presented at MATES Workshop at VLDB'18, Rio de Janeiro, August 2018

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

Andrea Autelitano, Barbara Pernici, Gabriele Scalia
Department of Electronics, Information and Bioengineering
Politecnico di Milano
Milano, Italy
[name.lastname]@polimi.it

## 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 for rapidly creating crisis maps during emergency events. While most of the social media support keyword-based queries, the amount and the accuracy of the retrieved posts depends largely on the keywords employed. The goal of the proposed methodology is to automatically and dynamically extract relevant keywords for ongoing events in order to ultimately crawl as much relevant posts as possible. This is accomplished taking into account the spatio-temporal features of the monitored event to better characterize it during its evolution and through cross-social crawling in order to exploit the specificities of a social media on the others. The methodology has been implemented on Flickr and YouTube and evaluated on two recent major emergency events demonstrating a large increment in the number of crawled posts with respect to using simple generic keywords and their high relevance for the scope.

Keywords: media mining, keyword extraction, adaptive crawling, emergency management, social media

**PVLDB Reference Format:**
Andrea Autelitano, Barbara Pernici, Gabriele Scalia. Spatio-temporal mining of keywords for cross-social crawling of emergency events. *PVLDB*, 11 (8): xxxx-yyyy, 2018.
DOI: https://doi.org/TBD

## 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 a great attention is emergency management, and information extracted from social media has proven 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 about the current situation of the area affected by the emergency. Rapid mapping professionals can find beneficial to be supported by on-site visual information, and images/videos taken by users and uploaded on their personal social media accounts represent a new valuable resource, as explored in the E$^2$mC European Project [8][1].

To gather these data what is typically done is crawling social media with some seed keywords relative to the event type. However, this approach has some important limitations: generic keywords could be unable to extract many posts, and, at the same time, those extracted could be not very precise (that is, relevant to the event) [16]. This is due to the noisy and dynamic nature of social media, characterized by ambiguities, fakes, and trending keywords which arise and evolve, often without a central coordination, to describe events and situations. In addition, the number of georeferenced posts is very low, usually less than 3%, and the posts containing images are only a fraction of those [6]: from our previous studies, the expected number of posts containing images which are useful for rapid mapping activities for an emergency event from Twitter is around 0.1%. Therefore, there is a need of increasing the number of posts that can provide useful visual insights about the event.

The goal of this paper is to present an approach to extract more (and more relevant) images and videos from social media during an emergency event, dynamically mining event-related keywords 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 the extracted keywords, followed by a cross-social crawling.

The proposed methodology is designed for ongoing events and to continuously refine the keywords as they evolve.

In Section 2 we discuss the state of the art related to the use of social media in emergency situations. In Section 3 we introduce the methodology for incrementally extracting media leveraging on keyword extraction and in Section 4 we illustrate the keyword generation methodology. The experimental evaluation of the approach 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]. In this section, we focus on analyzing the state of the

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


*Proceedings of the VLDB Endowment,* Vol. 11, No. 8

DOI: https://doi.org/TBD

art with respect to the requirements emerging from the use of social media in rapid production of crisis maps and we position our work.

A problem that arises 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, in our work we need to precisely locate the media contents, rather than other user information, starting from available features (in particular, text), when not enough georeferenced posts are available. To do this, we leverage our geocoder called CIME (Context-based IMage Extraction) developed in a previous work [10]. The CIME geolocation algorithm aims at geolocating posts starting from text and contextual information and is based on Stanford Core NLP and OpenStreetMap (OSM)[2]. Using OSM with respect to other commonly used gazetteers, such as GeoNames[3], has the advantage of providing an increased granularity of available locations.

In this work we perform iterative keywords generation based on a spatio-temporal analysis. [13] combines unsupervised machine learning techniques and spatio-temporal analysis for damage assessment in emergency events, obtaining relevant topics for the affected areas. Keywords obtained in our work can be considered topics to some extent, but our goal is not damage assessment but enhancing crawling, and a better assessment is an indirect consequence of more (and more relevant) media crawled.

Clustering is a key step in our methodology. Clustering in this domain is typical of event detection techniques [3] and to estimate the affected areas [2]. In particular, density-based clustering (as DBSCAN) has provided effective for this goal [15]. In our work we use density-based clustering for the same purposes (estimate the affected areas), but it is a mean for a different final goal, which is keywords extraction.

In [17], 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 we leverage some techniques described in this work, our methodology is totally unsupervised and focuses on tags in Flickr and YouTube rather than words in Twitter.

Several integrated frameworks to analyze crisis events through information extracted from social media exist, such as SensePlace3 [11], 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 the crawling phase, even if 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.

Our work can be considered a keyword-based *adaptive crawling*, as, for example, [16]. The difference is that our keywords generation is the result of a spatio-temporal modeling and analysis rather than an analysis of the post stream.

This work starts from the analysis and the considerations contained in [6], where a model for managing the evolving spatio-temporal information available in social media is proposed.

