W2E: A Worldwide-Event Benchmark Dataset for Topic Detection and Tracking

Tuan-Anh Hoang, Khoi Duy Vo, Wolfgang Nejdl
L3S Research Center, Leibniz University of Hanover, Germany
hoang,khoi,nejdl@l3s.de

ABSTRACT

Topic detection and tracking in document streams is a critical task in many important applications, hence has been attracting research interest in recent decades. With the large size of data streams, there have been a number of works from different approaches that propose automatic methods for the task. However, there is only a few small benchmark datasets that are publicly available for evaluating the proposed methods. The lack of large datasets with fine-grained groundtruth implicitly restrains the development of more advanced methods. In this work, we address this issue by collecting and publishing W2E - a large dataset consisting of news articles from more than 50 prominent mass media channels worldwide. The articles cover a large set of popular events within a full year. W2E is more than 15 times larger than TREC's TDT2 dataset, which is widely used in prior work. We further conduct exploratory analysis to examine the dynamics and diversity of W2E and propose potential uses of the dataset in other research.

KEYWORDS

Topic detection, topic tracking, benchmark dataset

ACM Reference Format:
Tuan-Anh Hoang, Khoi Duy Vo, Wolfgang Nejdl. 2018. W2E: A Worldwide-Event Benchmark Dataset for Topic Detection and Tracking. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3269206.3269309

1 INTRODUCTION

The ability of detecting and tracking topics in document streams is crucial for applications for seeking and navigating information. These applications play important roles in a wide range of context where the users would like to have comprehensive view of not only new and trending topics but also the evolution of all the topics. This helps users to gain more insight from the streams for further consideration or making appropriate decisions. For example, politicians want to track public attention through news and social media for better campaigning [14]. Similarly, companies and advertisers want to know customers' opinions expressed in their comments for designing better products and services [4]. Government offices and organizations want to closely follow the development of disasters and crises for better responses and planning [12].

As document streams grow rapidly in both size and frequency, topic detection and tracking has been a challenging problem. There are many prior works that propose different automatic methods for the problem. Although these methods are reported performing well, they have been evaluated mostly on small datasets. Often, prior work's experiments are conducted on TREC's TDTs datasets [8, 13, 15]. These datasets are obsolete and small: they were collected around year 2000 and each dataset has only around 20K documents with groundtruth topic(s) [5]. A few works employ more recent datasets, e.g., sets news articles collected by news aggregating services [9, 11, 15] and abstracts of scientific papers [16]. However, these datasets are also small and/or do not have well defined groundtruth topics for documents: topics are defined as high level categories such as *finance*, *entertainment*, and *sport* for news articles, and publishing venues for papers' abstracts. Some works use synthetic datasets in which documents are bag-of-words that are randomly sampled from a pre-defined probabilistic model [1, 6]. Other works leverage static corpora to synthesize document streams [9]. Lastly, there are works that rely on qualitative evaluation of returning topics using domain-specific datasets [2, 10, 16], which requires domain expertise, hence hard to generalize to other domains. There has been a lack of large and diverse datasets that are publicly available. This introduces an implicit obstacle for developing more efficient and effective methods for working with large and complex document streams in practice.

In this work, we aim to address the aforementioned obstacle by constructing and publishing a large dataset that consists of documents with detailed temporal information and groundtruth topic. We also want the topics to be well defined and dynamic so that the collected dataset is representative enough to serve as medium for evaluating topic detection and tracking methods. To do that, we collect news articles about events and use the events to define the articles' groundtruth topic. We select a large set of highly popular events worldwide and a number of prominent mass media to collect news articles from. These selections have several advantages. Firstly, worldwide popular events are often highly dynamic and attract much public attention, hence are well reported in news. This allows us to have enough number of articles for each event. Secondly, these events are often well documented, e.g., in Wikipedia, hence we can obtain well described topics. Thirdly, prominent mass media often follow world events closely and report more complete aspects of the events. Articles collected from these media therefore have better coverage for the events. Lastly, these media often express in their articles different views toward the same event. The collected dataset therefore contains diversified views of the events.

