International Journal of Computer Science Engineering & Technology
IJCSET
APPROVED BY
NATIONAL SCIENCE LIBRARY (NSL)
NATIONAL INSTITUTE OF SCIENCE COMMUNICATION AND INFORMATION RESOURCES (NISCAIR)
COUNCIL OF SCIENTIFIC AND INDUSTRIAL RESEARCH (CSIR)- NEW DELHI INDIA .
ISSN :2455-9091

BURSTY TOPIC DETECTION FROM TWITTER USING HOSVD

Asst.Prof. [1] P. Kalpana,
UG Students: [2] M. Chandru, [3] P. Dhanasekaran, [4] M.N. Naveen Kumar,
[1, 2, 3, 4] Department of Computer Science and Engineering,
[1, 2, 3, 4] Sri Krishna College of Technology,
[1, 2, 3, 4] Coimbatore, Tamil Nadu, India.

ABSTRACT: Twitter becomes one of the largest micro blogging platforms for users around the world. These studies have aimed a extracting the period and the location in which a specified topic frequently occurs. 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. A bursty topic in Twitter is one that triggers a surge of equalling 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 complication with immense practical value. Despite the wealth of analysis work on topic modelling and analysis in Twitter, it remains a challenge to detect bursty topics in real-time. In this paper, we use framework higher-order singular value decomposition (HOSVD) we focus on a system that detects hot topic in a local area and during a particular period. There can be a changes in the words used even though the posts are essentially about the same hot topic. Topic modelling and analysis in Twitter, it remains a challenge to detect bursty topics in real-time. Our experiments on a large Twitter dataset and synthetic datasets show that the proposed models can effectively mine the topic-specific behavioural factors of users and tweet topics. We further demonstrate that the proposed model consistently outperforms the other state-of-the-art content based models in retweet prediction over time.

1. INTRODUCTION

Social media provides a great source of information. The online conversations have undergone tremendous growth over the past few years. Some of the conversations are personal status updates, usually more significant to a user's social circle. While a large portion of the conversations in the social media space are instead responses triggered by events. Such events add natural disasters (e.g. hurricanes, earthquakes), political events (e.g. presidential elections), protests and marches, etc. Take Occupy Wall Street as an example, the OWS movement is a widely competed event known to use social media to advertise and spread nationwide. The movement is long lasting and widespread without central leadership, which creates challenges in understanding and acknowledging to the movement. Given the critical role of social media in the OWS movement, it is a great source to understand and search such events.

A topic in this work is represented as a distribution overwords. Particularly, in defining a bursty topic, we evaluatethe following two criteria: (I) There has to be a sudden surgein the topic's popularity which is

measured by the total number of relevant tweet. Those all-time popular topicstherefore would not count; (II) The topic must be reasonably famous. This would filter away the large number of trivialtopics which, despite the spikes in their popularity, areconsidered as noises because of the negligible number ofrelevant tweets.For criterion (I), we measure how bursty a topic is bythe acceleration of its popularity. Mathematically speaking, acceleration captures the change in the rate of the popularityof a topic. The more sudden the change is, the larger theacceleration is. We explain how to estimate theacceleration of a topic without even knowing which tweetsare associated to it. For criterion (II), once we found bursty
topic candidates, we simply count the relevant tweets ofthem, and filter out the trivial ones.Our task in this paper is, given a tweet stream, to detectbursty topics from it as early as possible.

2. LITERATURE REVIEW

An event is commonly considered as an existence at a specific time and place [1, 2, 3, 4, 5, 6, 7]. However, in the social media space, certain social campaign/movements do not wanted to happen in a physical location. We revise the definition of an event in the context of social media as an occurrence beginadjustment in the volume of text information that consults the associated topic at a specific time. This occurrence is characterized by topic and time, and often combined with entities such as people and location." Recent analysis has worked that one of the common uses of social media is reporting and discussing events users are experiencing: Sakaki et al. [11] demonstrated that mining of relevant tweets can be used to detect earthquake events and predict the earthquake center in real time. Becker et al. [12] proposed to identify real-world events through exploring a variety of methods for learning multi-feature similarity metrics for social media documents. Their evaluation results showed that events could be effectively noted from large-scale images provided by Flickr. Although numerous research papers have focused on presenting methods and systems for extracting event-related information from social media and newswire, few has reviewed the exposed systems from a task-specific perspective.

In previous process we useda two-stage integrated solution TopicSketch. In the first stage, we proposed a small data sketch which efficiently maintains at a low computational cost the acceleration of two quantities: the appearance 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 arranged such that the bursty topic inference would be triggered and achieved based on them. The fact thatwe can update these statistics efficiently and invoke the more computationally high values 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 used 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 meodsbased on hashing to achieve scalability and, at the same time, maintain topic quality with robustness.

