Noname manuscript No.
(will be inserted by the editor)

Spatio-temporal mining of keywords for social media cross-social crawling of emergency events

Andrea Autelitano · Barbara Pernici · Gabriele Scalia

Received: date / Accepted: date

Abstract Being able to automatically extract as much relevant posts as possible from social media in a timely manner is key in many activities, for example to provide useful information in order to rapidly create crisis maps during emergency events. While most social media support keyword-based searches, the amount and the accuracy of retrieved posts depend largely on the keywords employed. The goal of the proposed methodology is to dynamically extract relevant keywords for searching social media during an emergency event, following the event's evolution. Starting from a set of keywords designed for the type of event being considered (floods and earthquakes, in particular), the set of keywords is automatically adjusted taking into account the spatio-temporal features of the monitored event. The goal is to retrieve posts following the event's evolution and to benefit from cross-social crawling in order to exploit the specific characteristics of a social media over others. In the case considered in this paper, we exploit the precision of the geolocation of images posted in Flickr to extract keywords to search YouTube posts for the same event, since YouTube does not allow spatial crawling yet provides a richer source of information. The methodology was evaluated on three recent major emergency events, demonstrating a large increase in the number of retrieved posts compared with the use of generic seed keywords. This is a relevant improvement of relevance for providing information on emergency events, and the ability to follow the events development.

Keywords media mining · keyword extraction · adaptive crawling · emergency management · social media

A. Autelitano
Politecnico di Milano
E-mail: andrea.autelitano@mail.polimi.it

B. Pernici
Politecnico di Milano
E-mail: barbara.pernici@polimi.it

G. Scalia
Politecnico di Milano
E-mail: gabriele.scalia@polimi.it

1 Introduction

Extracting content from social media is becoming a success factor in many different domains, given the valuable and timely hints that these channels can provide.

One domain which has received great attention is emergency management. Indeed information extracted from social media has proven to be very useful and informative in many crisis situations [5].

One key activity in this domain is *rapid crisis mapping*, which has the goal of providing rescue teams and operators with information on the current situation of the area affected by the emergency. Rapid mapping professionals can find it beneficial to be supported by on-site visual information. Images/videos taken by users and uploaded on their personal social media accounts represent a new valuable resource [11]. The present work is developed within the E2mC[1] (Evolution of Emergency Copernicus Services) H2020 European project, which has the goal of exploring how to improve rapid mapping with the help of social media.

The goals are to retrieve as many images as possible (recall) and relevant to event, with images useful to assess the emergency situation which are geolocated or easy to geolocate automatically (precision).

In order to gather these data, the typical procedure involves crawling social media as the event evolves, by using certain seed keywords concerning the event type. However, this approach has some important limitations: generic keywords may fail to extract a sufficient number of posts, and the extracted posts may not be very precise [22]. In fact, the nature of social media is noisy and dynamic, characterized by ambiguities, fakes and trending keywords which arise and evolve, often without a central coordination in describing events and situations. In addition, images from the posts are useful only if they are precisely located. However, the number of georeferenced posts is very low, usually less than 3%, and the posts containing images are only a fraction of those [7]: from our previous studies, the expected percentage of Twitter[2] posts containing images which are useful for rapid mapping activities for an emergency event is around 0.1%, when social media are crawled with generic event-related keywords [6]. Therefore, there is a need to increase the number of posts that can provide useful visual insights about the event and which are geolocated or can be easily located with a geolocation algorithm on the basis of their textual contents (for a survey on geolocation algorithms see [13]).

A further consideration is that suitable keywords for an event may evolve in time. The reason is twofold and may depend on the type of event: some events (e.g. storms, or floods) move, therefore the situation evolves in different places. In other cases, awareness about the event changes over time, as mostly affected areas are identified (e.g., in earthquakes, for which the boundaries of the event are discovered in the first hours after the event). In this paper, we study how to provide faster awareness information in emergency events, as this analysis needs to be performed in a short time immediately after the event to support emergency responders and provide information on the situation to the public.

Social media present different characteristics with regard to volumes of data and the precision of geolocation information. From among the ones that can be

[1] `https://www.e2mc-project.eu`

[2] `https://twitter.com`

crawled without significant limitations, Flickr[3] is characterized by the association of precise location information to images in most cases, while YouTube[4] is characterized by large volumes of available videos, provided both by individual citizens and by the press. However no location information is associated to them.

The goal of this paper is to present an approach for extracting more (and more relevant) images and videos from social media and, in particular, from YouTube, during an emergency event, dynamically mining event-related keywords mainly from Flickr, in order to follow the evolution and the specificity of the event. The proposed methodology is based on the spatio-temporal characterization of the target event and on an iterative refinement of extracted keywords, followed by a cross-social crawling.

The proposed methodology is designed for ongoing events, continuously refining keywords as they evolve.

Section 2 discusses the state-of-the-art concerning the use of social media in emergency situations. Section 3 introduces the methodology to incrementally extract media leveraging on dynamic and unsupervised keyword generation, while Section 4 details the keyword generation and management approach. The experimental evaluation of the methodology in case studies related to emergency events is discussed in Section 5.

2 Related work

The use of social media in emergency situations has been advocated by many authors, as reported in the recent survey [5]. This section focuses on analyzing the state-of-the-art with respect to the requirements emerging from the use of social media in the rapid production of crisis maps, and positions the work.

A problem that srises in this context is *geolocation* when native geographical coordinates (also called georeferences or geotags or explicit geographical information) are not present in the post (meta)data. Geolocation has been studied from different points of view and through different techniques [1]. In particular, this work focuses on precisely locating media contents, rather than other user information, starting from available features (in particular, text), when the georeferenced posts available are not enough. In doing so, we leveraged our geocoder called CIME (Context-based IMage Extraction) developed in the E2mC project [8,19]. The CIME geolocation algorithm aims to geolocate posts starting from text and contextual information and is based on Stanford Core NLP[12] and OpenStreetMap (OSM)[5]. Using OSM with respect to other commonly used gazetteers, such as GeoNames[6], has the advantage to provide an increased granularity of available locations.

[18] combines unsupervised machine learning techniques and spatio-temporal analysis for damage assessment in emergency events, obtaining relevant topics for the affected areas. The keywords obtained in this work can be considered topics to some extent. However, the goal in the present work is not to assess damage but

[3] `https://www.flickr.com`

[4] `https://www.youtube.com`

[5] `https://www.openstreetmap.org`

[6] `https://www.geonames.org`

to enhance crawling, as a better assessment is an indirect consequence of more (and more relevant) media crawled. Topic extraction also involving spatial and temporal dimensions is discussed in [17], where LDA was used as a basis. In the present work the goal is to extract single keywords to be used as search words rather than topics, and the goal is to improve recall in search rather than location prediction, generating the search keywords set tailored to the event.

Clustering is a key step in this methodology. Indeed, clustering in this domain is typical of event detection techniques [3] and in estimating the affected areas [2]. In particular, density-based clustering (as DBSCAN) has proved to be effective for such a goal [21]. In our work we use density-based clustering for the same purposes (to estimate the affected areas), but it is a mean for a different final goal, which is to identify groups of posts that present common tags for keywords extraction and revision.

