International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 07
June 2017

Topic Sketch: Real-time Bursty Topic Detection from Twitter

1. S.SIREESHA, 2. B. PRUDHVI, 3.DR. K NAGESWARARAO

[1]Pg Scholar, Department of CSE, Mother Teresa Institute of Science and Technology, Sathupally
sireesha307@gmail.com
[2]Assistant Professor , Department of CSE, Mother Teresa Institute of Science and Technology, Sathupally
prithvireddyb@gmail.com
[3] Professor & HOD, Department of CSE, Mother Teresa Institute of Science and Technology, Sathupally
nageswararaokapu@yahoo.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 We call such events which trigger a surge of a large

number of relevant tweets "bursty topics" This work was done when the author was visiting Living Analytics

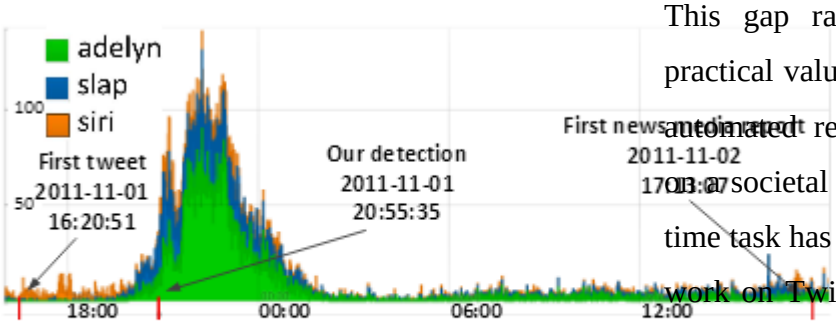

Fig. 1. The tweet volume of each of the top three keywords of the topic: "adelyn", "slap" and "siri".

Figures 1 shows an example of a bursty topic on Novem- ber 1st, 2011. A 14-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 at 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 ret1respective 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 or variational inference

. 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 TopicS-ketch.To our best knowledge, this is the first work to per- form real-time bursty topic detection in Twitter without pre- defined topical keywords.

It can be observed from Figures 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., finer granularity event description

III 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 and LDA 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, proposed a state machine to model the data stream, in which bursts appear as state transitions.proposed another solution based on a time-varying Poisson process model. Instead of focusing on arrival rates, 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 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 Twitter- like 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 while others focus on Topic Detection and Tracking (TDT) such as and . Yet these solutions do not scale to the overwhelming data volume like that of Twitter. In particular, 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 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.

IV Intuition

Actually, as mentioned in , the term "bursty topic" is very ambiguous, and can be viewed in very different ways. Various intuitions and corresponding definitions on it lead to diverse solutions The intuition behind this work comes from the observation that, the whole tweet stream is full of large amount of tweets about general topics such as car, music and food. Although they take a large proportion in the whole tweet stream, they are not helpful for our bursty topic detection task. Therefore, we try to separate the bursty topics from them. We found that, following daily routine, people usually tweet about general topics in a steady pace. In contrast, bursty topics are often triggered by some events such as some breaking news or a compelling basketball game, which get a lot of attention from people, and "force" people to tweet about them intensely. In physics, this "force" can be expressed by "acceleration" , which in our setting describes the change of "velocity" , i.e. arriving rate of tweets. Bursty topics can get significant acceleration when they are bursting, while the general topics usually get nearly zero acceleration. So the "acceleration" trick can be used to preserve the information of bursty topics but filter out the others. However, as the topics are hidden, we can not calculate their accelerations directly. A possible way is to estimate them by calculating the accelerations of words instead. Equation 1 shows how we calculate the "velocity" works like a soft moving window, which gives the recent terms high weight, but gives low weight to the ones far away, and the smoothing parameter is the window size. To capture the change of velocity, acceleration is defined as the difference of velocities with different window size . (Similar to the divergence of 5 day average and 10 day average in stock market, which is used to estimate the stock trend.) At that day, there was a compelling basketball game between San Antonio Spurs and Oklahoma City Thunder. At the beginning, this event got a big surge in Twitter, and at the end it got another even bigger wave of discussion on this Western Conference final. The daily volume in (a1) shows the popularity of each topic. We can see that, till at the end of the

day, "spurs" reaches the same scale as "obama" and "car". However, it will be too late if we wait till we observe the surge in volume to report this bursty topic. As shown in (b1), an earlier indicator is velocity, i.e. the arriving rate of a topic. Our idea of early detection is to monitor the acceleration of a topic which, compared against volume and velocity, gives an even earlier indicator of the popularity surge. The dash line in the plot shows the time when our detection system could be triggered. It is clear that at this time point, the daily volume and "car" nearly get zero acceleration in (c1), it is easy to distinguish "spurs" from them under the measurement of acceleration.

V. REALTIME D ETECTION TECHNIQUES

In this section, we present the technique details to achieve real-time efficiency for bursty topic detection in the huge- volume tweet stream setting.

A. Dimension Reduction

The first challenge is the high dimension problem as a result of the huge number of distinct words N in the tweet stream, which could easily reach the order of millions or even larger (see the experiments in Section VI-A). This results not only in an enormous data sketch (recall Y ″(t) in the sketch is an N×N matrix) but also an optimization problem of very high dimensions, i.e.

O (N·K).Since the problem is mainly because N is too large, one natural solution is to keep only a set of active words encountered recently, e.g. in the last 15 minutes. When a burst is triggered, consider only the words in this recent set. However, it turns out that the size of this reduced active word set for tweet stream is still too large to solve the optimization problem efficiently.

CONCLUSIONS

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 "sketch of topic", 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 efficiently. Compared with existing event detection system, from a different perspective – the "accelerations of topics", our

solution can detect bursty topics in real-time, and present them in finer-granularity.

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(Suppl1):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.