Web Semantics: Science, Services and Agents on the World Wide Web 26 (2014) 36–57

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

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

Linked knowledge sources for topic classification of microposts: A semantic graph-based approach

Andrea Varga [a,*], Amparo Elizabeth Cano Basave [b], Matthew Rowe [c], Fabio Ciravegna [a], Yulan He [d]

[a] Organisations, Information and Knowledge Group, The University of Sheffield, UK
[b] Knowledge Media Institute, The Open University, UK
[c] School of Computing and Communications, Lancaster University, UK
[d] School of Engineering and Applied Science, Aston University, UK

ARTICLE INFO

Article history:
Received 28 June 2013
Received in revised form
20 March 2014
Accepted 7 April 2014
Available online 13 April 2014

Keywords:
Linked knowledge sources
Semantic concept graphs
Topic classification

ABSTRACT

Short text messages a.k.a Microposts (e.g. Tweets) have proven to be an effective channel for revealing information about trends and events, ranging from those related to Disaster (e.g. hurricane Sandy) to those related to Violence (e.g. Egyptian revolution). Being informed about such events as they occur could be extremely important to authorities and emergency professionals by allowing such parties to immediately respond.

In this work we study the problem of topic classification (TC) of Microposts, which aims to automatically classify short messages based on the subject(s) discussed in them. The accurate TC of Microposts however is a challenging task since the limited number of tokens in a post often implies a lack of sufficient contextual information.

In order to provide contextual information to Microposts, we present and evaluate several graph structures surrounding concepts present in linked knowledge sources (KSs). Traditional TC techniques enrich the content of Microposts with features extracted only from the Microposts content. In contrast our approach relies on the generation of different weighted semantic meta-graphs extracted from linked KSs. We introduce a new semantic graph, called category meta-graph. This novel meta-graph provides a more fine grained categorisation of concepts providing a set of novel semantic features. Our findings show that such category meta-graph features effectively improve the performance of a topic classifier of Microposts.

Furthermore our goal is also to understand which semantic feature contributes to the performance of a topic classifier. For this reason we propose an approach for automatic estimation of accuracy loss of a topic classifier on new, unseen Microposts. We introduce and evaluate novel topic similarity measures, which capture the similarity between the KS documents and Microposts at a conceptual level, considering the enriched representation of these documents.

Extensive evaluation in the context of Emergency Response (ER) and Violence Detection (VD) revealed that our approach outperforms previous approaches using single KS without linked data and Twitter data only up to 31.4% in terms of F1 measure. Our main findings indicate that the new category graph contains useful information for TC and achieves comparable results to previously used semantic graphs. Furthermore our results also indicate that the accuracy of a topic classifier can be accurately predicted using the enhanced text representation, outperforming previous approaches considering content-based similarity measures.

1. Introduction

Social media posts, and in particular Microposts collected from Twitter, have been found to contain useful information for many applications including disaster detection [1], seasonal mood level changes [2], tracking influenza rates [3], box-office revenue forecast [4], political elections [5], stock market prediction [6], etc.

* Corresponding author. Tel.: +44 07508989022.
E-mail address: varga.andy@gmail.com (A. Varga).

http://dx.doi.org/10.1016/j.websem.2014.04.001
1570-8268/

For instance, during the widespread protest in Egypt in 2013, Microposts were found to provide early warning signals of violent events; such events were reported much faster than traditional media sources.[1] The real-time identification of such events could be extremely important to authorities and emergency professionals by allowing such parties to immediately respond.

However, the classification of such messages poses unique challenges, due to the special characteristics of the messages (i) the limited length of Microposts (up to 140 characters), restricting the contextual information necessary to effectively understand and classify them; (ii) the noisy lexical nature of Microposts, where new terminology and jargon emerges as different events are discussed; (iii) the large topical coverage of Micropost.

Existing approaches have addressed these challenges by proposing the use of social knowledge sources (KSs). These sources provide additional textual data on a growing number of topics, which can alleviate the sparsity of Microposts's content [7–12]. Furthermore these topic classifiers typically enhance the *lexical* (e.g. *bag-of-words* (*BoW*)) representation of text by incorporating additional contextual information about Microposts in the form of *semantic* (*bag-of-entities* (*BoE*)) features extracted from the content of Microposts only. Unlike these approaches, recently in [13] we proposed a TC framework which generates contextual information from graph structures surrounding concepts in multiple complementary linked KSs. Among the several useful graph structures defined in KSs [8], such as the *resource meta-graph* providing course grained classification of concepts by their type, or the *category meta-graph* which groups similar concepts together by their topic, our original framework exploited the *resource meta-graph* for context generation. Moreover, in [14] we also studied different content-based topic similarity (also called domain similarity or dataset similarity [15]) measures, which quantify the similarity between the KS data and Twitter data, serving as a proxy for the performance of a topic classifier on Twitter data. These content-based features correspond to simple *BoW* and *BoE* features derived from the Micropost content only.