In [23], dynamic keyword generation is proposed for event-related tweets, starting from seed keywords, with a semi-supervised approach based on a classification of the relevance of tweets. While some techniques described in this work were leveraged, our methodology is totally unsupervised and focuses on tags in Flickr and YouTube rather than words in Twitter.

Several integrated frameworks are available to analyze crisis events through information extracted from social media, such as SensePlace3 [15], which focuses on integrating both natively georeferenced posts and implicit geographical information derived from tweets in emergency situations by means of natural language processing and geocoding. This paper is focused only on improving the crawling phase, although it goes in the same direction combining georeferenced and text-based geolocated posts. The outcome of the proposed methodology could feed an interactive crisis monitoring application.

This work can be considered a keyword-based *adaptive crawling*, as, for example, [22]. The difference is that our keyword generation is the result of a spatio-temporal modeling and analysis rather than an analysis of the post stream. In addition, the goal is to leverage the characteristics of one social media (in the considered case mainly Flickr) to improve the keywords set for crawling another social media (YouTube).

This work starts from the analysis and considerations contained in [7], which proposes a model for managing the evolving spatio-temporal information available in social media.

With regard to social media, most papers focus on Twitter [5], either as a primary or secondary source [14]. On the other hand, other social media can provide useful information for rapid mapping.

[14] proposes a multi-social triangulation approach starting from Twitter to extract keywords to crawl Flickr, which provides information that is not available in Twitter. The concept of triangulation was used in the present work by exploiting Flickr posts, which come with a precise geolocation, in order to mine keywords that are then used on YouTube. The latter can provide a large number of relevant videos.

The identification of subevents from social media using the clustering of posts from Flickr and YouTube is advocated in [16], with the identification of subevents based on term similarity and post coordinates.

Social media present different characteristics for the purposes of this work, as summarized in Table 1: feasible searches (by keywords or by location), the presence

	Facebook Instagram	YouTube	Twitter	Flickr
Feasible searches	Not feasible[a]	Keyword-based	Keyword/Location-based (also in streaming)	Keyword/Location-based
Localization of posts	GPS (of post)/Manual tag	No coordinates	GPS (of post)/Manual tag	GPS (of media)
Time information	Post publication	Post publication Recording time[b]	Post publication	Post publication Media shooting

[a] No useful searches for the purpose and the scale of the domain.
[b] This field is not detailed (day granularity) and has proven to be not reliable from a preliminary analysis, therefore it was not used.

Table 1: Summary of social media characteristics (taking into account rapid mapping requirements) updated to May 2018

of native geolocation of posts, and time information. As shown, georeferences are not always available and they usually refer to posts rather than associated media (for example, in Twitter, image metadata are not provided even if present in the original photo). The time is usually the post publication time, which can be delayed with respect to when the associated image was taken. Flickr is the social media with the most detailed media information, since it includes shooting time and location, and is also known for accurate location data [10]. Therefore it is a good candidate to identify reliable areas of interest for an emergency event. YouTube does not provide such metadata; however, since both Flickr and YouTube use tags in their posts, this paper proposes a triangulation approach to mine keywords (tags) from Flickr posts and use them in YouTube. Other social media could be employed instead of Flickr, e.g. Twitter, but it would be necessary to introduce cleaning and filtering techniques to ensure the spatial and temporal quality of the considered posts. Social media like Facebook[7] and Instagram[8] can not be used because of existing crawling limitations for the purpose and the scale of the domain.

3 Media extraction methodology

The general methodology is presented in this section. In particular, Subsection 3.1 describes the methodology's main features, while Subsection 3.2 lists all the methodological steps.

3.1 Overview of the approach

The goal of the proposed methodology is to extract — in a timely manner — as many posts (and, in turn, *images* and *videos*) from social media as possible, concerning an ongoing emergency event.

Since the majority of posts can be crawled through keyword-based queries, a key to extract more (relevant) posts is, ultimately, to find those *emerging keywords* related to the event and discarding the irrelevant or old ones.

[7] `https://www.facebook.com`
[8] `https://www.instagram.com`

Three directions are explored to accomplish such goal:

Temporal. Given a target time frame, the amount of related posts (even if posted subsequently) increases over time. More importantly, analyzing the past can give hints on the best keywords to crawl an ongoing event, in terms of emerging topics in the target area.
Spatial. A target area of an event will be characterized by both specific topics and unrelated topics. By characterizing that area with respect to the rest of the world, it is possible to filter out unrelated keywords and keep only the event-related ones.
Cross-social. As mentioned, each social media has different features. However, by targeting the same event, it is possible to exploit information that exists only in a certain social media to obtain keywords exploitable on other social media, following a "triangulation" approach [14].

This study targeted Flickr and YouTube. As discussed in Section 2, they have different features and strengths. Flickr provides precisely georeferenced images with accurate timestamps. Therefore, it can be used as a primary source for the spatio-temporal approach in mining keywords that can be used subsequently for further crawling on both Flickr and YouTube to extract videos, which are not natively characterized by accurate timestamps nor by georeferences.

Another characteristic of this approach is its intrinsic language independence. As the focus is on a spatio-temporal analysis, the methodology aims to mine keywords from posts based solely on spatial and temporal characteristics, following the event's evolution. Language-dependence is limited to some filtering steps, which are optional, and to the text-based geolocation of YouTube posts (which is, however, external to the methodology and handled by a geolocation module).

A *sliding windows* approach to crawling was adopted to follow the event's *evolution*. This is relevant mainly for Flickr which distinguishes between the *date taken* (DT) and the *date uploaded* (DU) of a media. When targeting a certain interval of time, which is a *time frame*, more media can be obtained as time passes. In fact, posts with DT $\in$ *time frame* can be included even if DU follows the end of the time frame, since in many cases in Flickr there is a delay in posting pictures, causing an image taken one day to be posted on the same day or later. Therefore, the entire keyword generation procedure for a time frame can be repeated, enlarging the sliding DU and adding new media.

This is illustrated in Figure 1. An $x.y$ crawling means that the x-th time frame was crawled taking into account posts uploaded until the y-th. The first 24-hours time frame (from 10/02/2014 to 11/02/2014) was crawled three times: 1.1, 1.2 and 1.3, retrieving each time media with the corresponding DT and DU.

Since the distinction between DT and DU does not exist on YouTube, it was crawled only for each time frame.

For the current work, we considered time frames of 24 hrs and uploads delayed up to 72 hrs. These parameters can be varied, provided that the number of posts to be analyzed is sufficient to identify areas of interest (see Subsection 4.1).

