Real-time Filtering on Interest Profiles in Twitter Stream

Yue Fei Chao Lv Yansong Feng* Dongyan Zhao
{feiyue,lvchao,fengyansong,zhaody}@pku.edu.cn
Institute of Computer Science and Technology, Peking University

ABSTRACT

The advent of Twitter has led to the ubiquitous information overload problem with a dramatic increase in the amount of tweets a user is exposed to. In this paper, we consider real-time tweet filtering with respect to users' interest profiles in public Twitter stream. While traditional filtering methods mainly focus on judging relevance of a document, we aim to retrieve relevant and novel documents to address the high redundancy of tweets. An unsupervised approach is proposed to model relevance between tweets and different profiles adaptively and a neural network language model is employed to learn semantic representation for tweets. Experiments on TREC 2015 dataset demonstrate the effectiveness of the proposed approach.

CCS Concepts

•**Information systems** → **Document filtering;** *Data stream mining;*

Keywords

Real-time Filtering, Neural Network Language Model, Adaptive Thresholding

1. INTRODUCTION

Social media sites such as Twitter are increasingly becoming better sources of information owing to the timeliness of news and users' interactions with friends. However, the dramatic increase in the amount of information a user is exposed to, greatly improves the chances of the user experiencing information overload.

Our main objective is retrieving tweets relevant to user's interest without redundancy. To find the accurate relevance boundry for different profile, we retrieve a list of candidate posts for each profile from the background corpus and set the score at top k relevant posts as relevance threshold. To address the problem of vocabulary mismatch, we apply a neural network language mode to learn the semantics of texts and model the document representations. We evaluate our approach on the TREC 2015 microblog real-time filtering dataset. Experimental results show effectiveness of profile-biased adaptive thresholding method average pooling representation.

*Corresponding author

JCDL '16 June 19-23, 2016, Newark, NJ, USA

ACM ISBN 978-1-4503-4229-2/16/06.
DOI: http://dx.doi.org/10.1145/2910896.2925462

2. PROPOSED APPROACH

In this section we describe the proposed methodology for real-time filtering in tweet stream, which takes both relevance and redundancy into consideration. More specifically, given a real time tweet stream, where the tweet set $T = \langle t_1, t_2, t_3, ... \rangle$ arrives in a strictly chronological order, and users' interest profiles $P = \langle p_1, p_2, p_3, ... \rangle$, where each profile is denoted as a specific topic with a short phrase. our objective is to filter out relevant and novel tweet posts from the stream with respect to each interest profile.

2.1 Adaptive Thresholding

Traditional approaches filter tweets in a supervised manner [1, 2], which require labeled data and use fixed parameters across different profiles. Here we propose an unsupervised thresholding method to help identify interesting posts with regard to interest profiles in an adaptive manner. For each new coming document, our system will first decide whether it's relevant to the user's profile. The simplest way is to set a fixed threshold for every profile. However, a fixed threshold will result in two problems. On one hand, a fixed score threshold may not work well on different profile due to the variety of interest profiles. On the other hand, the content of relevant posts will change by time with a certain event involves or new subtopic emerges. Hence, the relevance threshold should be adaptive to both profile and time.

When initializing the relevance threshold for each profile, we run a Boolean Retrieval process to get candidate tweets relevant to the specific profile from the background corpus, i.e., the tweet collection of the last few days. The retrieved candidates are expected to be a reasonably accurate set since Boolean Retrieval is very strict. Then we compute the relevance scores between the candidates and the corresponding profile, rank the scores and cut out the top k scores as the initial threshold. Here k is an empirical parameter and can be set corresponding to the number of posts a user wish to read everyday. This threshold is timely updated with new coming posts in the same manner. In this way, we can get

a reasonable initial thresholds and adapt the threshold with the evolution of topics.

This redundancy problem is very typical in microblog, since users often post messages with the same, or very similar content, especially when reporting or commenting on news and events. We employ a tweet pool retaining posts that have already been pushed to help get rid of redundancy. When a new relevant tweet is identified, we compare it with all the delivered tweets and justify whether it's novel according to a fixed redundancy threshold. We choose a fixed weak threshold since we assume that users view redundancy not as sensitive as relevance so that only very "redundant" documents are considered redundant.