CIKM '18, October 22–26, 2018, Torino, Italy
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3269309

Our constructed dataset is called **W2E**, which abbreviated for “*w*orld*w*ide *e*vent", and made publicly available on our supporting webpage at *https://sites.google.com/site/w2edataset/*. The rest of the paper is organized as follows. We describe the data collection process in Section 2. Next, we present exploratory analysis to examine the collected dataset in Section 3. We then propose some potential uses of the dataset in research work other than topic detection and tracking in Section 4. Finally, we conclude the paper and discuss some directions for future works in Section 5.

2 DATA COLLECTION

We employ the following steps to construct W2E dataset. The result of each step is also publicly available on our supporting webpage mentioned above.

Event selection. We rely on Wikipedia's Current Event portal[1] (WCEP) to select the set of events. WCEP is a special service of Wikipedia that allows users worldwide to summarize popular events happen everyday. WCEP provides a platform for users to edit events' short summaries and link them with other pages containing detailed information. For each day, each event happened in the day is assigned with one of several categories, e.g., *Politics and elections*, *Sports*, and *Disasters and accidents*. Events of the same long-run story may be grouped together under the same topic. Both the categories and topics are defined by the users.

For this work, we select all the events happened in year 2016 that are summarized in WCEP. We choose year 2016 as there are many popular long-run stories happened in the year, including US presidential election, UK's European Union membership referendum, Middle East wars, Summer Olympic, and disasters in North America. For events that are not grouped under any topic, we create a new topic for each event. In total, we obtain 5,160 events, which are assigned into 10 categories and 3,083 topics.

Topic merging. In the previous step, events of the same long-run story may not be grouped into a single topic but multiple topics. To overcome this, we manually merge topics together to recover the long-run story. Precisely, starting from the last day of the selected events, for each topic t that starts on day d (i.e., the first event of t happens on d), we scan through all the topics t' that (1) start before day d, and (2) do not end earlier than 3 weeks before d (i.e., the last event of t' happens at most 21 days before d). We then carefully examine the description and detailed information of topics t and $t'$to determine if they belong to the same long-run story. If there is at least one topic that belong to the same long-run story with t then t is merged to the topic that starts most recent to d. That is, each topic can be merged to at most one other topic starting before it. This is repeated until no further merging is found. At the end, we obtain 2,264 topics.

News article collection. For each selected events, we employ Google's search engine to retrieve news articles about the event that are written in English. To do this, we first manually construct a query from examining the event's summarization and its long-run story (if there is one found in the previous step). We then use the query to search with time filter option is set to the duration from one week before to one week after the day when event happens. By this we also retrieve the published date of the returning articles. We also set the “site" option of the search to each of the 52 selected news agencies' website. The selected news agencies includes popular media companies worldwide (e.g., CNN, BBC, Fox News, and English version of Russian Today, China Daily, and Aljazeera, etc.), and well-known blogs and news publishers in different domains (e.g, TechCruch (for *Technology*), ESPN (for *Sport*), and Daily Kos and Breitbart News (for *Politics*), etc.). Please refer to our supporting webpage for full list of all selected new agencies.

Table 1: Basic statistics of TDT2 and W2E datasets

Dataset	TDT2	W2E
#Articles	12,729	207,722
#Categories	NA	10
#Topics	193	2,015
#Events	NA	4,501
Durations(#days)	177	377

In total, we retrieve from Google's search engine 1,123,721 URLs to 936,500 unique news articles. For each article, its relevance to the corresponding event is manually judged by researchers who are knowledgeable about the events and well informed about this research work. The relevance is determined based on the event's description, its long-story story (if there is one), and the article's title and content. At the end, we get 258,145 relevant articles. We also filter out articles that contain little text but mostly media objects (e.g., video and photo collections) and tables (e.g., election and sport results). This results in 218,067 remaining articles. We then use DiffBot's Article API[2] to extract textual content of the articles. The API takes as input the URL of an article and returns as output a JSON object containing the article's title, clean text, and other information such as links to media objects embedded in the article. Finally, we obtain 207,722 news articles with clean text that are relevant to 4,501 events of 2,015 topics.

3 DATA EXPLORATION

We show in Table 1 some basic statistics of W2E dataset. For comparison, we also show in the table the same statistics of TREC's TDT2 dataset[3] - a widely used dataset in prior work - when the information is available. For TDT2 dataset, the statistics are computed from articles that are annotated with grouptruth topic label(s). The table clearly shows that W2E dataset is not only more than 15 times larger than TDT2 dataset but also spans a double duration and contains richer information for each articles. In the following, we present exploratory analysis to examine the diversity and temporal dynamics of W2E dataset.