3.2 Methodological steps

The algorithm starts with a set of *seed keywords*, which are general, event-type related keywords. These should provide a "base set" of posts. Such set is the

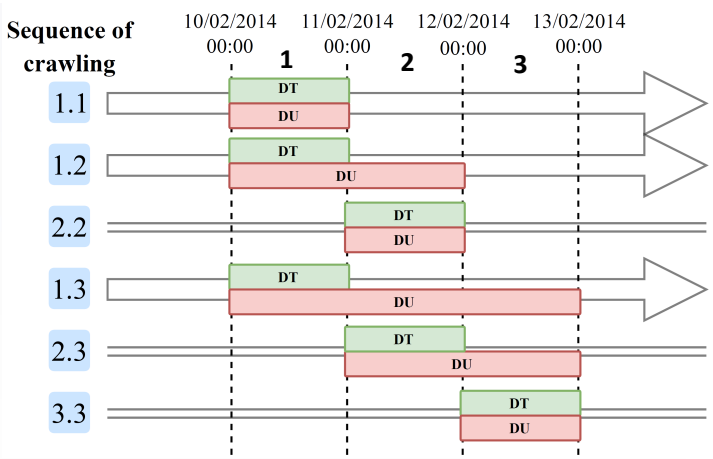

Fig. 1: Crawling sequence example

starting point to extract new keywords which lead to new posts, iteratively in a cycle. This procedure is repeated for each time frame, keeping a trace of the most relevant keywords from one time frame to the next, and taking into account new media which are published after the time frame as mentioned previously.

The methodology's main steps, depicted in Figure 2, are summarized as follows:

0. *Initial input.* Seed keywords (concerning the event type) and a past date/time — at least 24 hours prior to the beginning of the crawling — have to be provided and constitute the initial input. The input date/time corresponds to the beginning of the first time frame.
1. *Sliding window crawling.* For each 24-hours time frame, the crawling engine starts to create the *sliding windows*, and, as mentioned, also accounting for media uploaded after they were taken. Flickr and YouTube are incrementally crawled by keywords for the selected sliding windows.
   Flickr media are also enriched with additional georeferenced media extracted from the same albums and relevant groups[9].
2. *Spatio-temporal clustering.* Spatio-temporal density-based clustering is performed on all of the georeferenced Flickr media extracted up to that moment. In this way, the most event-affected areas can be detected. YouTube media geolocation may also be used (optional) through a text geolocation algorithm.
3. *Keyword generation.* Three parallel steps are performed for each area identified (cluster):
   – *Event-related keywords* generation. These are the emerging keywords with respect to the past for the target area, therefore characterizing the emerging event. This requires crawling for georeferenced media posted in the target area in the previous months. This step also enables the creation of "stories", which are "before/after" comparisons of interesting POIs.

[9]Relevant here means that the title contains at least a seed keyword. Groups and albums are retrieved only for media obtained through seed keywords and are limited to media posted on the same day.

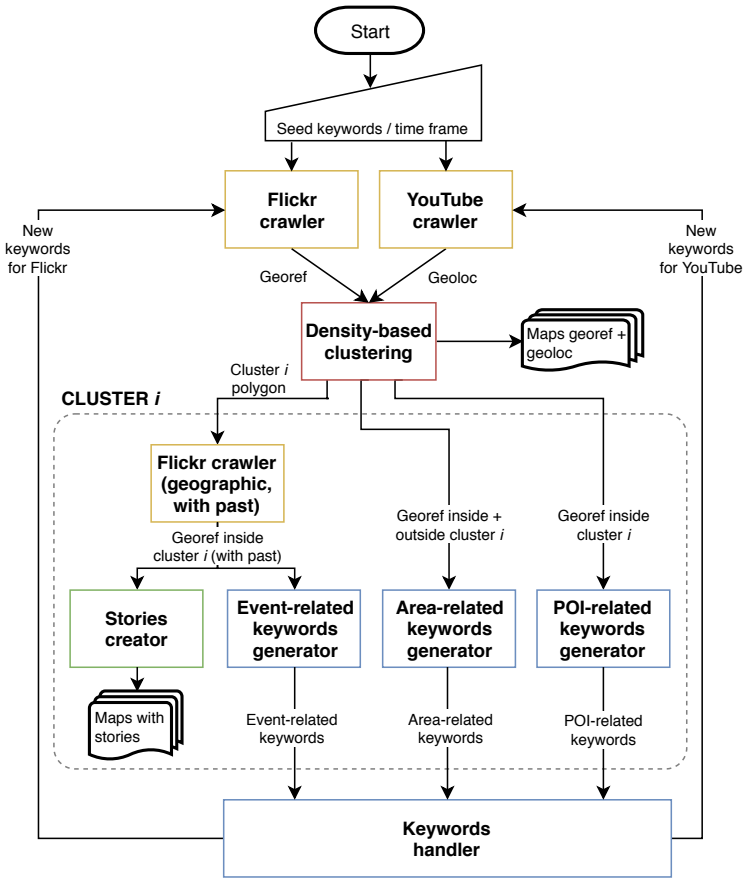

Fig. 2: General schema of the methodology

- *Area-related keywords* generation. Keywords characterizing a target area (cluster) with respect to the rest of the world, for the same time frame.
- *POI-related keywords* generation. Extracting relevant POIs for the area.

This step is detailed in Section 4, which also discusses keyword management for the removal of old keywords.

4. *Iteration.* Once the three sets of keywords have been generated, they are added to the seed keywords in order to re-crawl the same sliding windows and increase the (potentially relevant) results, restarting from step 1, iteratively.

4 Incremental keywords generation

This section details the keyword generation step, which is the core of the proposed methodology. It is characterized by two main steps: clustering (Subsection 4.1) and keyword generation (Subsection 4.2).

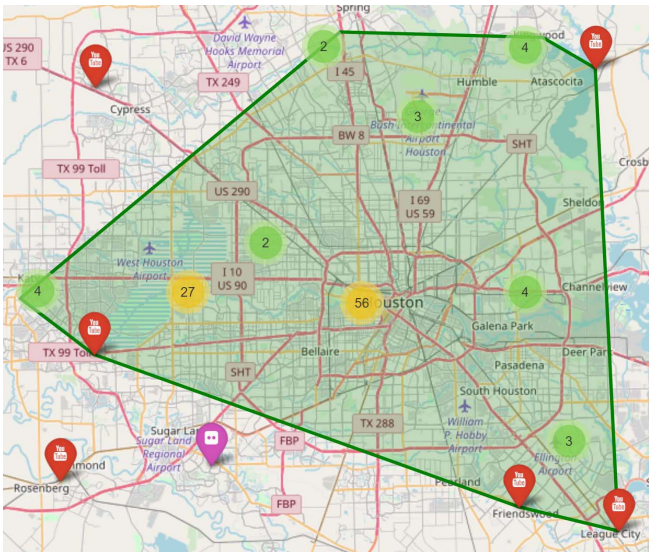

Fig. 3: Cluster example: the area of a cluster is given by the polygon which includes all the posts in the cluster (convex hull)

4.1 Spatio-temporal clustering

Spatio-temporal clustering identifies sub-areas related to the event. Points correspond to posts and each output cluster is a target sub-area.

Density-based clustering (DBSCAN [20]) was chosen for the task. Indeed, a higher density of posts related to domain-specific keywords is typically related to emerging events [18]. In addition, it accounts for affected areas which vary in number and shape.

