracking most frequent items dynamically. In Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pages 296–306, 2003.

[8] G. Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. Journal of Algorithms, 55(1):58–75, 2005.

[9] Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim. Finding bursty topics from microblogs. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, pages 536–544, 2012.

[10] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu. Parameter free bursty events detection in text streams. In Proceedings of the 31st international conference on Very large data bases, pages 181–192, 2005.

[11] T. Griffiths and M. Steyvers. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America, 101(Suppl 1):5228–5235, 2004.

[12] D. He and D. Parker. Topic dynamics: an alternative model of bursts in streams of topics. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 443–452, 2010.

[13] M. D. Hoffman, D. M. Blei, and F. Bach. Online learning for latent dirichlet allocation. Advances in Neural Information Processing Systems, 23:856–864, 2010.

[14] T. Hofmann. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 50–57, 1999.

[15] L. Hong, B. Dom, S. Gurumurthy, and K. Tsioutsiouliklis. A time-dependent topic model for multiple text streams. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 832– 840, 2011.

[16] A. Ihler, J. Hutchins, and P. Smyth. Adaptive event detection with time-varying poisson processes. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 207–216, 2006.

[17] C. Jin, W. Qian, C. Sha, J. Yu, and A. Zhou. Dynamically maintaining frequent items over a data stream. In Proceedings of the twelfth international conference on Information and knowledge management, pages 287–294, 2003.

[18] J. Kleinberg. Bursty and hierarchical structure in streams. Data Mining and Knowledge Discovery, 7(4):373–397, 2003.

[19] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 497–506, 2009.

[20] C. Li, A. Sun, and A. Datta. Twevent: segment-based event detection from tweets. In Proceedings of the 21st ACM international conference on Information and knowledge management, pages 155–164, 2012

AUTHOR'S PROFILE:

VELDI NIKHITHA

PG Scholar, Department of CSE, Vaagdevi College of Engineering, Bollikunta, Warangal, Telangana, Mail id: veldinikhitha93@gmail.com

Dr. RAVISANKAR MALLADI

Professor, Department of CSE, Vaagdevi College of Engineering, Bollikunta, Warangal, Telangana, Mail id:drmrsankar@gmail.com