Twitter Event Detection and Modeling with TEWS

Mykael Vigo, Zohra Bellahsene, Dino Ienco, Konstantin Todorov

Twitter Event Detection and Modeling with TEWS

Mykael Vigo[1], Zohra Bellahsene[1], Dino Ienco[2], Konstantin Todorov[1]

[1] LIRMM / University of Montpellier, France
{firstname.lastname@lirmm.fr}
[2] Irstea, UMR TETIS, Montpellier, France
dino.ienco@irstea.fr

Abstract. This paper introduces *TEWS*—**T**witter **E**vents on the **S**emantic **W**eb, pronounced like "news"—a semantic web tool for detection and representation of events taking as an input the social stream Twitter. The tool assists the user throughout a complete processing chain, starting from the detection of events on Twitter, their modeling and representation following the semantic web principles, to their storing in an RDF knowledge base that can be further published on the Web of Data.

Keywords: Linked Data, Event Extraction, Twitter, RDF Modeling

1 Introduction and Motivation

One of the goals of the Linked Open Data (LOD) initiative is to structure and interconnect data on the web by using semantic web technologies, such as the Resource Description Framework (RDF), thus taking the web of today up to the level of a veritable information network [1]. Although a considerable effort has been made in that direction throughout the last years, many sources of information on the web still remain unexplored, although they contain useful data beneficial for the LOD project. In this paper, we focus on the social medium Twitter—a platform for the publication of short messages (*tweets*) of maximal length of 140 characters. Twitter has became a major source of information about important events, made available in real time by media and ordinary users. The current paper describes *TEWS*, a tool that assists the user in the detection of events discussed on Twitter and their representation conforming to the semantic web principles in view of their publication on the web as RDF triples.

The originality of the approach lies in the use of tweets produced by conventional media (news agencies) in order to guide the event detection process on the tweets produced by ordinary users (the crowd). Contrary to previous works, in which the focus falls on event detection on Twitter [2] or the production of RDF data from text [3], our proposal covers both tasks.

2 Overview of TEWS

TEWS takes as an input a corpus of tweets. We have developed a protocol for the continuous constitution of a corpus by the help of the Twitter4J[3] API, which is designed to collect tweets from the social stream. As shown on the left hand side of Fig. 1 (steps 1–3), we are interested in two sources of information: (1) the conventional media, seen as a reliable and uninterrupted source of information about events (such as *ABC News*, *Al Jazeera*, *CNN*, *BBC*, etc.), and (2) the crowd (ordinary users). The latter is filtered by considering only tweets that contain a hashtag or a keyword (a location) from the media tweets corpus. The resulting set of tweets is continuously fed to *TEWS* in real time.

The overall events detection workflow and their RDF representation is given in Fig. 1. Steps 1–3 represent the corpus extraction procedure discussed above. *TEWS* is responsible for the actual extraction of events and their modeling and storing as RDF triples, given in steps 4–6. Step 7 illustrates the option that these data are further included to the web of data following the well-established publication protocol. As a design choice, the tool makes use of the MABED system [2] for detecting events and the WikiTimes [4] vocabulary for event modeling.

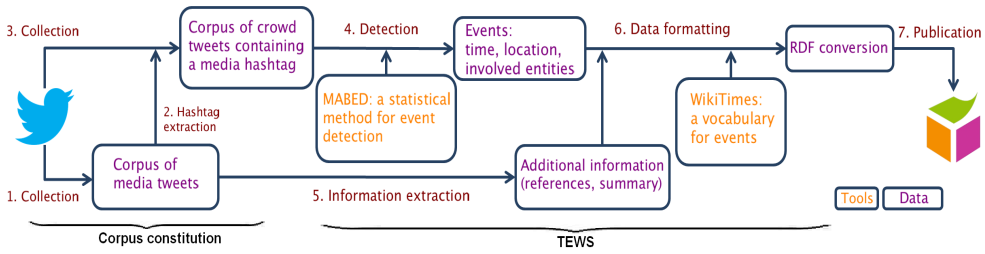


Fig. 1. The workflow of *TEWS* preceeded by the process of corpus constitution and succeeded by the data publishing step.

Event Detection. MABED is based on statistical methods for the identification of important events discussed by the users in a tweets corpus. In the first place, an event is defined by a salient topic identified by a sudden increase of the occurrence of certain key-words. An event is defined by a main key-word and a set of related terms. The efficiency of the approach is demonstrated in [2]. Note that the system is user-parametrizable with respect to time slices length and consideration of mentions (preceeded by the @ symbol). The length of the time slices has to be proportional to the length of the corpus constitution period.

Event Modeling. After comparing various existing ontologies, we have chosen the WikiTimes model for representing events, for it appeared to be the closest to our vision of an event. WikiTimes allows for the creation of two types of resources – an event and a story (a sequence of events). We focus on simple events only.