The clustering is spatial because posts are interpreted spatially as geographical points (latitude and longitude) on the map, using the haversine distance[10]. At the same time, it is temporal since posts are clustered at each timeframe independently, following the event's evolution.

As mentioned above, clustering is performed preferentially on georeferenced Flickr media. This should provide an approximate but faithful representation of the event-affected areas in the real world. In fact, as mentioned in Section 2, georeferences are the most accurate types of locations that can be extracted by social media, and YouTube does not provide them. In this phase it is better to avoid forming non-significant clusters (i.e. clusters do not reflect real world affected areas), since the following keywords generation phase has a higher performance if based on a faithful localization of the real world event situation. However, geolocated YouTube media are introduced when Flickr is not able to generate enough clusters (3 in the current implementation — more complex and flexible models could be employed). The geolocalization in this case is performed using the CIME approach [19].

[10] `https://en.wikipedia.org/wiki/Haversine_formula`

DBSCAN has two hyper-parameters, ϵ and *minPts*, which have an impact on the methodology. Specifically, a cluster is formed only when there are at least *minPts* points which are at a distance ϵ one from each other: these points form the cluster's *core*. Once a cluster is formed, every other point within a distance ϵ to a point in the core is also included in the cluster.

In this domain *minPts* must not be too low, since some fakes/not relevant media exist, and it is necessary to consider such noise. ϵ must account for the sparsity of posts: ideally, relevant areas should be constituted by close points, but factors such as the heterogeneity of the territory and of the users, imprecisions in the locations associated to the posts and specificities of the emergency events make this situation practically unachievable. Therefore setting ϵ is a trade-off between not being too low, because many areas would not be found due to missing data, and not too high, because the goal is to find affected areas within the event as precisely as possible.

An example of cluster is shown in Figure 3. The figure shows the polygon which includes all the posts in the cluster (convex hull). The cumulative number of a set of closely-located posts is indicated for legibility.

In the following, we use the clusters to define event-affected areas for the keyword generation phase, where an area is delimited by the polygon defined from the cluster.

4.2 Keywords generation

Three categories of keywords are generated starting from each *event-affected area* (cluster) identified:

- *Event-related keywords:* keywords generated comparing Flickr posts in the event-affected area with posts in the same area in the previous period. These keywords are likely to refer to the event, because they are the keywords emerging in the present with respect to the past.
- *Area-related keywords:* keywords generated comparing posts in the event-affected area with respect to all the other posts (for the same timeframe). These keywords should characterize an area of the emergency event, and therefore they often include geographical keywords related to the most affected places (e.g., villages, neighborhoods, streets).
- *POI-related keywords:* They are the POI names extracted from the gazetteer (OSM) [9] in the event-affected area for the considered timeframe.

In each timeframe several iterations are performed to refine the clusters identified, and in each new iteration the previously generated keywords are used besides the new ones. The first iteration is based on the seed keywords; further iterations are based on the existing keywords plus the generated keywords. Each new set of keywords enables to crawl new media, therefore helping to refine the clusters, which in turn could allow to compute new keywords in a cycle. Theoretically, the iterative generation of keywords can be carried out until a fixed point is reached. However, for practical reasons, a limit has to be set[11].

[11] In this paper, we consider in the experimentation timeframes of 24 hours, and iterations are repeated 3 times in each timeframe.

The following paragraph provides details about the generation of each category of keywords.

4.2.1 Event-related keywords generation

These keywords are extracted from the tags of the georeferenced posts, but also titles and descriptions are employed for their refinement.

All the tags contained in the posts in the current timeframe are candidate keywords. A temporary quality score is attributed to each tag, evaluating its relevance and coverage with respect to tags contained in past posts for the same area. Then, those scores are refined considering tokens (words) inside titles and descriptions, thus obtaining a final score for each tag and consequently their ranking. Top tags in the ranking are selected, filtered, and constitute the event-related keywords.

The steps are detailed as follows:

1. *Temporary quality scores.* For each tag extracted from the posts in the area, a *relevance* and a *coverage* is computed. A *temporary quality score* is then obtained starting from these two, creating a temporary tag ranking.
   In particular:
   - *Relevance* for each tag is defined as its relative entropy, following the approach presented in [23]:

$$\begin{aligned} r(t) &= p(t,C) \cdot \log \frac{p(t,C)}{p(t,P)} \\ &= \frac{|C(t)|}{|C|} \cdot \log \frac{|C(t)| \cdot |P|}{|C| \cdot |P(t)|} \end{aligned} \tag{1}$$

     where:

t	is a tag in the cluster
C	is the set of posts in the cluster, in the current frame
P	is the set of posts in the cluster, in the past[12]
$C(t)$	is the set of posts in the cluster, in the current frame, with tag t
$P(t)$	is the set of posts in the cluster, in the past, with tag t
$p(t,C) = \frac{\|C(t)\|}{\|C\|}$	is the probability that tag t is present in C
$p(t,P) = \frac{\|P(t)\|}{\|P\|}$	is the probability that tag t is present in P

   - *Coverage* for each tag is defined as [23]:

$$c(t) = \begin{cases} e^{(x-\gamma)} \text{ if } x = \frac{|C(t)|+|P(t)|}{|C(t)|} \leq \gamma \\ e^{(\gamma-x)} \text{ if } x = \frac{|C(t)|+|P(t)|}{|C(t)|} > \gamma \end{cases} \tag{2}$$

     where:

$x = \frac{\|C(t)\|+\|P(t)\|}{\|C(t)\|}$	is the *coverage ratio*
γ	is a hyper-parameter to be empirically set.[13]

[12] In this context, the past is the period of time before the beginning of the event, which models the normal situation. Three months are considered in the current implementation.

[13] The higher it is, the less a higher coverage is penalized, and therefore tags frequent also in the past are less penalized. γ is set as 1 in the current implementation.

– The *temporary quality score* for each tag is defined starting from Equations (1) and (2) as:

$$tagQS^*(t) = r(t) \cdot c(t) \tag{3}$$

In the end, we obtain a *temporary ranking of tags* based on the temporary quality scores.

2. *Title&description scores.* In principle, the previous step could be enough. However, since data are typically very scarce, top scores in the temporary ranking will often present ties. To yield a more significant ranking, titles and descriptions are employed and tokenized. For each token w, the quality score is produced by using the aforementioned method for tags.

$$tokenQS(w) = r(w) \cdot c(w) \tag{4}$$

3. *Tag quality scores.* The temporary quality score of each tag obtained under Step 1 is refined by the token score of the same candidate keyword (if any) found in titles and descriptions, thus obtaining a *final quality score* for each tag:

$$tagQS(t) = tagQS^*(t) \cdot (1 + tokenQS(t)) \tag{5}$$

A *final ranking of tags* based on the final quality scores is obtained.

