Information Systems 000 (2018) 1–15

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/is

Mining authoritative and topical evidence from the blogosphere for improving opinion retrieval

Jimmy Xiangji Huang[a,*], Ben He[b,*], Jiashu Zhao[c]

[a] Information Retrieval and Knowledge Management Research Lab School of Information Technology, York University, Ontario, Toronto M3J 1P3 Canada
[b] School of Computer & Control Engineering, University of Chinese Academy of Sciences Beijing 101408, China
[c] Data Science Lab, JD.com Beijing 100101, China

ARTICLE INFO

Article history:
Received 15 April 2017
Revised 8 February 2018
Accepted 11 February 2018
Available online xxx

Keywords:
Opinion mining
Blog retrieval
Sentiment analysis
Neural matching model

ABSTRACT

The rise of the Internet blogging has created a highly dynamic Web society that involves bloggers' views and opinions in response to real-world events. As an emerging research field, the blog post opinion retrieval requires finding not only relevant but also opinionated blog posts. Most of the current solutions are based on a dictionary of sentiment words for identifying subjective features from blog posts. In this paper, we propose to utilize novel evidence, namely the authoritative and topical evidence, for mining opinions from the blogosphere. We suggest that bloggers interested in controversial topics tend to express opinions in their posts, and therefore, it is beneficial to boost the ranking of blog posts written by such authors. We further improve our approach by extending with different sources of features, which is incorporated into a document-based neural matching model. Our experiments on the standard test data from the TREC 2006–2008 Blog track opinion finding task show that the proposed approach is capable of achieving remarkable improvements over strong baselines.

1. Introduction

The rise of the Internet blogging has created a highly dynamic Web publishing medium that attracts many grassroots Web authors. Indeed, numerous Web bloggers have formed an active social community, in which real-world events are promptly discussed, and opinions from different aspects are expressed. Most good-quality blogs are interactive, which allow visitors to leave comments on the blogs, and it is this interactivity that distinguishes them from other static websites. Therefore, blogging can be seen as a form of social networking. Indeed, bloggers do not only produce content to post on their blogs, but also build social relations with their readers and other bloggers. The so-called blogosphere, i.e. the collection of blogs on the Internet, is relatively new social media, which opens up several new interesting research areas.

A key feature that distinguishes blog content from other types of Web content is its subjective nature. Bloggers tend to express opinions and comments towards some given targets, such as persons, organizations or products. A study of a query log from a commercial blog search engine found that many blog queries seem to be related to uncovering public opinions about a given target [43]. For example, a user who is planning to buy a Nokia cell phone may wish to find out the opinions of other users in the blogosphere about how they rate its features.

Mining social media content to unveil latent information about people sentiment and opinions is drawing more and more attention [33,53]. There have been several studies on how to find opinions in the Natural Language Processing (NLP) community [4,26,34,35,38]. For example, Pang et al. [49] and Liu et al. [36] proposed to find opinions from movie reviews using machine learning and NLP techniques. Kang and Lee [26] mapped the content of texts in social media to Wikipedia-based features such as Wikipedia categories and article entities. Their approach is based on the assumption that the analyzed documents are already known to be relevant. However, building a retrieval system to uncover documents that are both opinionated and relevant remains a difficult challenge. Since 2006, the Text REtrieval Conference (TREC) has been running a Blog track and the corresponding opinion finding (**OF**) task for addressing this challenge, namely finding opinionated and relevant blog posts. The OF task is an articulation of a user search task, where a user is trying to uncover the public opinions that exist on the blogosphere, towards a given named-entity target.

Under the TREC Blog opinion finding task, the relevance of a blog post is defined at two levels. The first level, namely the topical relevance, assesses whether a given blog post, i.e. a permalink,

* Corresponding authors.
E-mail addresses: jhuang@yorku.ca (J.X. Huang), benhe@ucas.ac.cn (B. He), zhaojiashu1@jd.com (J. Zhao).

https://doi.org/10.1016/j.is.2018.02.002
0306-4379/

contains information about the target and is therefore relevant. The second level, namely the opinion relevance, assesses the opinionated nature of the blog post, if it was deemed relevant in the first assessment level [41,45]. The opinionated and relevant blog posts are a subset of the relevant ones. A blog post OF system is evaluated by looking at how the system performs over a so-called *baseline* for which no opinion feature is applied. A baseline is referred to as a retrieval system that aims to retrieve as many relevant documents as possible, regardless of their opinionated nature. In other words, a baseline system aims to optimize its performance for the topical relevance, without taking opinion relevance into consideration.

Various blog post opinion finding approaches have been proposed in the context of the TREC Blog track OF task [41,45]. However, the experimental results in this task have demonstrated considerable difficulty in improving strong retrieval baselines [46]. Indeed, only a handful of groups achieved an improvement over their baselines, using techniques such as NLP (for example integrating OpinionFinder [15]), dictionary-based statistical methods [14], or SVM classifiers [72].

