Neurocomputing 172 (2016) 168–179

Contents lists available at ScienceDirect

Neurocomputing

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

Online indexing and clustering of social media data for emergency management

Daniela Pohl [a,*], Abdelhamid Bouchachia [b], Hermann Hellwagner [a]

[a] Institute of Information Technology, Alpen-Adria-Universität Klagenfurt, Universitätsstr. 65-67, Klagenfurt, Austria
[b] Smart Technology Research Center, Bournemouth University, Poole House Talbot Campus, Fern Barrow Poole, BH12 5BB Bournemouth, UK

ARTICLE INFO

Article history:
Received 28 November 2013
Received in revised form
13 January 2015
Accepted 26 January 2015
Available online 9 May 2015

Keywords:
Event detection
Information filtering
Online indexing
Online clustering
Emergency management

ABSTRACT

Social media becomes a vital part in our daily communication practice, creating a huge amount of data and covering different real-world situations. Currently, there is a tendency in making use of social media during emergency management and response. Most of this effort is performed by a huge number of volunteers browsing through social media data and preparing maps that can be used by professional first responders. Automatic analysis approaches are needed to directly support the response teams in monitoring and also understanding the evolution of facts in social media during an emergency situation. In this paper, we investigate the problem of real-time sub-events identification in social media data (i.e., Twitter, Flickr and YouTube) during emergencies. A processing framework is presented serving to generate situational reports/summaries from social media data. This framework relies in particular on online indexing and online clustering of media data streams. Online indexing aims at tracking the relevant vocabulary to capture the evolution of sub-events over time. Online clustering, on the other hand, is used to detect and update the set of sub-events using the indices built during online indexing. To evaluate the framework, social media data related to Hurricane Sandy 2012 was collected and used in a series of experiments. In particular some online indexing methods have been tested against a proposed method to show their suitability. Moreover, the quality of online clustering has been studied using standard clustering indices. Overall the framework provides a great opportunity for supporting emergency responders as demonstrated in real-world emergency exercises.

1. Introduction

Access to information is fundamental during emergency management in order to deal efficiently with different sorts of incidents (e.g., traffic accidents, hurricanes, earthquakes, terror attacks). Collecting this information is not always an easy task, especially when relief units are not immediately on-site, e.g., due to the distance or street damages. Social media (e.g., Twitter) offers a new opportunity for supporting emergency management by enabling collection of data.

Studies [1,2] show the potential of social media in different emergency situations. People report on any kind of emergency situation they witness. Therefore, social media has become an important instrument to exchange information, thus providing additional perspectives on emergency situations [3].

However, intelligent analysis methods are needed to relieve emergency responders from a cumbersome manual browsing task through this data, which is potentially noisy. Methods should be able to summarize the ongoing situation and provide an overview of the emergency situation at hand. In this paper, we focus on the detection of sub-events, i.e., specific crisis-related hotspots (e.g., flooding in a specific district of a city, power outage in another district) that emergency personnel should be aware of when organizing their intervention.

In our early work [4], we examined clustering algorithms for their suitability to detect sub-events from social media. We used Flickr and YouTube data for aftermath analysis of the crisis situation. In particular our investigations relied on offline clustering which is inappropriate for real-time analysis during the emergency situation.

We introduced an online sub-event detection mechanism [5] which combines real-time clustering and online indexing (i.e., weighting and selection of indexing terms). The sub-events (clusters) are detected and tracked as new items from social media users become available. In [5], the mechanism is used to analyze data from the Hurricane Sandy 2012 in the form of batches. It handles data

* Corresponding author. Tel.: +43 463 2700 3688; fax: +43 463 2700 993688.
E-mail addresses: daniela@itec.uni-klu.ac.at (D. Pohl), abouchachia@bournemouth.ac.uk (A. Bouchachia), hellwagn@itec.uni-klu.ac.at (H. Hellwagner).

http://dx.doi.org/10.1016/j.neucom.2015.01.084
0925-2312/

collections from Twitter, Flickr and YouTube. We extract terms as features from the *textual* metadata of the incoming items. We do not process videos from YouTube and images from Flickr and do not analyze their contents, we rather extract their textual metadata (title, description, and tags) to be used along with tweets. Initial experiments on this data show the suitability for detecting topics related to the crisis at hand.

We integrated our online detection mechanism in a media exploration framework. For evaluation of the online processing method, we implement similar indexing methods and compare them with our indexing approach. Hence, the focus of this paper is on the examination of the indexing methods. In doing so, we adapted the online clustering algorithm described in [6] to meet the context of our present application. The experimental setting and the results regarding the different methods are described. They emphasize the suitability of our idea of online indexing for processing social media data.