3.1 Diversity Analysis

Figure 1 shows the numbers of topics, events, and articles by categories in W2E dataset. The figure clearly shows that the dataset is highly diverse in content. Figure 2 shows the number of articles collected from each selected news agency. We were able to collect articles from 48 news agencies, while TDT2 dataset's articles were collected from 9 news agencies. The figure shows that while the major part of articles is collected from agencies in English speaking countries, there is also a significant part from agencies in non-English speaking countries. Moreover, within the English-speaking countries, the agencies with different views contribute similar number

[1] https://en.wikipedia.org/wiki/Portal:Current_events

[2] https://www.diffbot.com/products/automatic/#article

[3] https://catalog.ldc.upenn.edu/LDC2001T57

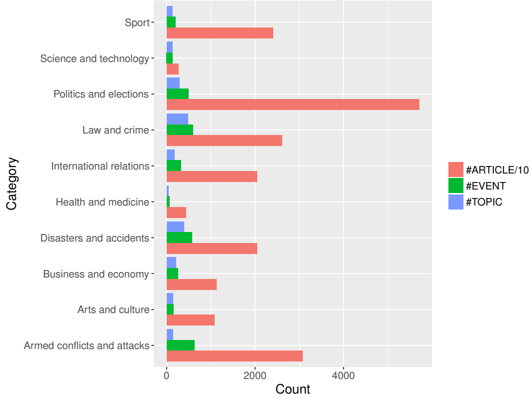


Figure 1: Numbers of topics, events, and articles by categories in W2E dataset. The number of articles is scaled down by 10 times for a better visualization

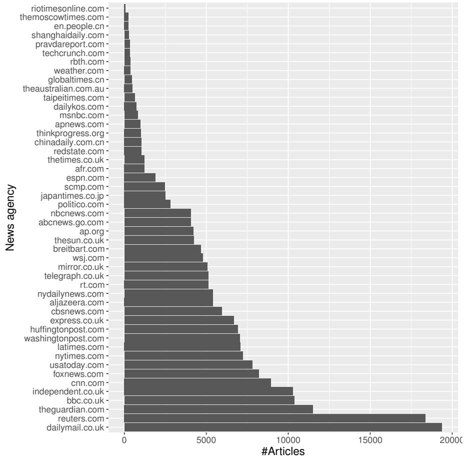
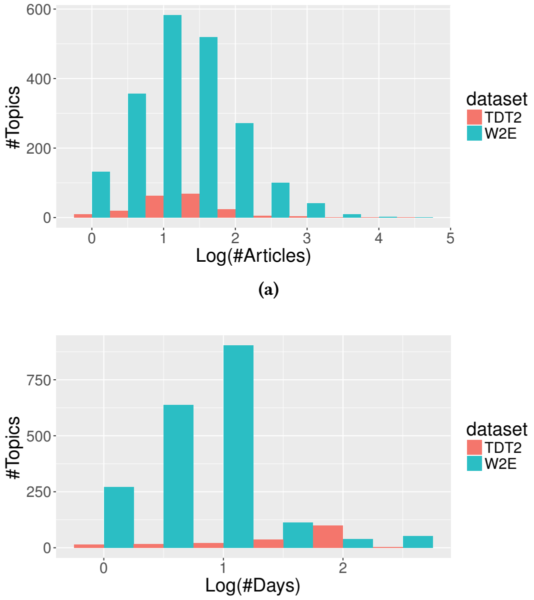


Figure 2: Number of articles collected from each news agency in W2E dataset

of articles (e.g, Washington Post and New York Times vs Fox News and Breitbart in US, and Daily Mail vs BBC and The Guardian in UK). Also Reuters - which is considered unbiased - contributes second most number of articles. W2E therefore contains diverse set of views. Lastly, Figure 3 (a) and (b) show the distribution of topics in both W2E and TDT2 datasets by size - i.e., the number of articles in each topic, and duration - i.e., the number of days from the first articles to the last articles of each topic, respectively. The figures show that both the two datasets have topics of varied size and duration, though W2E dataset has many more topics with larger sizes and span longer durations.