4. *Tag selection.* The top 5% tags in $tagQS(t)$ ranking are selected as *new candidate keywords* for crawling.
5. *Tag filtering.* In the end, two kinds of filters can be applied on the selected keywords. This optional step aims to reduce the frequency of useless keywords.
Pre-defined keywords filtering is used to exclude specific tags, used mainly by photographers and mass media, which are frequently present in Flickr and YouTube posts. These include camera names such as "nikon", "canon", etc. and social media names such as "flickr", "instagram", etc.
Distribution-based keyword filtering is used to exclude tags based on their crawling impact. This filtering is performed after the resulting keywords are used to crawl new posts, taking into consideration the geographical distribution of the posts obtained to detect not significant and risky keywords.
   – *Not significant keywords* are those leading to posts distributed too sparsely with respect to the event extension. This requires to: i) *Estimate the event extension.* Considering only the georeferenced posts previously extracted with seed keywords in the target timeframe, the centroid (coordinates) is computed, which is an approximation of the event centre. Then, the event extension is estimated calculating the average distance from the centroid to all the seed keywords media. ii) *Evaluate posts distribution.* If the average distance of the media extracted with new keywords with respect to the estimated centre of the event is lower than the estimated event extension the new keyword is kept, otherwise it is filtered out.
   – *Risky keywords* are defined as the ones for which the geographical distribution can not be evaluated, since there is no extraction of georeferenced media. Risky keywords are also excluded.

4.2.2 Area-related keywords generation

The whole generation process of area-related keywords is similar to the one described above for event-related keywords. The notable differences are:

- The social media used: only georeferenced media of Flickr are used to generate this kind of keywords. Since the goal is to characterize an area, posts must be originated from that area with $\approx$ 100% precision. If locations obtained by geolocating with a geolocation algorithm the not-georeferenced Flickr and YouTube posts were used, the confidence of the result would be lowered, due to potential inaccuracies of the geolocation algorithm.
- The data used: in Equations (1) and (2) the two different sets of posts to be used are no longer C and P but, respectively, In: set of posts which were taken in the current timeframe and are located *inside* the target cluster and Out: set of posts which were taken in the current timeframe and are located *outside* the target cluster.

4.2.3 Points Of Interest (POI) keywords generation

This step generates keywords concerning the relevant Points Of Interest in the identified areas.

POIs belonging to each event-affected area identified with clusters are retrieved from OpenStreetMap (OSM), through the Overpass API,[14] querying for places tagged as specific types, e.g., *railway=station*, *bridge=yes* and *place=square*.[15] Then, each post in the cluster (for the current timeframe) is associated to the possible POIs to which it refers.

Each POI in OSM can have a point, line or polygonal shape, and this difference must be considered for media to be associated to them. Moreover, a distance threshold must be set in order to consider a media near a POI. Figure 4 sketches how media are associated to POIs. In the current implementation, a threshold distance of 20 meters was set.

4.2.4 Generated keywords management

Once new keywords have been generated, they are handled by an algorithm, with the goal of maintaining a list of up-to-date keywords during the event's evolution. The handling algorithm assigns a score to the keywords for their reuse in subsequent 24-hour timeframes.

As described in Section 3, keywords for an event are generated iteratively at each timeframe and added to the previous set of keywords. As events evolve, and correspondingly relevant keywords may change, we introduce a Generated Keywords Management step to introduce a *memory* between timeframes, with the goal of evaluating whether, as time passes, keywords are still relevant for the ongoing event.

Intuitively, if the same keyword has been extracted for n subsequent timeframes — considered valid for the event, but it is not extracted in the $n+1$th one — it is still a good idea to keep such keyword for at least one extra timeframe in order to account for "holes" in the detection and to remove it only if it continues not to be extracted in subsequent timeframes.

[14] `https://wiki.openstreetmap.org/wiki/Overpass_API`

[15] The total amount of tags available in OSM (`https://taginfo.openstreetmap.org`) is huge and their type is not fixed. Tags referred to locations typically affected by emergency events were selected and are not listed here for the sake of brevity.

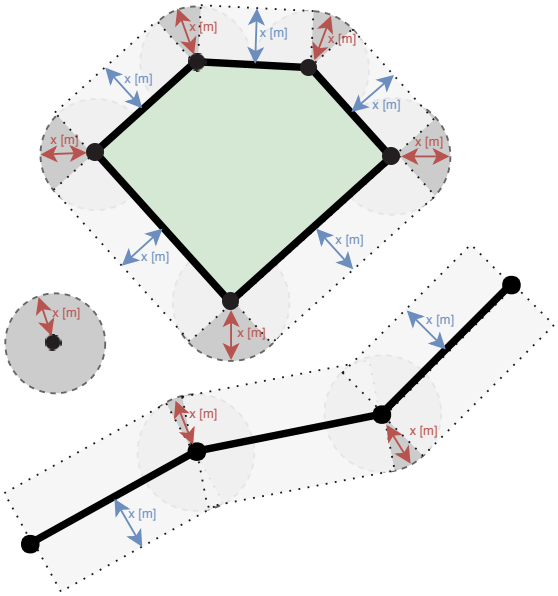

Fig. 4: Media association strategies to POIs

A scoring mechanism is employed for such implementation. Any new keyword has a validity score, which starts with 1 for new keywords. Each time a keyword is re-extracted, this score is increased by 1 (up to a maximum[16]); when an existing keyword is no longer extracted its score is reduced by 1. A keyword is effectively removed only when its score reaches 0.

This solution makes it possible to handle data scarcity and noise, and to also monitor ongoing (even minor) effects of previously extracted major occurrences.

5 Experimental evaluation

Three case studies, characterized by different features, were selected for the evaluation. While general results are provided for all of them, more detailed results and discussions focus on a single case study. The evaluation focuses on the *amount* of media extracted through the proposed methodology and their *relevance* for the emergency management task. Relevance is evaluated based on the manual annotations of the retrieved posts by three annotators. These look for images or videos that can be useful for emergency responders to assess the emergency event's evolution and for rapid mapping. The proposed methodology's performance is compared to an approach based only on a set of fixed seed keywords.

The proposed algorithm was developed in Python 3.6. Some external libraries supported the coding, e.g. scikit-learn (DBSCAN), scipy.geometry (convex hulls) and NLTK (tokenization of posts titles and descriptions). OpenStreetMap maps were produced for affected areas. Maps include stories built over the geotagged

[16]In the current implementation, the maximum has been fixed at 2.

posts extracted. The work was automated with Folium library, working with Leaflet. OSM was also exploited to extract POIs, through Overpass API (with the support of a Python wrapper: overpass-api-python-wrapper). In case of polygonal POIs associations, Shapely library accomplished the geometric analysis. The CIME algorithm developed in the E2mC project was used [8] to geolocate posts.

Section 5.1 presents selected case studies and the parameter setting, while Section 5.2 includes the results and their discussion. Finally, a discussion on the method's use in rapid mapping scenarios is provided in Section 5.3.

5.1 Case studies and experimental setting

The considered studies include 1) Hurricane Harvey,[17] which lasted from 17th August 2017 to 2nd September 2017 and was analyzed for the period 28-30th August 2017, 2) the 2013-2014 United Kingdom winter floods,[18] which lasted from around 5th December 2013 to 25th February 2014 and was analyzed for the period from 10-12th February 2014, and 3) the Central Italy earthquake, where the first major shock occurred on 24th August 2016 at 1:36 UTC, and was analyzed from August 24th to August 26th. Selected periods correspond to periods in which Copernicus Emergency Management Service (EMS[19]) activations for rapid mapping were performed (activation No. EMSR229, EMSR069, EMSR177, respectively).