[3] `http://twitter4j.org/en/index.html`

Each event is anchored in the time and is characterized by several properties, described in a specific vocabulary, such as place and date of occurrence and involved entities, as well as additional information (Fig. 2).

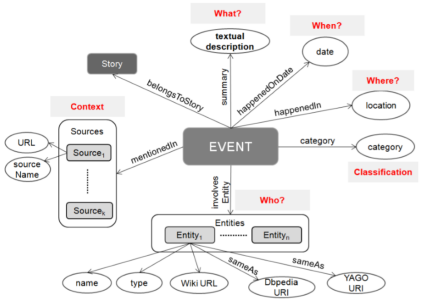


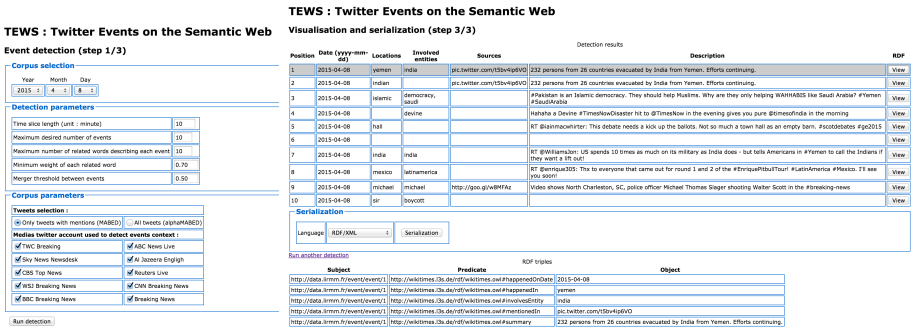
Fig. 2. The WikiTimes event resource model (taken from [4]).

3 TEWS in Use: an Example

We explain how *TEWS* works through a case study. A video illustrating this scenario is made available online[4]. The initial page of *TEWS* (Fig. 3(a)) allows to define all the parameters our tool needs. It has three main areas: i) *Corpus selection*: in this box we can specify the date (year, month, day) related to the tweets collection. The tool will retrieve only tweets published in the specified period. ii) *Detection Parameters*: this area allows to introduce the parameters related to the MABED system [2]. We can specify parameters like "maximum number of events", "maximum number of related words", etc. iii) *Corpus Parameters*: through this box we can specify the list of conventional media (news agencies), from which *TEWS* extracts hashtags or keywords to filter events from the general (crowd) stream of Twitter. In this case study, we choose to select all the conventional media sources. The interface allows to select a portion of them. Once parameters are chosen, the user can click on the *Run detection* button and *TEWS* will collect tweets and the events will be extracted.

Fig. 3(b) shows the results related to the specified period and the parameters introduced in the previous step. We can observe that *TEWS* detects ten events. Each event is described by the date, the location, the involved entities, the sources and the textual description. From the result list we can select one (or more) lines and visualize the corresponding RDF triples automatically generated by *TEWS*. In the example, we select the first event related to the *Yemen earthquake* and the tool produces the corresponding RDF knowledge. The RDF triples are visualized at the bottom of the page. The WikiTimes [4] vocabulary

[4] `https://www.dropbox.com/s/owc43372u47oe6k/tews_demo.mp4?dl=0`



(a) (b)

Fig. 3.

is employed to name the events properties. A standard procedure, based on the semantic web best practices, is deployed for creating resource URI's. At this step the user can serialize the produced knowledge in different formats (RDF/XML, N-Triplets, etc.) and the new set of RDF triplets can be successively published on the web of data or managed by standard triplestore engine.

4 Conclusion and Future Work

We introduced *TEWS*—a new Twitter event detection tool that produces RDF triples. The tool exploits conventional media sources to extract important keywords to trace phenomena on the stream of tweets. *TEWS* supplies a complete processing chain that combines state-of-the-art Twitter event detection with semantic web technologies to produce new knowledge to share on the web of data. As future work, we plan to improve the natural language processing aspects of the tool (such as named entity recognition and word sens disambiguation in order to filter out false entities, (i.e., exploit DBPedia). We also plan to integrate other sources of information to enrich the knowledge extraction from Twitter supplying a more complete description of the detected events.

References

1. C. Bizer, T. Heath, and T. Berners-Lee, "Linked data - the story so far," *IJSWIS*, vol. 5, no. 3, pp. 1–22, 2009.
2. A. Guille and C. Favre, "Mention-anomaly-based event detection and tracking in twitter," in *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pp. 375–382, IEEE, 2014.
3. R. Anantharangachar, S. Ramani, and S. Rajagopalan, "Ontology guided information extraction from unstructured text.," *IJVeST*, vol. 4, no. 1, p. 19, 2013.
4. G. Tran, M. Alrifai, T. N. Nguyen, and W. Nejdl, "Wikitimess knowledge extraction and enrichment process," 2014.