(a)

(b)

Figure 3: Distribution of topics by (a) number of articles and (b) duration

3.2 Dynamics Analysis

Figures 4 (a), (b), and (c) show the daily counts of articles, new topics, and active topics respectively in both the two datasets. Here, a topic is new or active for a day if it has the first article or at least one article published on the day respectively. The figures clearly show that, across time, the W2E dataset is much more dynamic then TDT2 dataset, in both number of articles, and number of new and active topics. Finally, Figure 5 shows the number of articles belong to the categories everyday. Here, an article belongs to a category if the article's corresponding event is assigned to the category. The figure shows that both the number and the proportion of articles belong to each category changes significantly from day to day. That means the content of W2E dataset is highly dynamic across time.

4 OTHER POTENTIAL USE CASES

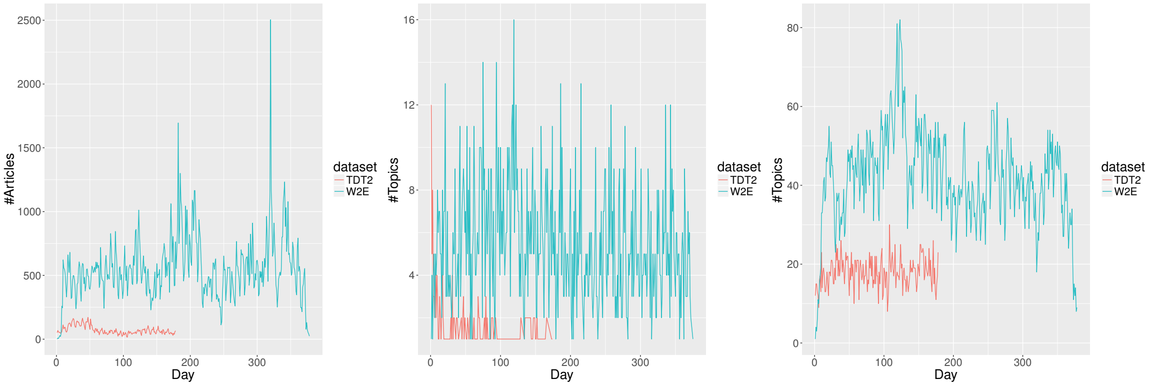
We propose here some potential use cases of W2E dataset other than in topic detection and tracking related research.

Multi-View clustering. Since each article in W2E dataset comes with a set of attributes including published date, event, topic, and category, the dataset can be used for evaluating multi-view clustering methods in which each attribute is served as a view [3].

Event and stream summarization. Remind that each event in W2E dataset is provided with a well written and concise summary in Wikipedia Current Event portal. The set of summaries of events hence forms a high quality temporal summary of articles in the dataset. Therefore, W2E dataset is reliable for evaluating event and stream summarization methods [7].

5 CONCLUSION

We have constructed and published a benchmark dataset for topic detection and tracking research. Our dataset consists of English articles from a number of prominent news agencies about a large set of worlwide popular events. The dataset is much larger and spans a longer time duration than existing publicly available ones, while containing richer information and covering much more diverse and temporally dynamics topics. For the future work, we would like to further enrich the constructed dataset. This includes collecting articles from more news agencies and written in different



(a) (b) (c)

Figure 4: Number (a) articles, (b) new topics, and (c) active topics in each day starting from the first day of each dataset

