Efficient Tracking of Breaking News in Twitter

Tuan-Anh Hoang
GESIS Leibniz Institute for the Social Sciences, Germany
tuan-anh.hoang@gesis.org

Thi-Huyen Nguyen, Wolfgang Nejdl
L3S Research Center, Germany
{nguyen,nejdl}@l3s.de

ABSTRACT

We present an efficient graph-based method for filtering tweets relevant to a given breaking news from large tweet streams. Unlike existing models that either require manual effort, strong supervision, and/or not scalable, our method can automatically and effectively filter incoming relevant tweets starting from just a small number of past relevant tweets. Extensive experiments on both synthetic and real datasets show that our proposed method significantly outperforms other methods in filtering the relevant tweets while being as fast as the most efficient state-of-the-art method.

ACM Reference Format:
Tuan-Anh Hoang and Thi-Huyen Nguyen, Wolfgang Nejdl. 2019. Efficient Tracking of Breaking News in Twitter. In *11th ACM Conference on Web Science (WebSci '19), June 30-July 3, 2019, Boston, MA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3292522.3326058

1 INTRODUCTION

Real-time acquisition of tweets related to breaking news in real-time is vital for many important applications, such as social sensing and public opinion monitoring. This task is challenging due to the large scale of the Twitter data stream and the prevalence of noise and the wide range of covered topics in Twitter. These challenges require efficient methods for filtering relevant tweets from large streams. Existing works on real-time tweet filtering however require much manual effort and/or strong supervision which are often not available for the breaking news. Moreover, most of these methods are not scalable as they rely on supervised learning models whose training process is computationally expensive. We address these shortcomings by developing a lightweight method that requires minimal supervision yet obtains high performance. Precisely, we consider the filtering task in the following context:

Given a small set of tweets relevant to breaking news, we have to automatically decide, in real-time, if subsequent tweets in the tweet stream are relevant to the news.

In the following sections, we present a brief overview of our proposed method and some major experiment results. Readers are encouraged to refer to the full version of this paper [2] for the detailed description of the proposed method and the experiments.

2 METHODOLOGIES

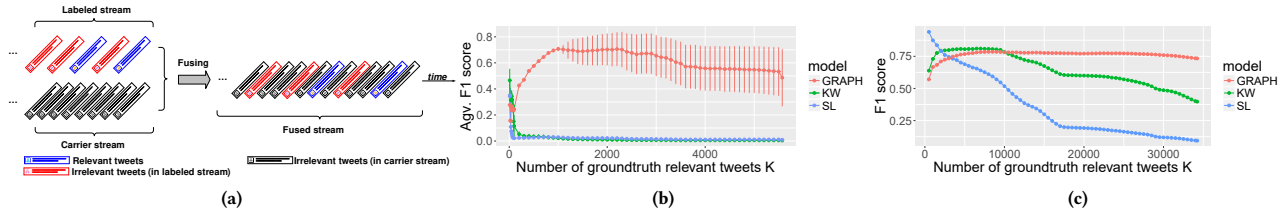
The main idea of our proposed method is to employ a graph based approach for measuring tweets' relevance to the breaking news and to the *background*. With background we refer to a representation of all topics in the tweet stream S that occur at around the same time with the news. As the stream consists of tweets in a vast variety of topics, we assume that incoming tweets are mostly relevant to the background and irrelevant to the news we want to track, and that news-relevant tweets are outliers. We therefore adopt a simple outlier detection approach to distinguish the news-relevant tweets based on the ratio of their relevance scores to the news and background. Our method consists of the following phases.

Initialization phase. We build the term graphs G_N and G_B from the set of initial relevant tweets T_N - which is given as input - and a set of tweets T_B that is randomly selected from all ones published within a short time window before the start filtering time. To do that, we preprocess each tweet by removing stopwords, punctuation marks, and special symbols (e.g., braces and quotations). The remaining tokens, which we call *terms*, are converted to lower-case, except the URLs embedded in tweets which are case sensitive. The node set then consists of all terms appearing in some preprocessed tweet(s). For a preprocessed tweet m and a pair of terms u and v, if both two terms appear in some window size L of m, then a undirected edge is drawn between u and v in the graph. A window size L of m is a sequence of at most L consecutive terms in m. The weight of edge (u, v) is the number of tweets in T that contain the edge. We then compute the importance of terms in G_N and G_B using Pagerank method. Finally, we compute the mean μ_N and variance σ_N of relevance scores of tweets in T_N - i.e., the relevant tweets - to the news, and compute mean μ_R and variance σ_R of *relevance ratio* of tweets in $T_N \cup T_B$. Here, a tweet's relevance ratio is the ratio between its relevance scores to the news and background. We will describe in detail the computation of tweets' relevance score in subsection below.

Filtering phase. For each incoming tweet m, its relevance scores to news r_N and background r_B are measured using the term graph G_N and G_B respectively. If $r_B > 0$, the ratio r_N/r_B is used to update μ_R and σ_R. Next, m's relevance label is determined as described below. If m is relevant then it is emitted as output and also added in T_N. The graph G_N is also updated using terms and edges induced by m. If m is irrelevant, with some probability $p < 1$, it is chosen for updating the background graph G_B and added into T_B. Regardless of m's relevance to the news, it is also used for checking whether updating of terms' importance is needed. The condition for this can be either time difference or number of (relevant) tweets found since the last update. If an update is needed, the oldest tweets in T_N and T_B are removed, and term graphs G_N and G_B are updated accordingly to the removed tweets. Also, term importance in the graphs is re-computed, and values of μ_N, σ_N, μ_R, and σ_R are re-assigned accordingly.

