International Journal of Research
Available at https://edupediapublications.org/journals

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 12
April 2018

Topic Sketch: Real-time Bursty Topic Detection from Twitter

[1]**B. Ravi Krishna**, raviome@gmail.com

[2]**S. Varsha** - varshasamala125@gmail.com

[3]**M. Sai Rahul**- msr0997@gmail.com

[4]**P.Sujan**- sujanpeddi@gmail.com

[5]**GV.Akhil**- akhilkumarhyd@gmail.com

[1, 2, 3,4,5] Dept.of Computer Science and Engineering, Vignan Institute of Technology and Science, Deshmukhi, Hyderabad. 508284

***Abstract**:*

Twitter has become one of the largest micro blogging 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 period of 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 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 would like to propose TopicSketch, a sketch-based topic model together with a set of techniques to achieve real-time detection. We are planning to evaluate our solution on a tweet stream with huge volume of tweets. Our approach is thatTopicSketch on a single machine can potentially handle hundreds of millions tweets per day, which is on the same scale of the total number of daily tweets in Twitter, and present bursty events in finer-granularity....

Keywords

Topicsketch, Tweet Stream, Bursty Topic, Realtime...

I. Introduction

With 200 million active users and over 400 million tweets per day, 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”. Figure 1 shows an example of a bursty topic on November 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.

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 m 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 variation 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 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 Figure1 that Topic Sketch 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 Topic Sketch. 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 Topic Sketch on a tweet stream containing over 30 million tweets and demonstrated both the effectiveness and efficiency of our approach. It has been shown that Topic Sketch 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.

II. Literature Survey

Online inference for the infinite topic-cluster model: Storylines from streaming text. We present the time-dependent topic-cluster model, a hierarchical approach for combining Latent Dirichlet Allocation and clustering via the Recurrent Chinese Restaurant Process. It inherits the advantages of both of its constituents, namely interpretability and concise representation. We show how it can be applied to streaming collections of objects such as real world feeds in a news portal. We provide details of a parallel Sequential Monte Carlo algorithm to perform inference in the resulting graphical model which scales to hundreds of thousands of documents.

Clustering: Given the high frequency of news articles in considerable excess of one article per second even for quality English news sites it is vital to group similar articles together such that readers can sift through relevant information quickly.

Timelines: Aggregation of articles should not only occur in terms of current articles but it should also account for previous news. This matters considerably for stories that are just about to drop off the radar so that they may be categorized efficiently into the bigger context of related news.

Content analysis: We would like to group content at three levels of organization: high-level topics, individual stories, and entities. For any given story, we would like to be able to identify the most relevant topics, and also the individual entities that distinguish this event from others which are in the same overall topic. For example, while the topic of the story might be the death of a pop star, the identity Michael Jackson will help distinguish this story from similar stories.

Online processing: As we continually receive news documents, our understanding of the topics occurring in the event stream should improve. This is not necessarily the case for simple clustering models — increasing the amount of data, along time, will simply increase the number of clusters. Yet topic models are unsuitable for direct analysis since they do not reason well at an individual event level.

Online learning for latent dirichlet allocation

We develop an online variation Bayes (VB) algorithm for Latent Dirichlet Allocation (LDA). Online LDA is based on online stochastic optimization with a natural gradient step, which we show converges to a local optimum of the VB objective function.

It can handily analyze massive document collections, including those arriving in a stream. We study the performance of online LDA in several ways, including by fitting a 100-topic topic model to 3.3M articles from Wikipedia in a single pass.

We demonstrate that online LDA finds topic models as good as or better than those found with batch VB, and in a fraction of the time.

We develop an online variation Bayes algorithm for latent Dirichlet allocation (LDA), one of the simplest topic models and one on which many others are based.

Our algorithm is based on online stochastic optimization, which has been shown to produce good parameter estimates dramatically faster than batch algorithms on large datasets.

Online LDA handily analyzes massive collections of documents and, moreover, online LDA need not locally store or collect the documents— each can arrive in a stream and be discarded after one look. In the subsequent sections, we derive online LDA and show that it converges to a stationary point of the variational objective function.

We study the performance of online LDA in several ways, including by fitting a topic model to 3.3M articles from Wikipedia without looking at the same article twice.

We show that online LDA finds topic models as good as or better than those found with batch VB, and in a fraction of the time (see figure 1). Online variational Bayes is a practical new method for estimating the posterior of complex hierarchical Bayesian models.

Streaming first story detection with ´ application to twitter.