3. PROBLEM DEFINITION

Finally, we evaluated Topic Sketch on a tweet stream containing over 30 million tweets and worked both the effectiveness and efficiency of our approach. It has been shown that TopicSketch on a single machine is able to potentially handle over 150 million tweets per day which is on the same level of the total number of tweetsgenerated daily in Twitter. We also presented case studieson interesting bursty topic examples which illustrate some desirable features of our approach, e.g., finer-granularity event description. Drawbacks identified below as Twitter for early detection of bursty topics has therefore become an

important research complication with immense practical value.

High computational complexity inherent in the topic models as well as the ways in which the topics are usually learnt.

Model bursty topics without the chance to shows the entire set of relevant tweets as in traditional topic modelling

In this paper, we focus on the task of identifying major events and sub-events when searching social media data. The timeline of events serves as a succinct summary of the massive social media space. Through further analysis of the responses to each individual event over time, one can gain wisdom regarding the event itself, people's opinions towards the event, as well as inferring causal relationships between events (sub-events) and responses. Our solution, called TopicSketch, is based on twomain techniques — a sketch-based topic model and a hashing-based dimension reduction technique. Our sketch-based topic model provides an integrated two-step solution. In the first step, it maintains as a sketch of the data the acceleration of two quantities: every pair of words, and every triple of words, which are early indicators of popularity surge and can be updated efficientlyat a low cost, making early detection possible. In the second step, based on the data sketch, it learns the bursty topics by tensor decomposition. To perform the detection efficiently in large-scale real-time setting, we propose a dimension reduction technique based on hashing whichProvides a scalable solution to the original problem without compromising much the quality of the topics.

4. BURSTY TOPIC DETECTION

Bursty Topic Detection Challenges of real time bursty topic disclosure arise from the following aspects: (1) How to efficiently maintain proper statistics to trigger detection; (2) How to model bursty topics without examining the entire set of relevant tweets; and (3) How to scale to the big volume of tweet stream. In our ICDM′ 13 work [8], we proposed a solution called TopicSketch which maintains a data sketch of the accelerations of three quantities at any time stamp t: (1) The whole tweet stream S ″(t), (2) Each word X ″(t) and (3) Each pair of words Y ″(t). The first two provide early signals of popularity surge while Y ″(t) are used to infer bursty topics from the keyword correlation embedded. We follow up with an idea similar to MACD(Moving Average Convergence) to estimate those quantities. Besides early detection, this data sketch also contributes to latent bursty topics inference. We model the whole tweet stream as a mixture of multiple latent topic streams and we are interested in identifying the top-K latent topics pk whose rate ak (t) is greater than a predefined threshold at any time stamp t. Once a predefined detection criteria is satisfied, we trigger the estimation of the parameters pk and ak (t) by solving a constrained optimization problem. The algorithm is detailed in.

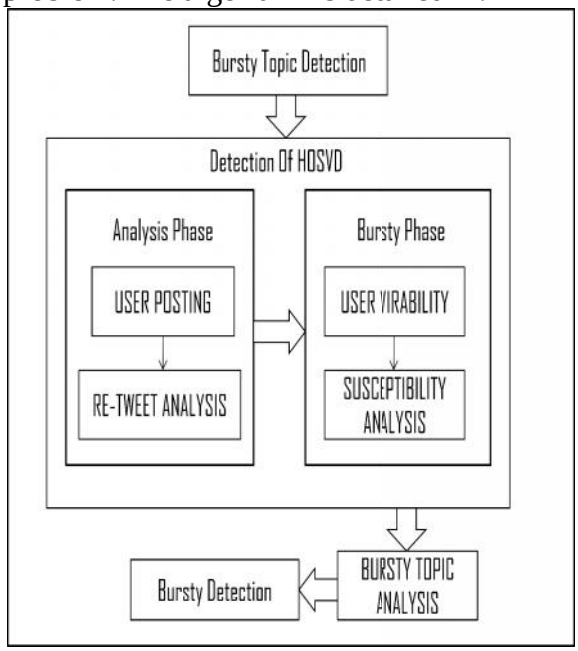

Figure1 - Architecture Diagram

We propose a methodology based on time-window analysis. We compute keyword frequencies, normalize by relevance and compare them in adjacent time windows. This comparison consists of analyzing variations in term arrival rates and their respective variation percentages per window. A similar notion