WebSci '19, June 30-July 3, 2019, Boston, MA, USA

ACM ISBN 978-1-4503-6202-3/19/06.
https://doi.org/10.1145/3292522.3326058



(a) (b) (c)

Figure 1: (a) Procedure for generating synthetic datasets, and experimental result on synthetic (b) and real datasets (c)

2.1 Computing Tweet Relevance Score

Given a term graph $G = (V, E)$ and a tweet m, our approach for measuring m's relevance to the news represented by G leverages both importance of terms and of edges in G. Precisely, the relevance score r of m is computed as follows.

$$r = \sum_{(u,v)\in E_m \cap E} \left[\pi(u)\frac{w(u,v)}{w(u,\cdot)} + \pi(v)\frac{w(u,v)}{w(\cdot,v)} \right] \quad (1)$$

In Equation 1, E_m is the set of edges induced by tweet m. $\pi(u)$ and $\pi(v)$ are importance of terms u and v in G respectively, $w(u,v)$ is the weight of edge (u,v), $w(u,\cdot)$ and $w(\cdot,v)$ are the summations of weights of all u's edges and v's edges respectively, i.e., $w(u,\cdot) = \sum_{(u,v')\in E} w(u,v')$ and $w(\cdot,v) = \sum_{(u',v)\in E} w(u',v)$.

2.2 Determining Tweets' Relevance Label

For a tweet m, if its relevance score to background $r_B = 0$, we decide that m is irrelevant. Otherwise, we determine relevance label of m based on its relevance scores to the news r_N and the ratio r_N/r_B. We assume that r_N follows a Gaussian distribution with mean μ_N and variance σ_N, while the ratio r_N/r_B follows a Gaussian distribution with mean μ_R and variance σ_R. We therefore measure the deviation d_N of respectively r_N and deviation d_R of r_N/r_B from their means as follows.

$$d_N = \frac{r_N - \mu_N}{\sigma_N} \quad \text{and} \quad d_R = \frac{(r_N/r_B) - \mu_R}{\sigma_R} \quad (2)$$

Then, m is assigned *relevance* label if $d_N \geq -1.3$ and $d_R \geq 1.05$. That means, only tweets whose relevance score is out of the bottom 10% and whose relevance ratio is among the top 15% are considered relevant to the news.

3 EXPERIMENTS

3.1 Datasets

Synthetic datasets with groundtruth. These datasets are synthesized following the procedure shown in Figure 1 (a). We re-used the set of tweets about the *Sandy Hurricane*[1] that were collected by [4] to simulate labeled stream. To simulate the carrier streams, we crawled tweets in 2017 using Twitter's realtime sample API. For each event, the labeled stream is fused into 15 different time durations of the carrier stream to create different datasets.

Real datasets with proxy groundtruth. Our real datasets consist of *2017 Westminster Attack*[2] and the tweet stream that are formed by crawling Twitter using its sample API as above. We used a Twitter-LDA topic model [5] to mine topics of tweets returned by the filtering methods.The obtained topics are then manually judged for relevance based on their top terms and top tweets. A tweet m is considered relevant if $p(z|m) \geq \theta = 0.6$ for some annotated-relevant topic z. Relevant tweets of all filtering methods are pooled to form the proxy groundtruth for evaluating their performance.

[1] https://en.wikipedia.org/wiki/Hurricane_Sandy
[2] https://en.wikipedia.org/wiki/2017_Westminster_attack

3.2 Results

Figure 1 (b) shows the F_1 scores over time of the filtering methods on 15 *Sandy Hurricane* datasets. Since all the datasets have the same groundtruth, we average their scores. Here we compare our proposed method -denoted by **GRAPH** with two baselines: **KW** proposed by Cotelo et al. [1] - which is the state-of-the-art keyword based methods, and **SL** proposed by Magdy et al. [3] - which is the state-of-the-art supervised learning based methods. Similarly, Figure 1 (c) shows the scores of the three methods on *2017 Westminster Attack* dataset. The figures show that both the two baseline methods have better performance than ours in a short time duration after the news happens when there are not many relevant tweets. However, their performance decreases rapidly later when there are many more relevant tweets. This is due to the fact that, at first, the baseline methods' filters are unigram-based and weakly trained by datasets with only a small number of truly relevant tweets (i.e., the set of input relevant tweets). The figures also show that our method is much more robust against the news' evolution as it obtains lower performance at earlier stages but significantly outperforms the baseline methods to obtain much higher performance consistently across subsequent states.

ACKNOWLEDGEMENT

This research is supported by the ERC Grant (339233) ALEXANDRIA and the DFG Grant (NI-1760/1-1) Managed Forgetting.

REFERENCES

[1] Juan M Cotelo, Fermin L Cruz, and Jose A Troyano. Dynamic topic-related tweet retrieval. *JAIST*, 65(3):513–523, 2014.
[2] Tuan-Anh Hoang, Thi-Huyen Nguyen, and Wolfgang Nejdl. Efficient tracking of breaking news in twitter. https://www.dropbox.com/s/uz54vtp8z3ort68/attt.pdf.
[3] Walid Magdy and Tamer Elsayed. Unsupervised adaptive microblog filtering for broad dynamic topics. *IPM*, 52(4):513–528, 2016.
[4] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *ICWSM*, 2014.
[5] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *ECIR*, 2011.