The seed keywords were: 'flood' and 'inundation' for the UK floods and 'hurricane', 'flood', 'inundation', 'huracán', 'inundar', 'inundación' (to also account for Spanish posts) for Hurricane Harvey. Seed keywords chosen for the Central Italy earthquake were 'earthquake' and, in Italian, 'terremoto' (earthquake), 'sisma', 'scossa sismica' (tremor) and 'faglia' (fault).

The considered case studies, apart from being different in their typology, are also different in terms of event duration. In fact, UK floods are a *long event*, as the event lasted several weeks, while Hurricane Harvey and the Central Italy earthquake are *short events*, with a sudden change of state in the area.

The following parameters were set: each time frame was equal to 24 hours; DBSCAN parameters were set as $minPts = 10$ and $\epsilon = 6km$; the period considered for extracting pre-event posts was three months.

5.2 Results

Fig. 5 shows the amount of unique media extracted through seed and generated keywords, together with the related increase given by generated keywords, for both Flickr and YouTube. Notice that the number of posts obtained through generated keywords does not consider posts already obtained with seed keywords, that is, they are *new* posts. For simplicity's sake, only each time frame's first day of crawling was considered in these summary results, i.e., we did not consider delayed extractions in the figures, but only the media immediately available in the same time frame.

[17] https://en.wikipedia.org/wiki/Hurricane_Harvey

[18] https://en.wikipedia.org/wiki/2013%E2%80%9314_United_Kingdom_winter_floods

[19] https://emergency.copernicus.eu/mapping/list-of-activations-rapid

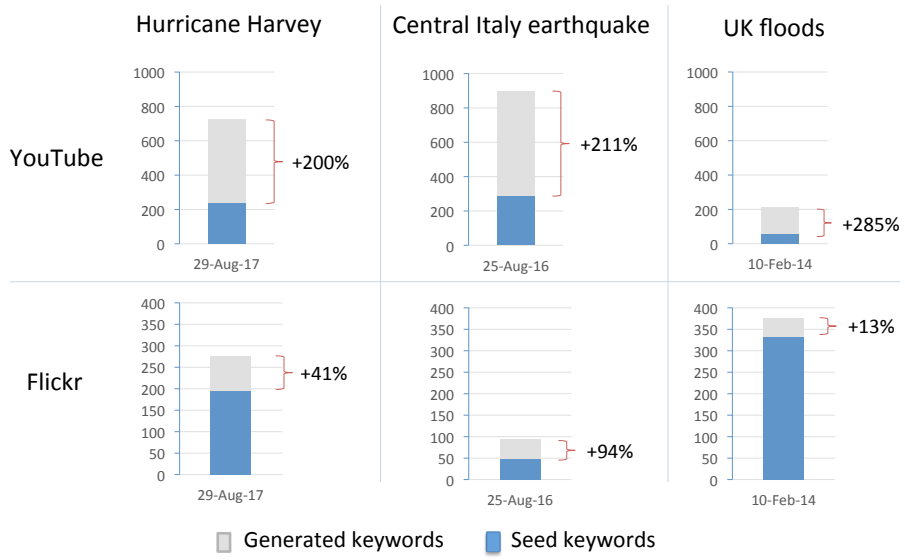

Fig. 5: Increase of retrieved posts

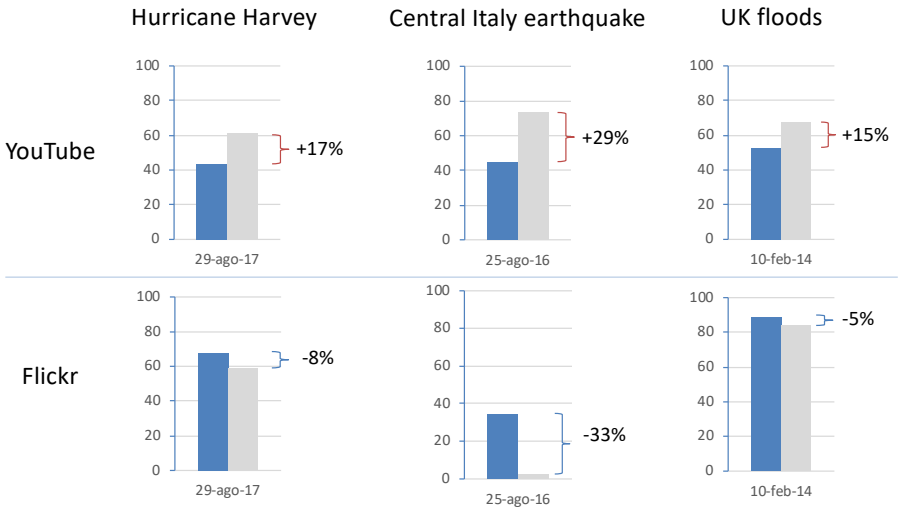

Fig. 6: Relevance evaluation

The manually-annotated relevance of media extracted through seed keywords and generated keywords is shown in Figure 6 (for YouTube, a random sample of 100 videos was selected for manual annotation), together with the variation (+/-) of relevance using the "generated" with respect to the "seed" keywords.

These results show that generated keywords can increase crawled media up to three times with respect to YouTube-only seed keywords. The increase is smaller in Flickr, but it is still significant ranging from 13% to 94%. These improvements in recall do not generally cause a loss of precision: YouTube media extracted with generated keywords are even more relevant than those extracted with seed keywords, and Flickr media extracted with generated keywords often have a comparable relevance (5-8% difference in two cases, 33% difference in one case). Results show the benefits of cross-social triangulation: keywords obtained mainly by starting from Flickr, lead to a higher improvement of precision and recall when used on YouTube.

Notice that the UK winter floods did not require the activation of the geolocation module, as the number of Flickr posts was sufficiently high to create clusters, while for Hurricane Harvey and the Central Italy earthquake, it was necessary to add-in some iterations, including locations derived from automatic geolocations of YouTube posts due to the sparseness of Flickr posts. Therefore, results show that the methodology can give results in both cases.

Fig. 7 shows the contribution of the three keyword generation methods in generating new keywords for the cases of Hurricane Harvey and the UK Floods. As the Figure clearly illustrates, this contribution is variable and it depends on the characteristics of each event, thus yielding different results.