With the recent rise in popularity and size of social media, there is a growing need for systems that can extract useful information from this amount of data.

We address the problem of detecting new events from a stream of Twitter posts. To make event detection feasible on web-scale corpora, we present an algorithm based on locality-sensitive hashing which is able overcome the limitations of traditional approaches, while maintaining competitive results.

In particular, a comparison with a stateof-the-art system on the first story detection task shows that we achieve over an order of magnitude speedup in processing time, while retaining comparable performance.

Event detection experiments on a collection of 160 million Twitter posts show that celebrity deaths are the fastest spreading news on Twitter.

We find that simply applying pure LSH in a FSD task yields poor performance and a high variance in results, and so introduce a modification which virtually eliminates variance and significantly improves performance.

We show that our FSD system gives comparable results as a state-of-the-art system on the standard TDT5 dataset, while achieving an order of magnitude speedup.

Using our system for event detection on 160 million Twitter posts shows that

i) The number of users that write about an event is more indicative than the volume of tweets written about it,

ii) Spam tweets can be detected with reasonable precision, and

iii) News about deaths of famous people spreads the fastest on Twitter.

Continuous time dynamic topic models.

We develop the continuous time dynamic topic model (cDTM). The cDTM is a dynamic topic model that uses Brownian motion to model the latent topics through a sequential collection of documents, where a “topic” is a pattern of word use that we expect to evolve over the course of the collection.

We derive an efficient variational approximate inference algorithm that takes advantage of the sparsity of observations in text, a property that lets us easily handle many time points. In contrast to the cDTM, the original discrete-time dynamic topic model (dDTM) requires that time be discretized.

Moreover, the complexity of variational inference for the dDTM grows quickly as time granularity increases, a drawback which limits fine-grained discretization. We demonstrate the cDTM on two news corpora, reporting both predictive perplexity and the novel task of time stamp prediction.

We consider time to be continuous. The continuous time dynamic topic model (cDTM) proposed here replaces the discrete state space model of the dDTM with its continuous generalization, Brownian motion.

The cDTM generalizes the dDTM in that the only discretization it models is the resolution at which the time stamps of the documents are measured.

The cDTM model will, generally, introduce many more latent variables than the dDTM. However, this seemingly more complicated model is simpler and more efficient to fit.

Topics over time: a non-markov continuous-time model of topical trends

This presents an LDA-style topic model that captures not only the low-dimensional structure of data, but also how the structure changes over time. Unlike other recent work that relies on Markov assumptions or discretization of time, here each topic is associated with a continuous distribution over timestamps, and for each generated document, the mixture distribution over topics is influenced by both word co-occurrences and the document's timestamp.

Thus, the meaning of a particular topic can be relied upon as constant, but the topics' occurrence and correlations change significantly over time. We present results on nine months of personal email, 17 years of NIPS research papers and over 200 years of presidential state-of-the-union addresses, showing improved topics, better timestamp prediction, and interpretable trends.

This paper presents Topics over Time (TOT), a topic model that explicitly models time jointly with word co-occurrence patterns. Significantly, and unlike some recent work with similar goals, our model does not discretize time, and does not make Markov assumptions over state transitions in time.

Rather, TOT parameterizes a continuous distribution over time associated with each topic, and topics are responsible for generating both observed timestamps as well as words. Parameter estimation is thus driven to discover topics that simultaneously capture word co-occurrences and locality of those patterns in time.

The model's generative storyline can be understood in two different ways. We fit the model parameters according to a generative model in which a per-document multinomial distribution over topics is sampled from a Dirichlet, then for each word occurrence we sample a topic; next a per-topic multinomial generates the word, and a per-topic Beta distribution generates the document's time stamp. Here the time stamp (which in practice is always observed and constant across the document) is associated with each word in the document.

III. System Analysis

A. *Existing System*

- SigniTrend can detect bursty keywords in real-time, but before it aggregates keywords into larger topics, it needs to wait until the end-of-day (or a fixed time period).
- Yang et al. use refined hierarchical and online document clustering algorithms to detect events from a news stream.

Disadvantages Of Existing System

- High computational complexity.
- It does not scale to the overwhelming data volume like that of Twitter, as a nearest neighbor search is costly on large data set.
- Usually a collection of bursty terms are detected from the document stream based on some criteria, and possibly later these bursty terms are grouped into several clusters which represent the bursty topics.