The structure of the paper is as follows. Section 2 discusses the related work. Section 3 addresses the terminology, i.e., difference between events and sub-events, and highlights it in the context of topic detection and tracking. Section 4 introduces our suggested "Multimedia Exploration Framework". Section 5 outlines the online sub-event detection, especially the interrelationship between online indexing and clustering algorithms. Section 6 describes the details of the online indexing (i.e., implementations and our learning and forgetting model). Section 7 depicts the used online clustering algorithm in this context. In Section 8, the experimental setting and the results are presented. Section 9 concludes the paper.

2. Related work

The present work is related to "topic detection and tracking" in the area of social media and to "indexing and feature selection" methods.

2.1. Topic detection and tracking

In fact, Twitter is very popular in social media analysis and detection. For example, Gao et al. [7] present an approach that colors geographical regions (social pictures) based on their importance for the topic of interests given by the messages related to these areas. The aggregation and coloring are based on a predefined algebra. The algebra also allows the combination of different social pictures (i.e., with multimedia processing like convolution or segmentation).

Lampos and Cristianini [8] identify important keywords from auxiliary sources, e.g., Wikipedia. These keywords are searched in tweets and scored according to the amount of keywords in the tweet (e.g., to identify the daily flu-rate based on incoming tweets) [9].

Krstajic et al. [10] show an event detection mechanism based on different scores that are calculated and combined by the preferences of the user. First, terms are extracted and combined to episodes (i.e., sets of tweets). After a predefined number of tweets shown to the system, the scores are calculated. If the combined score reaches a threshold, the episode is shown to the user as a new event.

Chakrabarti et al. [11] describe a detection mechanism based on initially learned terms and their importance for a specific event (e.g., football game). In contrast, Shen et al. [12] base their detection on general concepts (e.g., name of companies or persons). General concepts are aggregated together based on their contextual and lexical similarity. Tweets depending on the resulting bag-of-words clusters are analyzed via spike detection and shown to the user. Marcus et al. [13] summarize or identify events based on the peak detection mechanism (covering a one-minute time window). Klein et al. [14] analyze tweets in real time for emergency management. They introduce a graph analysis approach. It allows them to identify leading writing users as the origin of the information spreading. Cataldi et al. [15] describe in their work also a topic detection mechanism for Twitter considering the relation between users, i.e., followers. However, in emergencies people can write about the same events although there is no relation given between them.

Allan et al. [16] describe an approach for detecting and tracking specific events. Nallapati et al. [17] use agglomerative clustering to identify events in a static manner. Osborne et al. [18] describe an online story detection mechanism based on Twitter which uses Wikipedia to verify the identified stories. The framework described by O'Connor et al. [19] analyzes previously fetched tweets to identify and summarize topics. Starbird [20] introduces Tweak-the-Tweet, which defines and uses a predefined grammar for tweets to analyze them accordingly. Twitcident, by Abel et al. [21], is based on predefined keywords or manually inserted rules. CrisisTracker by Rogstadius et al. [22] (based on [13]) represents a crowdsourcing tool to support volunteers in processing of messages coming from the public during a crisis. It uses an initial *term frequency-inverse document frequency* (tf-idf) model based on a sample set of tweets.

Most of the approaches use additional or auxiliary material for detection, e.g., Wikipedia, previously processed training sets, or are based on a static analysis. Most of them (e.g., Wikipedia entries or a training set) are often not available during emergencies, especially in fast evolving scenarios.

2.2. Topic detection and tracking based on visual items

In addition to microblogs and text messages, visual items are important in the context of crisis management. Visual items (e.g., pictures and videos) give additional insights into the incident. For example, Chen and Roy [23] perform event detection based on tags annotating Flickr images. The approach allows them to identify periodic and non-periodic events. The tags are examined based on their temporal and spatial distribution and aggregated if they are similar (i.e., representing the same event). The approach allows them also to uncover the time and location of an event. In [24] an approach is proposed to identify disaster events from Flickr. It identifies bursty tags in a predefined time interval and fetches images related to a predefined number of tags. Rattenbury et al. [25] make also use of tags to identify events from Flickr. The identification process is based on a clustering algorithm that takes into account the distribution of tags over time. It is based on specific intra and inter-cluster relationship metrics to identify event-related clusters.

Another approach from Liu et al. [26] identifies events in Flickr images based on the number of items per day coming from unique users. If the number of the incoming items is above the median, a new event is declared. Petkos et al. [27] identify events from Flickr images/items based on Support Vector Machines. Support Vector Machines are proposed to decide/classify if two items belong to the same event. A graph representation is created, where nodes represent items and edges indicate if two items belong to the same event based on the decision of the Support Vector Machines. Community detection algorithms are applied to assign items to events. Rabbath et al. [28] investigate event detection from Facebook by locating photos of the same event shared by friends.

These studies use tags associated with images and were sometimes combined with visual features extracted from the images/videos (e.g., [28,27]) to detect events from social media. In a step forward, we rather use microblog texts in addition to textual annotations of the images and videos.