In general, most of the approaches proposed to utilize different sources of opinion-based evidence, mostly heuristically, such as a long list of pre-compiled subjective terms, or rare terms, for detecting opinionated documents. However, as Web blogs tend to be opinionated in nature, the interests of the bloggers, or general topics that the bloggers are usually talking about, can be a complementary source of evidence for mining opinions from their blog posts. This kind of evidence can be utilized to evaluate a blogger's willingness to express opinion to some extent. In this paper, we propose to utilize novel evidence, namely the authoritative and topical evidence, for mining opinions from the blogosphere. In particular, the authoritative evidence is interpreted as the relatedness of a blogger/feeds content to controversial topics, while the topical evidence is defined as the topical areas that the blog posts discuss about. We believe that the novel authoritative and topical evidence are useful for blog post opinion mining and retrieval in that they directly capture the opinionated nature of the blogosphere. More details about the evidence will be provided in Section 5.

A major contribution of this paper is to mine and utilize authoritative and topical evidence for improving the retrieval performance of opinionated blog posts. In particular, we propose statistical models to extract topical terms and calculate opinion weights of a blog. We suggest that a blogger's interests can be an important source of evidence for subjectivity. For instance, if a blogger is enthusiastic in blogging about certain controversial issues, such as the Iraqi war, or the U.S. foreign policy, the blogger is likely to share her opinions on these topics with others, and her blog posts are therefore likely to contain expressed opinions. In this case, blog posts written by this author should be boosted in the ranking. Note that the expressed opinions by bloggers can be either positive, negative, or even neutral, since writing in a blog that "I have no opinion about it" is something different from not mentioning any opinion. In our proposed approach, we build a profile for each blogger, which includes all her blog posts. We estimate the generation probability of a list of topical words extracted from the training queries taken from previous test queries used in the TREC Blog track opinion finding task. Finally, the generation probability is combined with the retrieval model to produce a final ranking of blog posts. Compared with the traditional dictionary-based approach, our proposed approach is very promising as shown in Section 9.2. Our proposed approach does not use any additional resources to extract opinion terms, which provides a new and promising avenue to mine opinion in the blogosphere. On the other hand, the topical terms extracted using our approach can be served as a source of evidence for opinions complementary to the dictionary-based approaches. As shown in Section 6, our approach can be extended with different sources of information, including term proximity, opinion lexicon, and pseudo relevance feedback to perform comparably to the state-of-the-art.

The remainder of this paper is organized as follows. Section 2 explains the TREC paradigm for experimentation and evaluation of the opinion retrieval systems. Section 3 shows the novelty of our work by a summary of previous research on the blog post opinion finding. Section 4 introduces the Okapi system that is used as a baseline throughout this paper. Next, our approach to the blog post opinion finding by mining authoritative and topical evidence is proposed in Section 5. Furthermore, our proposed approach is extended with quite a few sources of information wrapped up in a neural matching model in Section 6. Section 7 describes the experimental setup for the evaluation of our proposed approach, and Section 8 presents the related results. In addition, Section 9 compares our approach with other baselines and leads to the related discussion. Section 10 provides a discussion on why topical and authoritative evidence are important for blog post opinion finding. Finally, Section 11 concludes the work and suggests future research directions.

2. The TREC paradigm for experimentation of opinion retrieval

This task uses the Blog06 collection, including 100,649 blog feeds spanned from December 2005 to February 2006. During that time, XML feeds, their corresponding homepages (the blog as seen by the users) and permalink documents (single blog posts and their corresponding comments), were fetched and saved [40]. The collection is 148GB in size, with three main components consisting of 38.6 GB of XML feeds (i.e. the blog), 88.8 GB of permalink documents (i.e. a single blog post and all its associated comments) and 28.8 GB of HTML homepages (i.e. the main entry to the blog). The permalink documents are used as a retrieval unit for the OF task. There are over 3.2 million permalink documents in the Blog06 collection. In this paper, we follow the TREC setting and experiment on the permalink documents.

The TREC blog OF task had been running since 2006 until 2008. Each year, a dozen of research groups and organizations participated in the task by submitting their experimental results on the Blog06 collection using their own retrieval systems. Each participating system is evaluated using a set of topics and their associated relevance assessment. In total, there are 150 topics from the TREC 2006–2008 OF tasks, where 50 topics are used in each year. For example, a Blog OF topic is included in Fig. 1.

Each participating system in the OF task is required to return the top-1000 ranked blog posts and their associated comments for each topic. The relevance assessment procedure for the documents retrieved for the topics has two levels, namely the topical relevance and the opinion relevance as explained in the previous section.

A system's performance in retrieving opinionated blog post is evaluated by how the system performs over a baseline, whose aim is to optimize the retrieval performance for topical relevance, instead of opinion relevance. Note that in the TREC OF task, submission of baselines was encouraged in TREC 2006, and has been mandatory since[1].

Our experiments in this paper follow this paradigm. We examine if our proposed method brings improvement when running on top of strong and robust baselines. Moreover, we experiment at the second relevance assessment level, which takes into account only blog posts that are both relevant and opinionated.

[1] http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG.

Please cite this article as: J.X. Huang et al., Mining authoritative and topical evidence from the blogosphere for improving opinion retrieval, Information Systems (2018), https://doi.org/10.1016/j.is.2018.02.002