Identification of Events in Social News Streams

Tabiya Manzoor Beigh
taha.beigh@gmail.com

Shuchita Upadhyaya
shuchita_bhasin@yahoo.com

Girdhar Gopal
girdhar.gopal@kuk.ac.in

Department of Computer Science & Applications, Kurukshetra University, Kurukshetra-136119

Abstract

Current age can be well described as the age of digitization. The digital universe uses enormous amount of structured as well as unstructured data. The data can be in any form, of any shape and size. The data being generated evolves from almost everything and every source which could be imagined in present era. Social media produces enormous amounts of data. Social media provides information about the latest happenings of the world. Simultaneously it provides the opinion of every individual. Analyzing the news streams which are temporally sequenced has been an interesting issue. In this paper, timely detection of bursty events and their evolution is carried out A news stream is represented as feature streams of thousands of features (i.e., keyword). Each news story consists of a set of keywords. A bursty event therefore is composed of a group of bursty features, which show bursty rises in frequency as the related event emerges. In this paper, we give a formal insight to the above problem and present a solution. We extensively evaluate the proposed methods on the Reuters Corpus Volume 1. Experimental results show that our methods can detect bursty events in a timely way and effectively discover their evolution.

Keywords:- Big data, Bursty Feature, Bursty Event, Event Detection, News Streams, Social media

1. Introduction

IDC- a premier global market intelligence firm termed 21st Century as “Digital universe”. Data has many different uses – real-time fraud detection, web display advertising and competitive analysis, call center optimization, social media and sentiment analysis, intelligent traffic management and smart power grids, to name just a few. All of these analytical solutions involve significant (and growing) volumes of both multi-structured and structured data. Various data sources include social media, sensors, health sector, digital images, energy and utilities. As per the Internet, it was predicted that digital universe is set to explode to an unimaginable 8 Zeta bytes by the year 2015. This huge amount of data was actually produced. This would roughly be a stack of DVDs from Earth all the way to Mars. The term “Big data” was coined to address this enormous volume of data storage and processing. Any data having property of 3V ’s is said to be big data. Any data having extremely large size mostly in Tera bytes or Peta bytes represents volume. Data arriving or evolving at an alarming speed is the velocity of the data. The data can be in any form, shape or size. It represents the variety of data.

Temporally sequences news streams has been widely increasing. The available stream of news reports threatens to be over whelming. The available news reports boosted researchers to analyze and utilize the news reports. Topic Detection and Tracking (TDT) Community has been studying for a decade to give practical solutions for effectively monitoring news streams for important events[1].

The rest of the paper is organized as follows:- Section 2 describes in detail the factors that gave rise to the issue of event detection in news streams. Section 3 describes related work on event detection in social media. Section 4 describes in detail all the facets of the proposed methodology and the Inference. . Section 5 summarizes our work and concludes the paper.

2. Problem Formulation

According to Nielsen, Internet users continue to spend more time with social media site than any other type of site. The total time spent on social media in the U.S. across PC and mobile devices increased by 99% to 121 billion minutes in July 2012 compare to 66 billion minutes in July 2011. Social media includes activities that are directly influenced by real world happenings. To identify events in social media is a challenging task. It should be also taken into consideration whether to identify events in social media from online stream or archived stream. The content on social media includes photographs, videos, and text documents. The content may be related to an event either well known or smaller. An event is something that occurs in a certain place during a particular interval of time[2]. An event can be as big as possible such as the marriage ceremony of a celebrity. their related social media content helps in powerful local event browsing and search. Events cover real life as well as web happenings, and may comprise only one o several topics. The web representation will appear after the real life event, the two event times are expected to be very close, as users tend to immediately report events in the social media, especially in Twitter.[3] .

With the increasing number of real-world events that are originated and discussed over social networks, events are stored in a news steam. Detecting news events and summarizing their evolutions along the timeline will provide a conceptual structure for news stories in the news stream and greatly facilitate users navigation in news spaces. In the TDT community, there are mainly two lines of research related to news event detection, i.e., retrospective news event detection (RED)[4] and new event detection (NED).[5] RED detects previously unidentified events in a news corpus. NED is more related to our work, which detects news stories about previously unseen events in a stream of news stories.