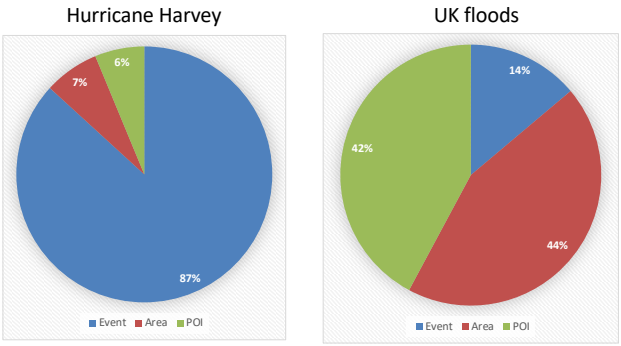

Fig. 7: Evaluation of the contribution of the three keywords generation methods

More detailed results for the Hurricane Harvey case are illustrated in Table 2, which shows the detailed number of unique posts crawled in the first two days, with their relevance, and in Table 3, where, for a single day, posts crawled through generated keywords are detailed highlighting the single contribution of each keyword and its *source*: A (area-related), E (event-related) or P (POI-related). Keywords coming from previous time frames are denoted by an asterisk. In Table 3, x/y denotes a total of y posts crawled, among which only x are new with respect to seed keywords. This table highlights that all the sources contribute to the total amount of unique posts already shown in Table 2, and that keywords comprise POIs, common words and event-specific hashtags.

Day	Extr. method	New unique		Relevance [%]	
		Flickr	YouTube	Flickr	YouTube
28/08	Seed	271	249	-	61.67%
	Generated	50	203	-	72.58%
	% Var.	**+18%**	**+82%**	-	**+10.91%**
29/08	Seed	196	243	67.76%	43.59%
	Generated	80	485	59.31%	61.07%
	% Var.	**+41%**	**+200%**	**-8.45%**	**+17.48%**
30/08	Seed	380	203	-	-
	Generated	204	289	-	-
	% Var.	**+54%**	**+142%**	-	-

Table 2: Hurricane Harvey (28-30th August 2017): summary of crawling results

Extr. method		Keywords	New media/Extracted	
			Flickr	YouTube
Seed			196/196	243/243
Generated	A*	houston	50/124	39/63
	E*	harvey	31/181	35/76
	E*	hurricane harvey	6/147	34/64
	E*	flooding	3/105	48/93
	E*	floods	3/106	47/58
	E*	flooded	3/106	53/65
	E	tropical storm harvey	3/19	75/77
	E	houston flood	2/63	93/108
	E	street flooding	0/1	34/42
	E	buffalo bayou flood	0/11	32/34
	E	tsharvey2017	0/10	0/0
	E	houston2017	0/10	0/0
	E	allen parkway	0/10	35/36
	E	tropical storm flood	0/16	0/0
	P	buffalo bayou park	0/10	24/25
	P	memorial park	2/12	8/8
	P	spotts park	0/1	3/3
			New unique [increase]	
			80 [**41%**]	485 [**200%**]

Table 3: Some more-detailed results on Hurricane Harvey (29th August 2017) - A: Area-related, E: Event-related, P: POI-related

5.3 Application scenarios

Finally, we discuss some qualitative findings originating from the analysis of the three case studies in association with the goal of performing rapid mapping activities.

In the case of unexpected events, such as an earthquake, search keywords evolve rapidly over time. In fact, the events social media characterization changes as awareness on the event evolves. The names of affected localities are not yet

known in the very initial hours after the event. Therefore, in the case of the studied earthquake, there was an evolution from only considering the main affected location i.e. 'Amatrice' as a tag in posts, to later using keywords such as 'centralitalyearthquake' for more precise information on the event. In addition, video posting also changes over time. Fig. 8 shows the evolution of the retrieved videos over time. Initially, the difficulty is to discriminate between videos coming from large nearby cities (in the considered case, Rome) which were affected in a minor way. In this case, many videos were posted showing images from interiors, like, for instance, oscillating chandeliers. This can be noticed in the low relevance shown in Fig. 6 for posts on YouTube with seed keywords. As a consequence, automatically deriving new tags (autotagging) only by applying clustering techniques on posts originating from the same social media in this case would result in a selection of posts with low relevance for the rapid mapping tasks and local awareness information. The proposed approach can be a help in finding videos really coming from the area of interest, which is not well-defined initially, and not from nearby areas which are not affected by the event.

In addition, it must be noted that the type of crawled videos changes its nature as events of this type evolve, moving from event coverage directly from the field, to a focus on supporting operations and reports from officers in charge. The method proved particularly effective to quickly identify the videos posted by emergency responders in the first hours after the event to increase awareness. These posts, and in particular the ones containing aerial images, can be very useful for rapid mapping before images are acquired from satellites.

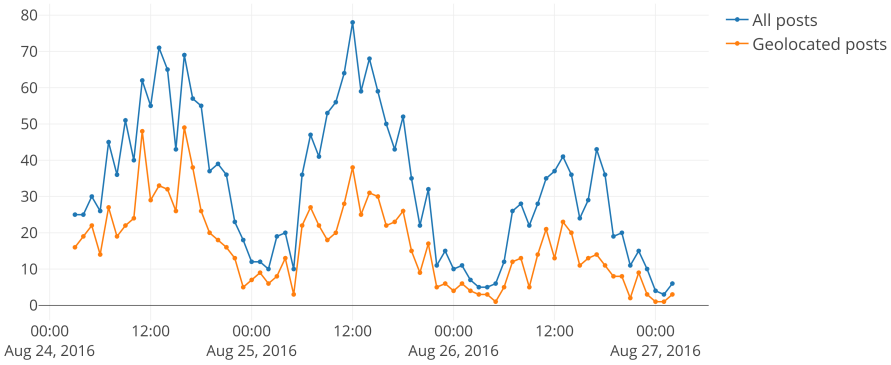

Fig. 8: YouTube posts crawled for the central Italy earthquake in the first three days

The goal is to create crisis maps to support first emergency responders, like the one shown in Fig. 9. This depicts witness maps created in the E2mC project to display posts (from Twitter, YouTube, and Flickr) located as precisely as possible. The figure illustrates how it is possible from the map to retrieve the images and videos connected to a point directly from social media, and to add other annota-

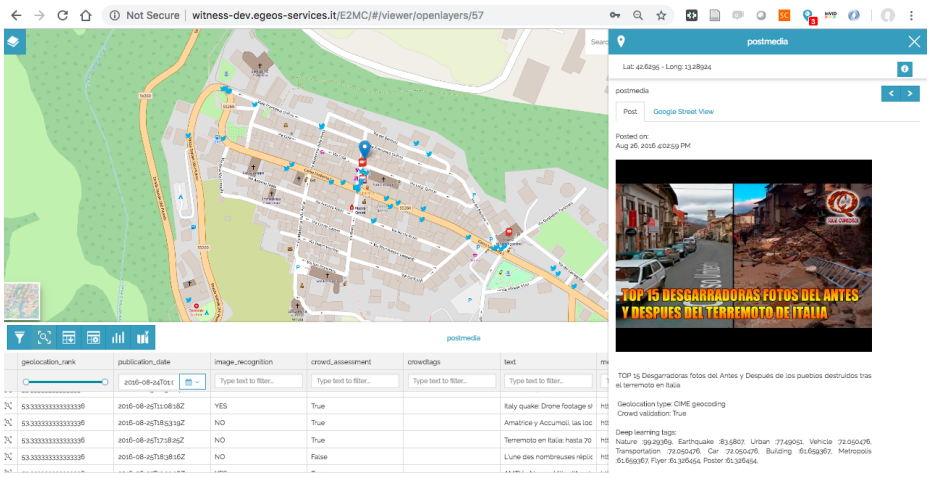

Fig. 9: E2mC witness presentation of a crisis map with visual content

tions to each post, such as the crowd's validation and automatic image analysis evaluations. Details of the E2mC approach to integrate different post evaluation tools can be found in [11].