(Discrete Second Derivative) has been used in the context of espial of bursts in academic citations [4]. We define Relevance Rates (RR) as the probability of occurrence of a non-stopword term in a window. We use RR to generalize burst detection making them independent of the appearance rate. Even though the public Twitter API only provides a stream which is said to be less than 10% of the actual tweets posted on Twitter, we believe our technique can be easily adapted for the complete data stream using a MapReduce schema [see Section 4.2.3]. Arrival rates vary periodically (in a non-bursty way) during the day; depending on the hour, time zone, user region, global events and language (shown in Figure 4). Bursty keywords are ranked according to their relevance variation rate. Our method avoids the use of statistical distribution analysis methods for keyword frequencies; The main reason is that this approach, commonly used in state-of-the-art approaches, increases the complexity of the process. We show that a simple relevance variation concept is sufficient for our purposes if we use good stopword filter and noise minimization analysis. To study the efficiency of our algorithm, which we name Window Variation Keyword Burst Detection, we implement a proofof-concept system. The processes involved in this system are five. These modules are independent of each other and they have been structured for processing in threads. These modules are: Stream Listener, Tweets Filterer, Tweets Packing, Window Processing and Keyword Ranker. This architecture allows us to process all the input data with linear complexity, making it scalable for on-line processing

5. PROPOSED SYSTEM

A bursty topic in Twitter is one that triggers a surge of significant 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 analysis problem with huge practical value. Despite the wealth of research work on topic modelling and search in Twitter, it remains a challenge to detect bursty topics in real-time. In this paper, we use framework higher-order singular value decomposition (HOSVD) we focus on a system that detects hot topic in a local area and during a particular period. There can be a variation in the words used even though the posts are essentially about the same hot topic. Topic modelling and analysis in Twitter, it remains a challenge to detect bursty topics in real-time. Our experiments on a large Twitter dataset and synthetic datasets show that the proposed models can effectively concerns the topic-specific behavioural factors of users and tweet topics. We further demonstrate that the proposed model consistently outperforms the other state-of-the-art content based models in retweetforecost over time.

6. ALGORITHM HIGHER-ORDER SINGULAR VALUE DECOMPOSITION (HOSVD)

In many areas of science and methods, data structures have more than two dimensions, and are naturally represented by multidimensional arrays or tensors. Two-dimensional matrix methods, such as the singular value decomposition (SVD), are wide range and well studied mathematically. However, they do not take into account the multidimensionality of data. In some scientific areas, notably chemometrics and psychometrics, tensor methods have been developed and used with great success.The SVD may be generalized to higherorder tensors or multiway arrays in several ways. The two main techniques are the so-called Tucker/HOSVD decomposition and the CP expansion (from canonical decomposition (CANDECOMP) and parallel factors (PARAFAC) [4]). The CP expansion is a special case of the Tucker/HOSVD decompositions. For simplicity, we begun these decompositions for tensors of order q 5 3. This shows all fundamental differences to

the case of a conventional matrix (q 5 2), and generalizations to q.3 are rather direct.

Advantages Discussed In This Proposed Methodology

- Efficiently maintains at a low computational cost the acceleration of two quantities
- These accelerations provide as early as possible the indicators of a potential surge of tweet popularity.
- Update these statistics efficiently and invoke the more computationally expensive topic inference part only
- It possible to achieve real-time detection in a data stream of Twitter scale.
- Our solution can detect bursty topics in real-time, and present them in finer-granularity.
- Variety of difficulties in mining sequential patterns from massive data represented by a huge set of symbolic features.
- Reduces the representation of the sequential data by uncovering significant, hidden temporal structures.

7. EXPERIMENTAL RESULT

The experimental results suggest that topic identification by HOSVD a set of keywords works fairly well, using either of the investigated similarity measures. In the present experiment a recently proposed distribution of terms associated with a keyword clearly gives best results, but computation of the distribution is relatively expensive. The reason for this is the fact that co-occurrence of terms is (implicitly) taken into account.

Techniques	Accuracy	Throughput	Delay
K-means	78%	80%	70%
SVD	82%	86%	62%
HOSVD	90%	92.3%	51%

Table1: Performance measure table

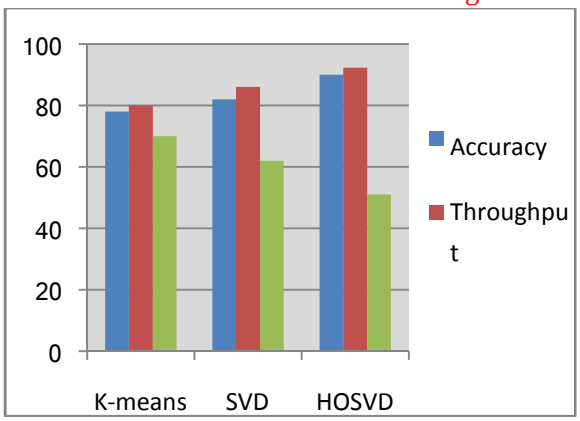

Figure 2: Performance Graph

This study has shown that fairly simple techniques can achieve very high quality results, but that substantial work is needed to reduce the errors to manageable numbers. Fortunately, that the problem focuses on Broadcast News and not on arbitrary forms of information means that there is hope that more carefully crafted approaches can improve the tracking results substantially.