3. Related work

Numerous approaches have been proposed to detect the events in social streams. Various algorithms and techniques based on clustering are used to detect events in social media. Dynamic clustering techniques are mostly used in which event identification is modeled as an online incremental clustering task[6]. For each document, its similarity to existing events (clusters of documents) is computed and the document is assigned to either an existing event, or to a new event based on predefined criteria. Indexing -tree has been to detect news event in the given stream of stories.[7] Indexing-tree speeds up the detection process as it has been used dynamically used. The tree uses term re-weighting based on previous story clusters, or statistics on training data, to learn a model for each class of stories. Single pass clustering and online adaptive filtering has been used to handle evolving events within a stream of broadcast news stories.[8] Hierarchical and non-hierarchical clustering has been used to automatically detect novel events from a stream of news stories.[9] Document streams are transferred into feature streams using spatial approach.[10] A method has been proposed that clusters bursty features to form bursty events and associate each event with a power value which reflects its burst level. Event identification has also been attempted through statistical methods. A scalable system has been used that employs Latent Dirichlet Allocation in order to track events and sub-events, while excluding non-relevant (to the event of interest) text portions[11]. Classification algorithms are also extensively used in the event detection field. A supervised learning algorithm is used to classify the on-line document stream into pre-defined broad topic categories and then perform topic-conditioned detection for documents in each type.[12] Event detection has been done through graph analysis. An event can be defined as a set of relations between social actors on a specific topic over a certain time period and represent the social text streams as multi-graphs, where each node represents a social actor and each edge represents a piece of text communication that connects two actors. [13]Events are detected by combining text-based clustering, temporal segmentation and graph cuts of social networks. It has been hypothesized that documents describing the same event contain similar sets of keywords and the graph of keywords for a document collection contains clusters of individual events. In this context, they built a network of 1. keywords based on their co-occurrence in documents 2. and proposed a graph-based event detection algorithm, that uses community detection methods to discover and describe events.

4. Proposed work

4.1 Feature Selection in News stream

News stream is defined as collection of stories or documents. Each document comprises of a set of features in a vocabulary F = {f_1, f_2,…..$f_{t,.....}$}. News stream consists of thousands of time series of streams of features in F. Trial of a feature f_i can be written as a discrete time series f[1, 2, …, t, …], where each element $f_i[t]$ denotes the value of feature f_i at time point t .The unit of time point may be one hour, half a day or one day, etc Bursty event is a minimal set of bursty features that occur together in a certain time with strong support of documents in the text stream . These features of the same bursty event share a similar bursty pattern and intersect highly in documents when the event happens.

A feature is identified as bursty within a certain window when the aggregate value of its feature trail within the window is much larger than the aggregate values in most other windows of the same size. Therefore the key of online bursty events detection is to automatically indentify minimal sets of bursty features in the current time window. After identifying bursty events E={e1, e2, …, ei, …} in the current window, we also want to find the closely related events (i.e., event's evolution) in history.

The evolutionary trail of a bursty event e_i can also be written as discrete time series e_i [1, 2, …, t, …]. Each entry of the trail represents the relative strength of the event in the corresponding time point. The avail;able stream of news needs to be preprocessed.

4.2 Data Pre-processing

The data set is available in the XML format. The data is cleaned because of the availability of noise. The useful Xml tags are extracted so that events could be identified appropriately. XML tags which do not contribute to the detection of events are simply rejected.

4.3 Model Tuning

-After cleaning the data, data is now available in the readable format. Appropriate model should be applied to readable format data to detect the event in news streams.

The model will produce set of bursty features and eventually bursty events

4.4 Inference

Experimental results show that our methods can detect bursty events in a timely way and effectively discover their evolution.