2.2 Tweet Representation

We apply the average pooling of word embedding vector to tackle the vocabulary mismatch problem caused by the shortness of tweets [5] . The average pooling of word embedding vector utilizes word embeddings in a low-dimensional continuous space where relevant words are close to each other. The relationships among words are embedded in their word vectors, providing a simple way to compute aggregated semantics for word collections such as paragraphs and documents. In the previous study, a simple average pooling approach was proposed to derive document vectors from word embeddings. Letting $c_{i,j}$ denote the word embedding of the j-th word token of document i, the document vector v_i can be computed as Eq. 1, where J_i denotes the number of word tokens in the document i.

$$v_i = \frac{1}{J_i} \sum_{j=1}^{J_i} c_{i,j} \tag{1}$$

3. EVALUATION

We evaluate our approach on the real-time filtering task of the TREC 2015 Microblog track [4]. The dataset is a sample from public tweet stream which begins from July 20, 2015 and lasts until July 29, 2015. We conduct preprocessing on the raw data including non-English elimination, stop words removal and stemming. There are 13,988,358 tweets left after preprocessing, while ignoring that some tweets are missed during sampling.

Totally 56 interest profiles are provided for evaluation. Each profile contains three parts: title, description and narrative. During the experiments, we find that both description and narrative introduce more noise than information for profiles. Hence, we will leverage the title information only for each profile. Two metrics are introduced for evaluating this task: the expected latency-discounted gain (ELG) and the normalized cumulative gain (nCG)[4]. ELG is a precision oriented metric while NCG values more recall.

We compare our tweet representation method with the BOW representation and paragraph vector representation [3]. Experimental results of the three representations with and without adaptive thresholding method are shown in table reftab:result. It's obvious that AP gains a statistically significant improvement in terms of ELG over BOW and PV, which implies the average pooling representation can better model semantic relations between tweet texts. Besides, the adaptive thresholding method significantly improves the filtering effectiveness using either filtering measures over those without. Moreover, our method is also comparable with the TREC 2015 best automatic run[4] in the last line. Figure 1 shows the performance of adaptive thresholding strategy with k ranging from 5 to 50, which illustrates the robustness of our method. Overall, the results show the effectiveness of our average pooling and adaptive thresholding strategy.

Table 1: Experimental results with different strategy. † denotes a statistically significant increase over BOW. Statistical significance is estimated with a paired t-test at ($p < 0.05$).

Run	ELG	nCG
BOW	0.2690	0.2415
BOW+Ada	0.2818 †	0.2722 †
PV	0.2709	0.2435
PV+Ada	0.3009 †	0.2745 †
AP	0.3037 †	0.2449
AP+Ada	**0.3067** †	**0.2806** †
TREC 2015 BEST	0.3150	0.2679

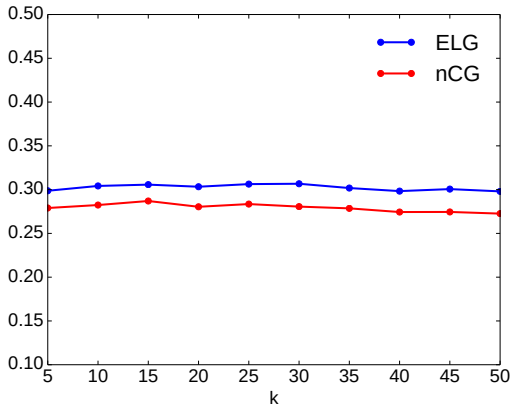

Figure 1: Sensitivity Analysis of k

4. ACKNOWLEDGMENTS

The work reported in this paper was supported by the National Natural Science Foundation of China Grant 61370116.

5. REFERENCES

[1] M. Albakour, C. Macdonald, I. Ounis, et al. On sparsity and drift for effective real-time filtering in microblogs. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 419–428. ACM, 2013.

[2] Y. Fei, Y. Hong, and J. Yang. Handling topic drift for topic tracking in microblogs. In *Advances in Information Retrieval*, pages 477–488. Springer, 2015.

[3] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*, 2014.

[4] J. Lin, M. Efron, Y. Wang, G. Sherman, and E. Voorhees. Overview of the trec-2015 microblog track. In *Proceedings of TREC*, volume 2015, 2015.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.