Data & Knowledge Engineering 88 (2013) 1–24

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

Event identification in web social media through named entity recognition and topic modeling

Konstantinos N. Vavliakis *, Andreas L. Symeonidis *, Pericles A. Mitkas

Aristotle University of Thessaloniki, Dept. of Electrical and Computer Engineering, GR54124 Thessaloniki, Greece
Information Technologies Institute, Centre for Research and Technology — Hellas, GR57001 Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 14 December 2011
Received in revised form 26 August 2013
Accepted 27 August 2013
Available online 4 September 2013

Keywords:
Event identification
Social media analysis
Topic maps
Peak detection
Topic clustering

ABSTRACT

The problem of identifying important online or real life events from large textual document streams that are freely available on the World Wide Web is increasingly gaining popularity, given the flourishing of the social web. An event triggers discussion and comments on the WWW, especially in the blogosphere and in microblogging services. Consequently, one should be able to identify the involved entities, topics, time, and location of events through the analysis of information publicly available on the web, create semantically rich representations of events, and then use this information to provide interesting results, or summarize news to users.

In this paper, we define the concept of *important event* and propose an efficient methodology for performing *event detection* from large time-stamped web document streams. The methodology successfully integrates named entity recognition, dynamic topic map discovery, topic clustering, and peak detection techniques. In addition, we propose an efficient algorithm for detecting all important events from a document stream. We perform extensive evaluation of the proposed methodology and algorithm on a dataset of 7 million blogposts, as well as through an international social event detection challenge. The results provide evidence that our approach: a) accurately detects important events, b) creates semantically rich representations of the detected events, c) can be adequately parameterized to correspond to different social perceptions of the event concept, and d) is suitable for online event detection on very large datasets. The expected complexity of the online facet of the proposed algorithm is linear with respect to the number of documents in the data stream.

1. Introduction

The World Wide Web has transcended from a read-only to a read-write web. Nowadays, anyone can actively participate in content creation by providing personal opinions, reporting an event, or commenting on a posted article. Information is uploaded in real-time; the minute an event that attracts people's attention occurs, its description becomes available on some (micro) blogging service. Thus, we should be able to identify online and real-life events (at least the important ones), just by looking at the web.

The definition of *event* depends on context and granularity. Dictionary.com defines an *event* as "something that occurs in a certain place during a particular interval of time". The term may also refer to a significant occurrence or happening, or a social gathering or activity [1]. WordNet defines an event as "something that happens at a given place and time" [2]. In knowledge representation, the event concept is an activity that involves an outcome or an arbitrary classification of a space/time region by a

* Corresponding authors at: Aristotle University of Thessaloniki, Dept. of Electrical and Computer Engineering, GR54124 Thessaloniki, Greece.
E-mail addresses: kvavliak@issel.ee.auth.gr (K.N. Vavliakis), asymeon@eng.auth.gr (A.L. Symeonidis).

0169-023X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.datak.2013.08.006

cognitive agent. In general, one could argue that an event is "a notable activity that happens". This definition is intentionally vague, since the event concept is socially defined, meaning that an event may be important to a group of people and unimportant to others.

In our work, events cover real life as well as web "happenings", and may comprise only one (e.g. a natural disaster), or several topics (e.g. the Olympic Games and all the games that take place during them). Thus, our methodology can detect any major event that triggers enough discussion in the web. In the case of real life events, we detect their virtual representations, which we consider to provide an accurate description of the real life event. Although the web representation will appear after the real life event, the two event times are expected to be very close, as users tend to immediately report events in the social media, especially in Twitter [3]. Thus, in our work we use the real life event time and the time of the web event representation interchangeably, and we consider as future work the quantitative analysis of the difference between them.

To our understanding, an event is a "significant" action that has a duration (beginning–peak–end), involves a set of entities (legal or physical), and is associated with one or more locations. An event can be described by a topic, which we define as the matter (subject) dealt in the text that is used to identify the event. We consider that an event takes place when a sudden peak of the mentioned topic/entity occurs in the web.

Having defined the notion of event within the context of our work, we discuss the notion of *event identification* (or *event detection*[1]). According to [4], event identification is the problem of identifying stories in several continuous news streams that amount to a new or previously unidentified event. Identification may imply discovering previously unidentified events in an accumulated collection ("retrospective identification"), or flagging new events from live news-feeds in an on-line fashion ("on-line identification"). Event identification comprises numerous challenges: one has to integrate information from multiple document streams, extract the spatial and temporal information associated with each document, identify and distinguish possible unknown entities, and classify the event to multiple event types.

The definition of *importance* is also subjective. Events that may be deemed interesting by many people are often available in various web sources. Although they are authored by different "reporters" that may use different vocabulary and express diverse opinions, they all share common features. Documents (articles, news stories, blogposts, comments, tweets, etc.) referring to the same event are reported at time periods close to the actual event. They also contain similar information (topics and named entities) that define the reporting event. We argue that these information snippets are the principal components that indicate events. We define an event as *important*, if the event has affected enough people to be reported or commented on in the Web. The minimum number of reports (per time unit) can be a tunable threshold and depends on the specific application/domain. In this work, we propose a methodology to detect previously unknown *important events*, as reported through social media interactions. We take advantage of publicly available information in the blogosphere and identify the *time* and *space* of events, as well as their semantics through a collective intelligence process. By *time*, we mean the time the event is reported on the web, which is after the event took place, but very close to the real event [3]. By *space*, we refer to the physical location(s) the event takes place and we try to identify, wherever available, the specific location entities involved. The event significance is calculated by the number of entity/topic occurrences over some time period. In addition, we propose an efficient unsupervised algorithm that detects the important events described in a dataset. The overall framework integrates a variety of unsupervised learning techniques, such as named entity recognition, dynamic topic map discovery, topic clustering, and peak detection, in order to identify events. Our methodology is suitable for fast detection of socially defined topics, both in an online, as well as in a retrospective manner. The contribution of our work can be summarized in the following: a) We propose a methodology that accurately detects interesting events and augments each event with semantic information pointing out the topic, the entities involved, the place, and the time period the event was observed on the web. b) We also propose an algorithm that can be adequately parameterized to accommodate different perceptions of the event concept and has expected complexity that grows linearly with the number of documents in the stream. This makes the algorithm suitable for online event detection on very large datasets. Our implementations of the proposed methodology and algorithm were evaluated on a dataset of 7 million blogposts and have outperformed other approaches in an international social event detection challenge.

The rest of the paper is organized as follows: related work on event and peak detection, as well as on topic extraction, is discussed in Section 2. Section 3 describes in detail all the facets of the proposed methodology, while Section 4 introduces the tunable algorithm for event detection. Section 5 evaluates the methodology through extensive experimental action and compares it against other approaches. Section 6 summarizes our work, discusses future directions, and concludes the paper.

2. Background and related work

The main objective of the event identification problem is to identify events from temporally-ordered streams of documents and organize these documents according to the events they describe. Towards this direction, numerous algorithms and techniques have been proposed, with most of them in the unsupervised learning category.

In the majority of cases, dynamic *clustering* techniques are used. One common approach is to model event identification as an online incremental clustering task [5]. For each document, its similarity to existing events (clusters of documents) is computed and the document is assigned to either an existing event, or to a new event based on predefined criteria. Following a different approach, Zhang et al. [6] propose a news event detection model that speeds up the detection task by using a dynamic news

[1] In this paper we use the terms event identification and event detection interchangeably.