Data used to test our approach is taken from Reuters Corpus Volume 1[14]. RCV 1 is an archive of 806791 manually categorized newswire stories made available by Reuters Ltd. for research purposes. These news stories are from '96/8/20' to '97/8/19' (totally 365 days) using a daily resolution. The Simple Analyzer of open source full text indexing and searching toolkit. Only the text in <title> and <body> fields were processed. We implemented all experiments in R. R is a programming language and a software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. Packages used in this approach include XML, Rminer and Kknn. Models used in this approach include ksvm and k-nearest neighbors

. 5 . Conclusion

Online monitoring of temporally-sequenced news streams for interesting patterns and trends has gained popularity in the last decade. In this paper, we studied a particular news stream monitoring task: timely detection of bursty events which have happened recently and discovery of their evolutionary patterns along the timeline. Here, a news stream is represented as feature streams of tens of thousands of features (i.e., keyword. Each news story consists of a set of keywords.). A bursty event therefore is composed of a group of bursty features, which show bursty rises in frequency as the related event emerges. In this paper, we give a formal definition to the above problem and presented a solution. We extensively evaluate the proposed methods on the Reuters Corpus Volume 1. Experimental results show that our methods can detect bursty events in a timely way and effectively discover their evolution.

References

[1] (2008) Topic Detection And Tracking Project. [Online]. http://www.itl.nist.gov/iad/mig//tests/tdt/

[2] R.Jain, "Eventweb: developing a human-centered computing system, Computer," *IEEE*, pp. 42-50, 2008.

[3] M.Okazaki, Y.Matsuo T.Sakaki, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th International Conference on World Wide Web WWW'10*, New York NY USA, 2010, pp. 851-860.

[4] Y.M., Pierce, T., Carbonell, J.G. Yang, "A Study on Retrospective and On-line Event Detection.," in *Conf. on Research and Development in Information Retrieval*, 1998, pp. 28-36.

[5] J., Papka, R., Lavrenko, V. Allan, "Online New Event Detection and Tracking.," in *Conf. on Research and Development in Information Retrieval*, 1998, pp. 37-45.

[6] J. G. Carbonell, R.D. Brown, T.Pierce, B.T. Archibald, X.Lui, Y.Yang, "Learning approches for detecting and tracking news events," *IEEE*, vol. 14, pp. 32-43, 1999.

[7] J.Zi, L.G.Wu K.Zhang, "New event detection based on Indexing Tree and named entity," in *Proceedings of 30th Annual International ACM SIGIR Conference on Research and developm ent in Information Retrieval*, New York NY, USA, 2007, pp. 215-222.

[8] R. Papka, V.Lavrenko J.Allan, "On-lone new event detection and tracking," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and development in information Retrieval*, New York NY USA, 1998, pp. 37-45.

[9] T. Pierce, J. Carbonell Y. Yang, "A study of retrospective and online event detection," in *Proceedings of the 21th Anuual International SIGIM Conference On Research and development in Information Retrieval*, New York NY USA, 1998, pp. 28-36.

[10] C.Chen, L, C.Wang,J.,-j.Bu W.Chen, "Online detection of bursty events and their evolution in news streams," in *J.Zheijang Univ,-Sci C 11*, 2010, pp. 340-355.

[11] Y. Mejova, C.Harris, P.Srinivasan, V.Ha-thuc, "Event Intensity Tracking in weblog collections," in *3rd Int'l AAAI Conference on Weblogs and Social media*, ICWSM, 2009.

[12] J.Zhang, J. Carbonell, C.Jin Y. Yang, "Toopic Conditioned novelty Detection ," in *Proceedings of the Eighth ACM SIGKDD International Conference On Knowledge Discovery and data mining , KDD* , New York NY USA , 2008, pp. 688-693.

[13] P.Mitra Q.Zhao, "Event Detection and Visualization for Social Streams ," in *1st Int'll AAAI Confernce on Weblogs and Social Media* , 2007.

[14] D.D., Yang, Y.M., Rose, T.G., Li, F., Lewis, "RCV1: a new benchmark collection for text categorization research.," *J. Mach. Learn. Res.*, pp. 361-397.