Proposed System

- We propose a new detection framework called Topic Sketch. It can be observed from that Topic Sketch 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.
- First, we proposed a two-stage integrated solution Topic Sketch.
- In the first stage, we proposed a small data sketch which efficiently maintains at a low computational cost the acceleration of two quantities: the occurrence of each word pair and the occurrence of each word triple. 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.
- Second, we proposed dimension reduction techniques based on hashing to achieve scalability and, at the same time, maintain topic quality with robustness.
- Finally, we evaluated Topic Sketch on a tweet stream containing over 30 million tweets and demonstrated both the effectiveness and efficiency of our approach. It has been shown that Topic Sketch on a single machine is able to potentially handle over 150 million tweets per day which is on the same scale of the total number of tweets generated daily in Twitter.

Advantages Of Proposed System

- More sophisticated sketch structure, which captures not only the information of word pairs, but also the word triples;
- More effective inference algorithm, i.e. tensor decomposition, which is an important contribution on top and more comprehensive evaluations.

IV. System Design and Architecture

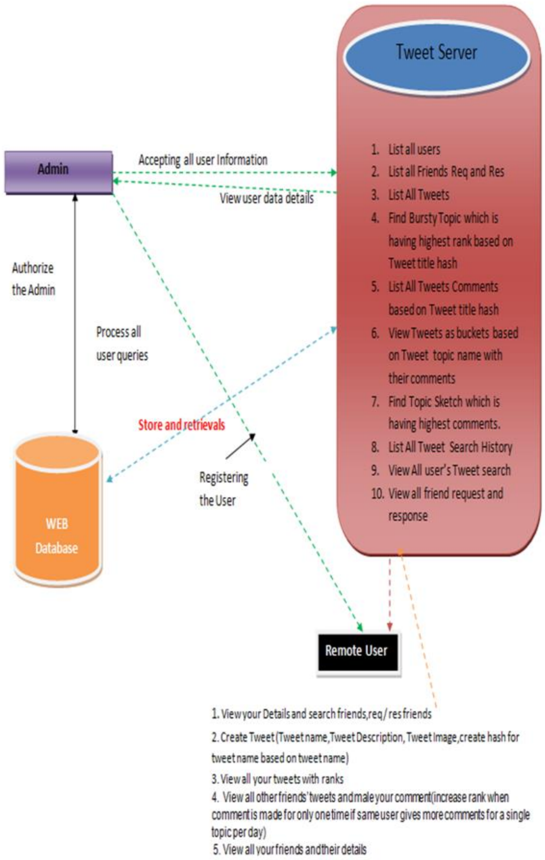

Tweet Server
1. List all users
2. List all Friends Req and Res
3. List All Tweets
4. Find Bursty Topic which is having highest rank based on Tweet title hash
5. List All Tweets Comments based on Tweet title hash
6. View Tweets as buckets based on Tweet topic name with their comments
7. Find Topic Sketch which is having highest comments.
8. List All Tweet Search History
9. View All user's Tweet search
10. View all friend request and response
Admin
Accepting all user Information
View user data details
Authorize the Admin
Process all user queries
Store and retrievals
Registering the User
WEB Database
Remote User
1. View your Details and search friends,req/res friends
2. Create Tweet(Tweet name,Tweet Description, Tweet Image,create hash for tweet name based on tweet name)
3. View all your tweets with ranks
4. View all other friends' tweets and male your comment(increase rank when comment is made for only one time if same user gives more comments for a single topic per day)
5. View all your friends and their details

Fig 1: System Architecture

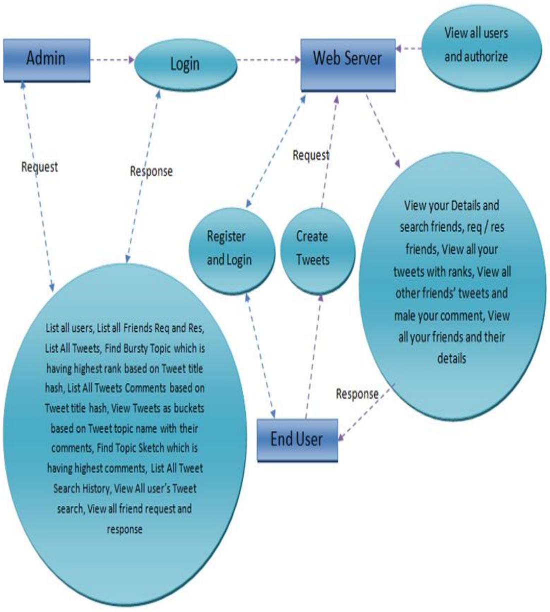

Admin
Login
Web Server
View all users and authorize
Request
Response
Request
Register and Login
Create Tweets
View your Details and search friends, req / res friends, View all your tweets with ranks, View all other friends' tweets and male your comment, View all your friends and their details
List all users, List all Friends Req and Res, List All Tweets, Find Bursty Topic which is having highest rank based on Tweet title hash, List All Tweets Comments based on Tweet title hash, View Tweets as buckets based on Tweet topic name with their comments, Find Topic Sketch which is having highest comments, List All Tweet Search History, View All user's Tweet search, View all friend request and response
Response
End User

Fig 2: Data Flow Diagram

Software Requirements:

Listed below are the minimum software requirements that must be fulfilled in order for the software to run on your system.