As regards the coverage of events, news sources, such as newspapers and TV channels, provide a wide and prompt video coverage of major ongoing emergency events. Live coverage is common in events like storms, for which the landfall and event's unfolding is forecast, and several sites are available to show the expected evolution. It should be noted that live coverage in streaming retrieved through Twitter might later be transformed into videos posted by news channels on YouTube. For instance, The New York Times provides a video with pre- and post-event images, comparing Google Street View with the current situation (`https://www.youtube.com/watch?v=YzQGgyrxXiI`). This type of information is important to increase awareness in the first emergency phases and to show how the event unfolds.

Using the methodology's intermediate results, it is also possible to automatically create *stories* from retrieved posts, including pre-event timeframes, and creating before/after image comparisons of interesting locations at different times. An example is shown in Fig. 10, illustrating the evolution of a road flood over time. In this way, it could be possible in the future to automatically generate reports similar to the one previously described. More details about stories generation can be found in [4].

6 Concluding remarks

The aim of this paper is to increase the recall and precision of social media crawling from Flickr and YouTube in order to provide awareness information in emergency events, by dynamically mining search keywords during an emergency. Keywords mining is mainly based on spatial and temporal features, and language dependent functionalities are limited solely to (optional) filtering. The proposed methodology

Fig. 10: An example of history on a map

could also be extended to other social media used in the various emergency phases depending on the information they provide, ultimately to increase the number of posts crawled.

Future work will further analyze the impact of different parameters on results, in particular, different time frame lengths, clustering methods, seed keyword sets and thresholds. More comparisons to other related methods should also be considered. Ongoing work also aims to automatically analyze images in order to improve relevance and to compare and match media automatically, further increasing precision and recall.

Acknowledgements This work was funded by the European Commission H2020 project E^{2}mC "Evolution of Emergency Copernicus services" under project No. 730082. This work expresses the opinions of the authors and not necessarily those of the European Commission. The European Commission is not liable for any use that may be made of the information contained in this work. The authors thank Chiara Francalanci and Paolo Ravanelli for their support throughout this work and Nicole Gervasoni for her support in ground truth analysis and annotations.

References

1. O. Ajao, J. Hong, and W. Liu. A survey of location inference techniques on twitter. *Journal of Information Science*, 41(6):855–864, 2015.
2. J. Ao, P. Zhang, and Y. Cao. Estimating the locations of emergency events from twitter streams. In *Proceedings of the Second International Conference on Information Technology and Quantitative Management, ITQM 2014, National Research University Higher School of Economics (HSE), Moscow, Russia*, pages 731–739, 2014.

3. F. Atefeh and W. Khreich. A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164, 2015.
4. A. Autelitano. Spatio-temporal cross-social media mining for emergency events, Master's Thesis, Politecnico di Milano, Milan, Italy, 2018.
5. C. Castillo. *Big crisis data: Social media in disasters and time-critical situations*. Cambridge University Press, 2016.
6. C. Francalanci, P. Guglielmino, M. Montalcini, G. Scalia, and B. Pernici. IMEXT: A method and system to extract geolocated images from tweets — analysis of a case study. In *2017 11th International Conference on Research Challenges in Information Science (RCIS)*. IEEE, May 2017.
7. C. Francalanci, B. Pernici, and G. Scalia. Exploratory spatio-temporal queries in evolving information. In *Mobility Analytics for Spatio-Temporal and Social Data - First International Workshop, MATES 2017, Munich, Germany, September 1, 2017, Revised Selected Papers*, pages 138–156, 2017.
8. C. Francalanci, B. Pernici, G. Scalia, and G. Zeug. Talking about places: Considering context in geolocation of images extracted from tweets. In *GI-Forum 2018, Issue 1, Salzburg, July 2018, Short paper*, pages 243–250, 2018.
9. M. M. Haklay and P. Weber. OpenStreetMap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, 2008.
10. C. Hauff. A study on the accuracy of Flickr's geotag data. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 1037–1040. ACM, 2013.
11. C. Havas, B. Resch, C. Francalanci, B. Pernici, G. Scalia, J. L. Fernandez-Marquez, T. V. Achte, G. Zeug, M. R. R. Mondardini, D. Grandoni, B. Kirsch, M. Kalas, V. Lorini, and S. Rüping. E2mC: Improving emergency management service practice through social media and crowdsourcing analysis in near real time. *Sensors*, 17(12):2766, 2017.
12. C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, Baltimore, MD, USA*, pages 55–60, 2014.
13. S. E. Middleton, G. Kordopatis-Zilos, S. Papadopoulos, and Y. Kompatsiaris. Location extraction from social media: Geoparsing, location disambiguation, and geotagging. *ACM Trans. Inf. Syst.*, 36(4):40:1–40:27, 2018.
14. G. Panteras, S. Wise, X. Lu, A. Croitoru, A. Crooks, and A. Stefanidis. Triangulating social multimedia content for event localization using Flickr and Twitter. *Transactions in GIS*, 19(5):694–715, 2015.
15. S. Pezanowski, A. M. MacEachren, A. Savelyev, and A. C. Robinson. Senseplace3: a geovisual framework to analyze place–time–attribute information in social media. *Cartography and Geographic Information Science*, 45(5):420–437, 2018.
16. D. Pohl, A. Bouchachia, and H. Hellwagner. Automatic identification of crisis-related sub-events using clustering. In *11th Intl. Conf. on Machine Learning and Applications, ICMLA, Boca Raton, FL, USA, Volume 2*, pages 333–338, 2012.
17. Q. Qu, C. Chen, C. S. Jensen, and A. Skovsgaard. Space-time aware behavioral topic modeling for microblog posts. *IEEE Data Eng. Bull.*, 38(2):58–67, 2015.
18. B. Resch, F. Uslaender, and C. Havas. Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartography and Geographic Information Science*, 45(4), 2018.
19. G. Scalia. Network-based content geolocation on social media for emergency management, Master's Thesis, Politecnico di Milano, Milan, Italy, 2017.
20. E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu. Dbscan revisited, revisited: Why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):19, 2017.
21. K. Tamura and T. Ichimura. Density-based spatiotemporal clustering algorithm for extracting bursty areas from georeferenced documents. In *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, pages 2079–2084. IEEE, 2013.
22. X. Wang, L. Tokarchuk, F. Cuadrado, and S. Poslad. Exploiting hashtags for adaptive microblog crawling. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 311–315. ACM, 2013.
23. X. Zheng, A. Sun, S. Wang, and J. Han. Semi-supervised event-related tweet identification with dynamic keyword generation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1619–1628. ACM, 2017.