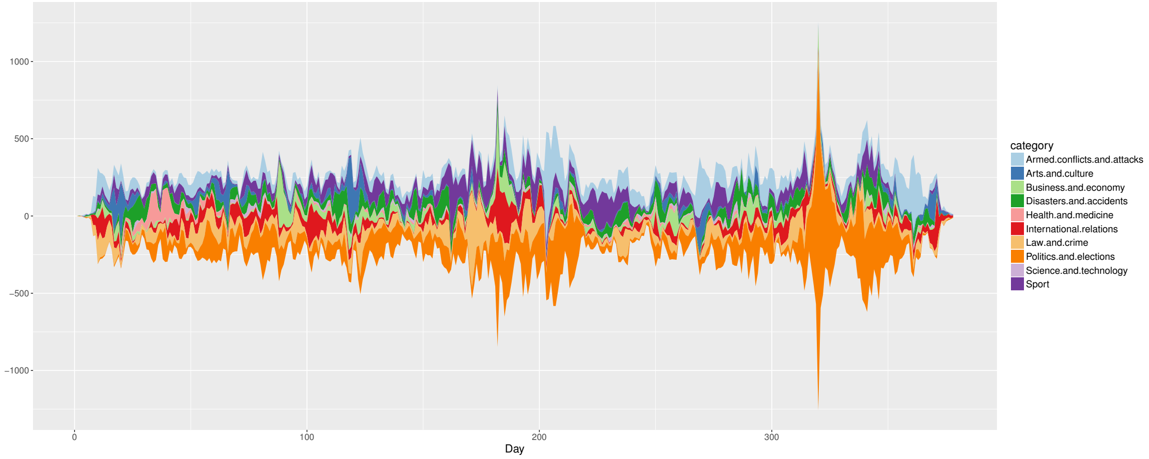


Figure 5: Number of articles (as represented by the stream's width) by categories in each day in W2E dataset

languages, and also collecting media content other than text about the selected events.

ACKNOWLEDGMENT

This research is supported by the ERC Grant (339233) ALEXANDRIA.

REFERENCES

[1] Amr Ahmed and Eric P Xing. 2010. Timeline: A Dynamic Hierarchical Dirichlet Process Model for Recovering Birth/Death and Evolution of Topics in Text Stream. In *UAI*.

[2] Adham Beykikhoshk, Ognjen Arandjelović, Svetha Venkatesh, and Dinh Phung. 2015. Hierarchical Dirichlet process for tracking complex topical structure evolution and its application to autism research literature. In *PAKDD*.

[3] Steffen Bickel and Tobias Scheffer. 2004. Multi-View Clustering. In *Proceedings of the Fourth IEEE International Conference on Data Mining*.

[4] Yan Chen, Hadi Amiri, Zhoujun Li, and Tat-Seng Chua. 2013. Emerging topic detection for organizations from microblogs. In *SIGIR*.

[5] Christopher Cieri, Stephanie Strassel, David Graff, Nii Martey, Kara Rennert, and Mark Liberman. 2002. Corpora for topic detection and tracking. In *Topic detection and tracking*. Springer, 33–66.

[6] Avinava Dubey, Ahmed Hefny, Sinead Williamson, and Eric P Xing. 2013. A nonparametric mixture model for topic modeling over time. In *SDM*.

[7] Tao Ge, Lei Cui, Baobao Chang, Sujian Li, Ming Zhou, and Zhifang Sui. 2016. News stream summarization using burst information networks. In *EMNLP*.

[8] Qi He, Kuiyu Chang, Ee-Peng Lim, and Arindam Banerjee. 2010. Keep it simple with time: A reexamination of probabilistic topic detection models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 10 (2010), 1795–1808.

[9] ICDM. 2015. Modeling emerging, evolving and fading topics using dynamic soft orthogonal nmf with sparse representation.

[10] Yookyung Jo, John E Hopcroft, and Carl Lagoze. 2011. The web of topics: discovering the topology of topic evolution in a corpus. In *WWW*.

[11] Xiangfeng Luo, Junyu Xuan, and Guangquan Zhang. 2016. Measuring the semantic uncertainty of news events for evolution potential estimation. *TOIS* (2016).

[12] Leysia Palen and Kenneth M Anderson. 2016. Crisis informatics - New data for extraordinary times. *Science* 353, 6296 (2016), 224–225.

[13] Ankan Saha and Vikas Sindhwani. 2012. Learning Evolving and Emerging Topics in Social Media: A Dynamic Nmf Approach with Temporal Regularization. In *WSDM*.

[14] Ben Sayre, Leticia Bode, Dhavan Shah, Dave Wilcox, and Chirag Shah. 2010. Agenda setting in a digital age: Tracking attention to California Proposition 8 in social media, online news and conventional news. *Policy & Internet* (2010).

[15] Carmen K Vaca, Amin Mantrach, Alejandro Jaimes, and Marco Saerens. 2014. A time-based collective factorization for topic discovery and monitoring in news. In *WWW*.

[16] Xiaolong Wang, Chengxiang Zhai, and Dan Roth. 2013. Understanding Evolution of Research Themes: A Probabilistic Generative Model for Citations *(KDD)*.