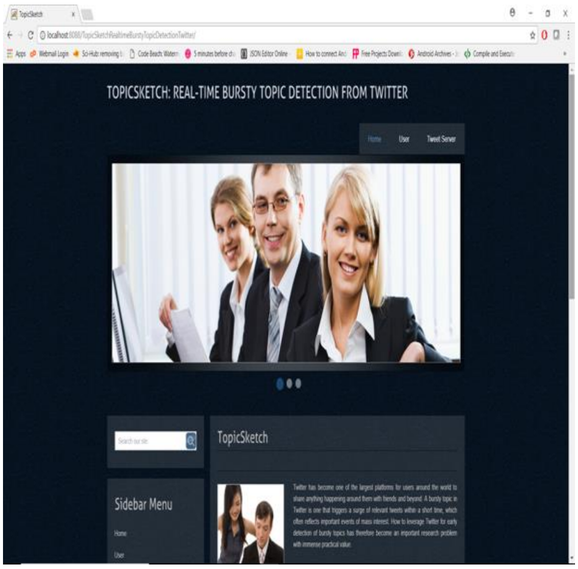
Software	Minimum Requirement
Operating System	Windows
Technology	JAVA or J2EE
Web Technology	HTML, CSS and JavaScript
IDE	Eclipse
Web Server	Tomcat
Database	MySQL
JAVA Version	J2SDK 1.5

Hardware requirements:

Listed below are the minimum hardware requirements that must be fulfilled in order for the software to run on your system.

Hardware	Minimum Requirement
Processor	Pentium
Speed	2.1 GHz
RAM	2 GB
Hard Disk	20 GB

V. Results



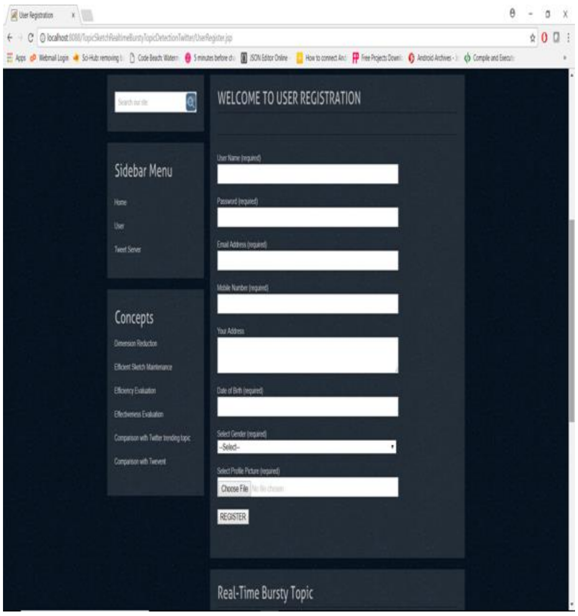
TOPICSKETCH: REAL-TIME BURSTY TOPIC DETECTION FROM TWITTER

Home User Tweet Server

Sidebar Menu
Home
User

TopicSketch

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.



localhost:8080/TopicSketchRealtimeBurstyTopicDetectionTwitter/UserRegister.jsp

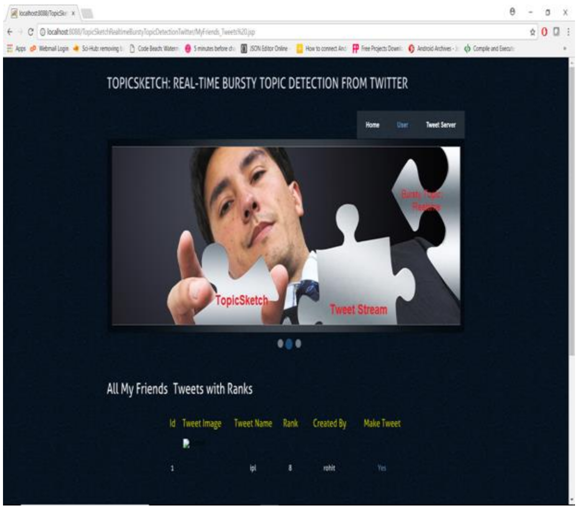
Sidebar Menu
Home
User
Tweet Server

Concepts
Dimension Reduction
Efficient Sketch Maintenance
Efficiency Evaluation
Effectiveness Evaluation
Comparison with Twitter trending topic
Comparison with Twevent