Unfortunately, current approaches still present some limitations. The majority of the approaches model the entities using very generic concept types. For example, in the case of the entity *Obama*, the generic class *Person* is considered. When detecting Microposts related to the *war* topic, however, a more fine grained categorisation of this entity, such as *President of United States* (*Presidents_of_the_United_States*), could be more useful.

Further, the use of fine grained information in KSs provided by the *category meta-graph* has been exploited for many other problems, such as document classification [8], entity disambiguation [16], and semantic relatedness [17], and shown that it carries rich semantic information. However, to date no study has been conducted to investigate the usefulness of this *category meta-graph* graph structure for TC.

In this paper we thus present an extension of our TC framework [13], which exploits this new semantic graph, called *category meta-graph*, providing a more fine grained classification of concepts based on their topics. We introduce a set of novel semantic features derived from this graph, and present a comparative evaluation against those obtained from the *resource meta-graph*.

Furthermore our goal is also to understand which semantic feature contributes to the performance of a topic classifier. For this reason we propose an approach for automatic estimation of accuracy loss of a topic classifier. We introduce novel topic similarity measures, which in contrast to our previous content-based similarity measures [14], aim to measure the similarity between the KS documents and Microposts at a conceptual level, considering the enriched representation of these documents.

To evaluate the usefulness of exploiting this new *category meta-graph* for both TC and topic similarity, we present an extensive analyses of our extended framework using a ground truth data in the Emergency Response (ER) and Violence Detection (VD) domains.

The main research questions we investigate are the following:

- *How does the performance of a topic classifier vary using different concept graphs? Which concept graph provides the most useful semantic features for TC of Microposts?*
- *Are there differences in the roles (generalisation patterns) of the concept graphs in the different TC scenarios?*
- *Can we predict the performance of a topic classifier? Which topic similarity measure provides best estimate on the performance of a topic classifier?*

1.1. Contributions

To address the above research questions, we present an approach which facilitates the exploitation of multiple semantic meta-graphs from linked KSs for TC of Microposts. In particular, in contrast to our previous work [13], in this paper our main focus is to understand the differences between the different semantic concept graphs, and present a comparative evaluation of these graphs at different stages of our three-stage approach. The main stages of our approach can be summarised as follows: (i) *context modelling*; (ii) *topic classification* and (iii) *topic similarity analysis*.

The *context modelling* stage enriches the text using different concept abstraction techniques. For this reason we extract various semantic features about entities appearing in the text from two distinct concept graphs built from linked KSs.

The second stage *topic classification* involves the creation of statistical TC models, which incorporate various semantic features obtained in the context modelling step. In this stage we investigate two different scenarios: the Twitter only scenario in which we build a topic classifier on Twitter data only, and the cross-source TC case where we make use of the information from multiple linked KSs. This allows us to analyse which concept graph provides better semantic features for TC, and also whether the role of the semantic features differ according to the TC scenarios. In particular, we investigate whether the same semantic features which account for modelling the specificity of the topic in the Twitter only scenario, serve the same role in the cross-source scenarios.

The final stage *topic similarity analysis* uses the enhanced representation of the documents (in both the KSs and Twitter) following context modelling to provide an estimate on the performance of the topic classifier on new, unseen Micropost data. This allows us to analyse which semantic concept graph is better suited to measure the topic similarity between KS documents and Microposts for TC. In this stage, we also investigate whether this novel representation of the documents provides a better measure for topic similarity than our previous content-based statistical measures [14].

The main contributions of this paper are four fold:

- We introduce a novel set of semantic features derived from the *category meta-graph* of KSs;
- We present a systematic comparison of different semantic concept graphs for TC of Microposts;
- We present an analysis of the different roles of semantic concept graphs on ground truth data in the VD and ER domains;
- We propose a novel set of topic similarity measures for estimating the performance of a topic classifier.

[1] http://irevolution.net/2013/07/07/twitter-political-polarization-egypt/.