CONCLUSION

We proposed TopicSketch a framework forreal-time detection of bursty topics from Twitter. Due tothe huge volume of tweet stream, existing topic modelscan hardly scale to data of such sizes for real-time topicmodeling tasks. We developed a “sketch of topic”, whichprovides a “snapshot” of the current tweet stream and canbe updated efficiently. This makes our approach easy to use and promising for on-line processing in comparison to other state-of-the-art methods. Experimental results indicate that our algorithm can scale to high tweet arrival rates while still producing high-quality results. Overall, our system produces an extraordinary keyword overlap against LDA, using very limited resources and memory

REFERENCES

[1] M. JunaidMajeed; M. Man Malik; M. Taimoor Khan; Shehzad Khalid “Real-time government policy monitoring framework using expert guidedtopic modeling”2016 Sixth

International Conference on Innovative Computing Technology (INTECH)
[2] N. Du, M. Farajtabar, A. Ahmed, A. J. Smola, and L. Song.Dirichlet-hawkes processes with applications to clustering continuous-time document streams. In Proceedings of the 21th ACMSIGKDD International Conference on Knowledge Discovery and DataMining, Sydney, NSW, Australia, August 10-13, 2015, pages 219–228,2015.
[3] Li Zhao, Yan Li, Xinran Liu, Hong Zhang The National Computer Network Emergency Response Technical Team Coordination Center of China (CNCERT) Beijing, China "A Graph-based Bursty Topic Detection Approach in User-Genterated Texts" © 2014 IEEE DOI 10.1109/WISA.2014.
[4] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky.Tensor decompositions for learning latent variable models. Journalof Machine Learning Research, 15(1):2773–2832, 2014.
[5] W. Feng, C. Zhang, W. Zhang, J. Han, J. Wang, C. Aggarwal, andJ. Huang. STREAMCUBE: hierarchical spatio-temporal hashtagclustering for event exploration over the twitter stream. In 31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul,South Korea, April 13-17, 2015, pages 1561–1572, 2015.
[6] Q. Diao, J. Jiang, F. Zhu, and E. Lim. Finding bursty topicsfrom microblogs. In The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14,2012, Jeju Island, Korea - Volume 1: Long Papers, pages 536–544, 2012.
[7] M. Cataldi, L. D. Caro, and C. Schifanella. Personalized emergingtopic detection based on a term aging model. ACM TIST, 5(1):7,2013.
[8] F. Alvanaki, S. Michel, K.Ramamritham, and G. Weikum.See what's enblogue: real-time emergent topic identificationin social media. In 15th International Conference on ExtendingDatabase Technology, EDBT '12, Berlin, Germany, March 27-30, 2012, Proceedings, pages 336–347, 2012.
[9] A. Ahmed, Q. Ho, C. H. Teo, J. Eisenstein, A. J. Smola, andE. P. Xing. Online inference for the infinite topic-cluster model:Storylines from streaming text. In Proceedings of the FourteenthInternational Conference on Artificial Intelligence and Statistics,AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011, pages101–109, 2011.
[10] J. Chang, J. L. Boyd-Graber, S. Gerrish, C. Wang, and D. M.Blei. Reading tea leaves: How humans interpret topic models. In Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009.Proceedings of a meeting held 7-10 December 2009, Vancouver, BritishColumbia, Canada., pages 288–296, 2009.
[11] D. M. Blei and J. D. Lafferty. Dynamic topic models. InProceedings of the 23rd international conference on Machine learning,pages 113–120, 2006.
[12] G. Cormode and S. Muthukrishnan. An improved data streamsummary: the count-min sketch and its applications. Journal of Algorithms, 55(1):58–75, 2005.
[13] T. Brants and F. Chen.A system for new event detection. InSIGIR 2003: Proceedings of the 26th Annual International ACM SIGIRConference on Research and Development in Information Retrieval, July28 - August 1, 2003, Toronto, Canada, pages 330–337, 2003.
[14] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. TheJournal of machine Learning research, 3:993–1022, 2003.
[15] J. Allan, R. Papka, and V. Lavrenko. On-line new event detectionand tracking. In SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development inInformation Retrieval, August 24-28 1998, Melbourne, Australia, pages37–45, 1998.