Information Sciences 328 (2016) 76–96

Contents lists available at ScienceDirect

Information Sciences

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

Categorizing events using spatio-temporal and user features from Flickr

Steven Van Canneyt[a,*], Steven Schockaert[b], Bart Dhoedt[a]

[a] Ghent University - iMinds, Gaston Crommenlaan 8, Ghent, Belgium
[b] Cardiff University, 5 The Parade, Cardiff, United Kingdom

ARTICLE INFO

Article history:
Received 14 January 2015
Revised 11 August 2015
Accepted 21 August 2015
Available online 28 August 2015

Keywords:
Ensemble learning
Semi-structured data
Events
Social media
Flickr

ABSTRACT

Even though the problem of event detection from social media has been well studied in recent years, few authors have looked at deriving structured representations for their detected events. We envision the use of social media for extracting large-scale structured event databases, which could in turn be used for answering complex (historical) queries. As a key stepping-stone towards this goal, we introduce a method for discovering the semantic type of extracted events, focusing in particular on how this type is influenced by the spatio-temporal grounding of the event, the profile of its attendees, and the semantic type of the venue and other entities which are associated with the event. We estimate the aforementioned characteristics from metadata associated with Flickr photos of the event and then use an ensemble learner to identify its most likely semantic type. Experimental results based on an event dataset from Upcoming.org and Last.fm show a marked improvement over bag-of-words based methods.

1. Introduction

Several authors have shown that social media can successfully be used to detect events [1,2,4,16,25,34], even before they have been reported in traditional media [28]. However, it is difficult to evaluate queries such as 'In which countries did U2 perform during 2013?' against a set of events that have been detected in this way. Answering such queries requires access to a structured representation of events. The absence of such structured representations limits the applicability of current methods for event extraction from social media. In particular, it is of interest to learn structured representations of the kind that have traditionally been considered in template-based information extraction [9,10]. The most relevant template for a given event is often based on the semantic type of that event. For instance, for a football match we want to encode the final score. In contrast, we want to know the magnitude and number of casualties of an earthquake. In this paper, we study how the semantic type of events can be extracted from social media, as a first step towards automatically extending and creating structured event databases.

Evidence about the semantic type of an event can be obtained by analyzing social media documents, such as Flickr photos taken at the event, which we consider in this paper, or tweets that have been sent about the event. In particular, similar as in e.g. [1,2,25], we represent an event as a set of social media documents related to that event, together with its associated characteristics. A set of social media documents related to an event may for instance be automatically extracted from social

* Corresponding author. Tel.: +32 (0)9 33 14940.
E-mail addresses: Steven.VanCanneyt@ugent.be, steven.vancanneyt@intec.ugent.be (S. Van Canneyt), SchockaertS1@cardiff.ac.uk (S. Schockaert), Bart.Dhoedt@ugent.be (B. Dhoedt).

http://dx.doi.org/10.1016/j.ins.2015.08.032
0020-0255/

media [1,2,4,25] or may be extracted from existing event databases such as Upcoming[1]. Most initial work about discovering the semantic types of events only used textual information [7,8,27], which may lead to poor performance when the text is noisy (e.g. in some Twitter posts) or absent (e.g. in some Flickr photos). However, social media documents also contain metadata which provide an indication about the spatio-temporal and attendees features of an event. The hypothesis we consider in this paper is that in many cases the event type can be discovered by looking at properties, such as timing, the type of venue or characteristics of attendees, which can be readily obtained from social media sources. For example, when an event occurs on a Saturday inside a sport complex and it has basketball players as main actors, it is very likely that this event is of type 'basketball game'.

Even though our methods can be applied more generally, we will restrict ourselves in this paper to experiments with Flickr photos. In particular, the considered characteristics of an event are estimated using its associated Flickr photos, and these characteristics are then used to describe the event. To estimate the type of a given event, we use an ensemble of classifiers, one for each of the considered descriptors. Subsequently, we consider two use cases. First, these trained classifiers are used to analyze in detail to what extent our methodology is able to discover the semantic type of known events that have no associated semantic type. This is useful, for example, to improve existing event databases such as Upcoming, for which we found that about 10% had no known type. Second, the model is used to estimate the semantic type of events which have been automatically detected from Flickr, which could substantially increase the applicability of existing methods for automated event detection.

The remainder of this paper is structured as follows. We start with a review of related work in Section 2. Next, in Section 3, we describe our methodology for classifying events based on their characteristics. Subsequently, Section 4 presents the experimental results. Finally, we discuss and conclude our work in Sections 5 and 6.

2. Related work

Early work on extracting structured data from text focused largely on news articles. The Message Understanding Conferences (MUC) were organized during the 1990s [10] to encourage the development of new and better methods to extract information from documents. The main task of these conferences was to automatically fill in a template with information about the event described in a given news article. For each event type considered, a template was constructed by the organizers with characteristics specific to it. For example, the template of the 'airplane crash' event type contained characteristics such as the place and the consequences of the event. The standard methodology to handle this task consisted of two major parts. First, the system extracted facts, i.e. entities and actions, from the text through local text analysis. Second, global text analysis was used to merge the discovered facts or to produce new facts through inference. The obtained knowledge was finally used to fill in the event templates. More details on this method are described in [9]. An interesting project related to event detection using news media is the GDELT project[2]. GDELT monitors the world's broadcast, print and web news in real-time. It identifies and connects people, locations, organizations, themes, emotions, quotes and news-oriented events which are stored in a structured event database. This information gives a global perspective on what is happening, its context, who is involved, and how the world is feeling about it. This data was for instance used to visualize the protests and unrest around the world on a map in real-time.

In the last few years, the focus has shifted somewhat from news articles to social media due to the latter's large data volume, the broad user base and its real-time aspect. However, social media documents tend to be noisy and are often very short compared to news articles, which has led to new challenges.

There has been a lot of interest in detecting events and their associated documents using social media. In [4], for example, the authors analyzed the temporal and locational distributions of Flickr tags to detect bursty tags in a given time window, employing a wavelet transform to suppress noise. Afterwards, the tags were clustered into events such that each cluster consists of tags with similar geographical distribution patterns and with mostly the same associated photos. Finally, photos corresponding to each detected event were extracted by considering their related tags, time and location. EDCoW [34] used wavelet transformations to measure the bursty energy of each word used in Twitter posts. It then filtered words with low energy in a given time window t. Finally, the remaining words were clustered using modularity-based graph partitioning to detect events in t. Twevent [16] improved the approach of EDCoW by first splitting the incoming tweets in n-grams. An n-gram was then considered as an event segment in a given time window when the occurrence of that n-gram was significantly higher than its expected occurrence. The obtained event segments were finally clustered into events and ranked based on the importance of their event segments in Wikipedia.

Becker et al. [1] represented an event as a cluster of social media documents related to that event. To detect events, they clustered social media documents based on their textual, time and location similarity features. They used a classifier with these similarity scores as features to predict whether a pair of documents belongs to the same cluster. To train the classifier, known clusters of social media documents were used, which were constructed manually and by using the Upcoming database. When the probability that a document belongs to an existing cluster is smaller than a threshold, a new cluster is generated for this document. Becker et al. [2] introduced an additional step which classifies the clusters corresponding to candidate events as 'event' or 'non-event' based on e.g., the burstiness of the most important words in the clusters and the coherence of the content of the social media documents in the cluster. Using the methodology described in [1,2], the authors were able to detect events using Flickr and Twitter data. Their methodology was evaluated in [1] by comparing the detected photo clusters and the photo

[1] http://upcoming.org/
[2] http://gdeltproject.org/