Information Processing and Management 52 (2016) 513–528

Contents lists available at ScienceDirect

Information Processing and Management

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

Unsupervised adaptive microblog filtering for broad dynamic topics

Walid Magdy [a], Tamer Elsayed [b]

[a] Qatar Computing Research Institute, HBKU, Doha, Qatar
[b] Department of Computer Science and Engineering, College of Engineering, Qatar University, Doha, Qatar

ARTICLE INFO

Article history:
Received 28 March 2015
Revised 17 November 2015
Accepted 24 November 2015
Available online 6 January 2016

Keywords:
Twitter
Microblog filtering
Unsupervised adaptive filtering
Broad dynamic topics
Arabic tweets

ABSTRACT

Information filtering has been a major task of study in the field of information retrieval (IR) for a long time, focusing on filtering well-formed documents such as news articles. Recently, more interest was directed towards applying filtering tasks to user-generated content such as microblogs. Several earlier studies investigated microblog filtering for focused topics. Another vital filtering scenario in microblogs targets the detection of posts that are relevant to long-standing broad and dynamic topics, i.e., topics spanning several subtopics that change over time. This type of filtering in microblogs is essential for many applications such as social studies on large events and news tracking of temporal topics. In this paper, we introduce an adaptive microblog filtering task that focuses on tracking topics of broad and dynamic nature. We propose an entirely-unsupervised approach that adapts to new aspects of the topic to retrieve relevant microblogs. We evaluated our filtering approach using 6 broad topics, each tested on 4 different time periods over 4 months. Experimental results showed that, on average, our approach achieved 84% increase in recall relative to the baseline approach, while maintaining an acceptable precision that showed a drop of about 8%. Our filtering method is currently implemented on TweetMogaz, a news portal generated from tweets. The website compiles the stream of Arabic tweets and detects the relevant tweets to different regions in the Middle East to be presented in the form of comprehensive reports that include top stories and news in each region.

1. Introduction

Microblogging sites, such as Twitter, are currently one of the main platforms for exchanging real-time information and discussions. This large amount of communicated information motivated many social scientists to study public response towards major events and topics on social media (Ali, Magdy, & Vogel, 2013; Chen et al., 2013; Conover, Ratkiewicz, Francisco, & Gonçalves, 2011; Magdy, 2013; Magdy & Elsayed, 2014; Phuvipadawat & Murata, 2010; Teevan, Ramage, & Morris, 2011; Vieweg, Hughes, Starbird, & Palen, 2010). However, selecting relevant information out of social posts requires, in many cases, scalable and adaptive information filtering techniques, since the topics to be tracked can be of broad and dynamic nature. This has appeared in many recent applications where it was essential to follow broad topics for long periods of time from the social media. These applications are such as news reporting (Elsawy, Mokhtar, & Magdy, 2014; Magdy, 2013), political events analysis (Borge-Holthoefer, Magdy, Darwish, & Weber, 2015; Conover et al., 2011), crisis management (Olteanu, Castillo, Diaz, & Vieweg, 2014; Vieweg et al., 2010), and many others. Most of these topics consist of multiple subtopics that get changed and drifted over time. For example,

E-mail addresses: wmagdy@qf.org.qa (W. Magdy), telsayed@qu.edu.qa (T. Elsayed).

http://dx.doi.org/10.1016/j.ipm.2015.11.004
0306-4573/

following microblogs related to "US Elections" requires tracking posts about several subtopics and related entities such as candidates, campaigns, political views, election process, etc. Moreover, the subtopics are also dynamic. For example, "debates" is an important one before elections, while "election results" is the most important during voting. Sometimes subtopics span very short period of time, e.g., press statements by candidates. The broad topics can even be more general or broader than the earlier example, e.g. "US Politics", which includes "US Elections" as a temporal subtopic. Similarly, in a study about public response to a long-standing event such as "the Syrian conflict", tracking relevant microblogs is not a straightforward task, since it requires the coverage of as much of the related posted content as possible to cope with the developing sub-events. Tracking such kind of topics, which last for a long period of time and consist of several sub-events that change dramatically over time, requires a set of selective queries (rather than just a single one) to be updated periodically for effectively covering different aspects of these topics.