WELCOME TO USER REGISTRATION

User Name (required)
Password (required)
Email Address (required)
Mobile Number (required)
Your Address
Date of Birth (required)
Select Gender (required)
--Select--
Select Profile Picture (required)
Choose File No file chosen
REGISTER

Real-Time Bursty Topic

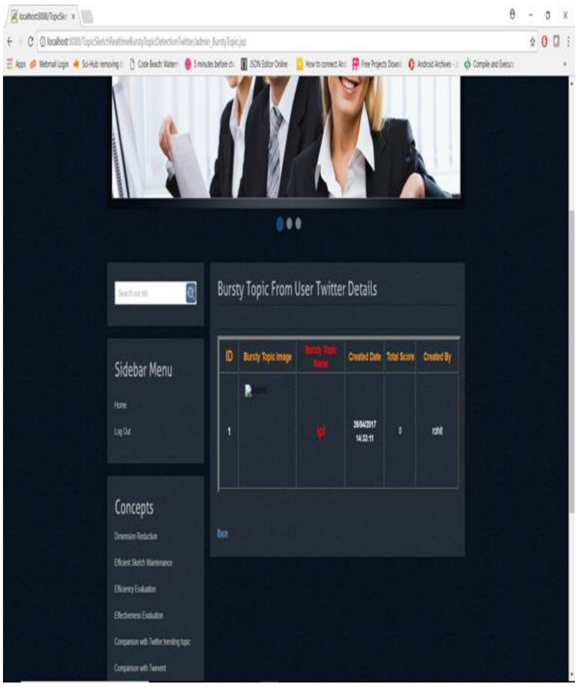


TOPICSKETCH: REAL-TIME BURSTY TOPIC DETECTION FROM TWITTER

Home User Tweet Server

All My Friends Tweets with Ranks

Id	Tweet Image	Tweet Name	Rank	Created By	Make Tweet
1		ipl	8	rohit	Yes



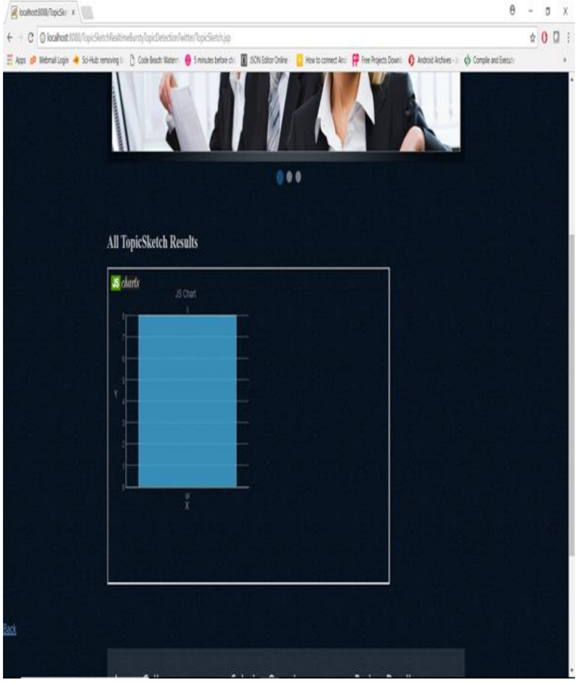
Sidebar Menu
Home
Log Out

Concepts
Dimension Reduction
Efficient Sketch Maintenance
Efficiency Evaluation
Effectiveness Evaluation
Comparison with Twitter trending topic
Comparison with Twevent

Bursty Topic From User Twitter Details

ID	Bursty Topic Image	Bursty Topic Name	Created Date	Total Score	Created By
1		ipl	28/04/2017 14:32:11	8	rohit

Back



localhost:8080/TopicSketchRealtimeBurstyTopicDetectionTwitter/TopicSketch.jsp

All TopicSketch Results

Back

VI. Conclusion

In this paper, we proposed Topic Sketch 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. Compared with existing event detection system, our experiments have demonstrated the superiority of Topic Sketch in detecting bursty topics in real-time.

VII. 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. Barbar´a, 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.