Two common user-based features that are provided by microblogging platforms are widely used for filtering. The first is the "follow" feature that allows a user to follow other accounts of entities, persons, or events to get their tweets into the user's timeline. The other method for following specific microblogs is searching for given hashtags (i.e., the character "#" followed by a tag (e.g., #Syria) that generally indicates the topic of the mentioning tweet), which is a common way for users to get updates on topics that are indicated by those hashtags. This method is less strict in filtering information, where more tweets are generally presented to user. However, many off-topic tweets would be retrieved because of the misuse of hashtags by some users. Moreover, many tweets that are relevant to the topic may not include the hashtag itself, and thus will be missed.

In this paper, we present an unsupervised approach for microblog filtering that aims at following broad and dynamic topics. Our main objective is to boost *recall* by retrieving a large number of relevant microblogs, while preserving high *precision* to avoid bothering users with irrelevant feeds. The main challenge lies in capturing relevant microblogs to temporal short-term subtopics that might only appear for a short period of time. Our approach initially gets a user-defined fixed query or set of queries that cover the most static part of the target topic and retrieves initial set of hopefully-relevant microblogs using simple Boolean search. The initial retrieved set of microblogs is used in a novel, *unsupervised*, and *adaptive* manner to train a binary classifier that is used for detecting other relevant content within a stream of microblogs. This classifier is trained and updated automatically and regularly to adapt to the dynamic nature of the tracked topic without any user intervention.

We tested the approach on six hot broad and dynamic topics to simulate the topics typically studied in social studies. We tested our approach with the topics on four different days from four different months to measure the performance consistency of the approach over time. This forms a set of 24 testing points for our approach. Thousands of relevance assessments were created for the evaluation process. Topics were selected from three different domains: politics, war, and sports. Three of the test topics were each represented by a single hashtag, and the other three topics were well-formulated by a rich set of accurate Boolean queries to achieve an initial high-precision set of microblogs. A stream of 3–4 million Arabic tweets per day was used in our experiments with the goal of identifying tweets that are relevant to each of the test topics. Our approach achieved to retrieve large number of relevant microblogs that do not include any of the search queries but still are on topic based on the events on-going on the tested dates. Compared to the baseline approach; experimental results showed a boost in recall of average 84%, while precision is dropped by only 8%.

Our filtering approach is implemented in the backend of TweetMogaz[1] (Magdy, 2013), which is a framework for following news happening in different regions from Twitter, and presents them in the form of comprehensive report that includes pseudo-articles (Elsawy et al., 2014) and top posts about the topic. The service is live and shows the effectiveness of on practical use on real-life streams of data.

We make the developed test set along with 6 broad topics and relevance assessments for these topics over 4 different days publically available for potential future research studies[2].

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 defines our problem. Section 4 describes our proposed approach. Section 5 explains experimental setup. Section 6 reports the results and highlights the success of the approach. Section 7 introduces TweetMogaz, a platform for tweets search and filtering, which uses our filtering approach and demonstrates its effectiveness in practice. Finally, Section 8 concludes the paper and provides possible future directions.

2. Related work

In this section we show examples of various applications that depend on following broad topics on social media. Then we present some prior work on microblog filtering.

2.1. Microblog filtering applications

The purpose of collecting microblogs, especially for broad and dynamic topics, extends beyond reading them; it actually includes analysis, summarization, prediction, or classification. Some of the applications for such task are news monitoring (Elsawy et al., 2014; Magdy, 2013; Phuvipadawat & Murata, 2010), crisis management (Olteanu et al., 2014; Vieweg et al., 2010), customer

[1] www.tweetmogaz.com

[2] http://alt.qcri.org/~wmagdy/Resources/FilteringData.htm