Regarding to social media, most of the papers focus on Twitter [5], either as a primary or secondary source [9]. On the other hand, other social media can provide useful information for rapid mapping.

[9] proposes a multi-social triangulation approach starting from Twitter to extract keywords to crawl Flickr, which provide information not available in Twitter. We use the concept of triangulation by exploiting Flickr posts, which have precise geolocation, to mine keywords then used on YouTube, which provide many relevant media.

The identification of subevents from social media using clustering of posts from Flickr and YouTube has been advocated in [12], with the identification of subevents based on terms similarity and post coordinates.

Social media present different characteristics for the purposes of this work, as summarized in Table 1: the feasible searches (by keywords or by location), the presence of native geolocation of posts, time information. As shown, georeferences are not always available and they are usually referred 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. We notice that 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 [7], therefore it is a good candidate for identifying reliable areas of interest for an emergency event. YouTube does not provide such metadata, but since both Flickr and YouTube use tags in their posts, we propose in this paper a triangulation approach for mining keywords (tags) from Flickr posts and using them in YouTube.

The present work is developed within the E2mC (Evolution of Emergency Copernicus Services) European project [8], which has the goal of improving rapid mapping with the help of social media.

# 3. MEDIA EXTRACTION METHODOLOGY

## 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, relevant to 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, finding those *emerging keywords* related to the event.

To accomplish this goal, three directions are explored:

*Temporal.* Given a target time frame, the amount of posts related to it (even if posted later) increases over time. More importantly, analyzing the past can give hints about 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. Being able to characterize that area with respect to the rest of the world, it is possible to filter out the unrelated keywords keeping only the event-related ones.

[2] http://www.openstreetmap.org/

[3] http://www.geonames.org/

| | Facebook | YouTube | Instagram | Twitter | Flickr |
|---|---|---|---|---|---|
| **Feasible searches** | Not feasible[a] | Keyword-based | Not feasible[a] | Keyword/Location-based (also in streaming) | Keyword/Location-based |
| **Localization of posts** | GPS (of post)/Manual tag | No coordinates | Manual tag | GPS (of post)/Manual tag | GPS (of media) |
| **Time information** | Post publication | Post publication Recording time[b] | Post publication | 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, so it has not been used.

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

*Cross-social.* As explained before, each social media has different features. However, 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 [9].

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

Another characteristic of our 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 evolution of the event. 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).

To follow the *evolution* of the event, we adopt a *sliding windows* approach to crawling. This is relevant mainly for Flickr which distinguishes between the *date taken* (DT) and the *date uploaded* (DU) of a media. If we target a certain interval of time, which is a *time frame*, we can get more media as time passes since there could be posts with DT $\in$ *time frame* even if DU is after the end of the time frame, since in many cases in Flickr there is a delay in posting pictures, so an image taken one day could be posted 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. A crawling $x.y$ means that the $x$-th time frame is crawled taking into account posts uploaded until the $y$-th. The first 24-hours time frame (from 10/02/2014 to 11/02/2014) is crawled three times: 1.1, 1.2 and 1.3, retrieving each time media with the corresponding DT and DU. The second time frame is also crawled three times: 2.2, 2.3 and 2.4, and so on.

Since the distinction between DT and DU does not exist on YouTube, it is crawled only for 1.1, 2.2., etc.

For the current work, we considered windows of 24 hrs and 3 days of iteration. These parameters can be varied, provided that the number of posts to be analyzed is sufficient to identify areas of interest (see Section 4.1).

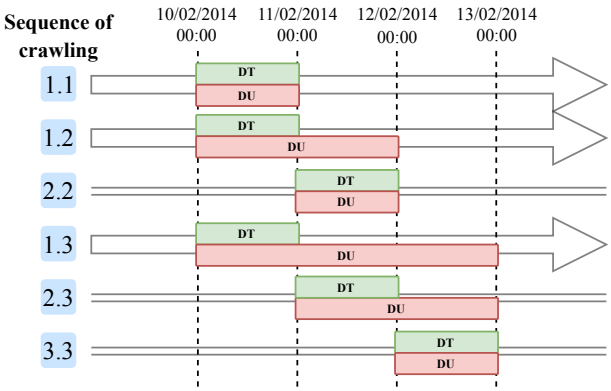


Figure 1: Crawling sequence example

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

The main steps of the methodology, depicted in Figure 2, are summarized in the following:

0. *Initial input.* Seed keywords (relative to the event type) and a past date/time preceding at least 24 hours the beginning of crawling have to be manually decided and constitute the initial input.

1. *Sliding window crawling.* For each 24-hours time frame, the crawling engine starts creating the *sliding windows* to account also for media uploaded after when they were taken as explained previously. Flickr and YouTube are crawled by keywords for the selected sliding windows. Flickr media are enriched also with additional georeferenced media extracted from same albums and relevant groups[4].

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

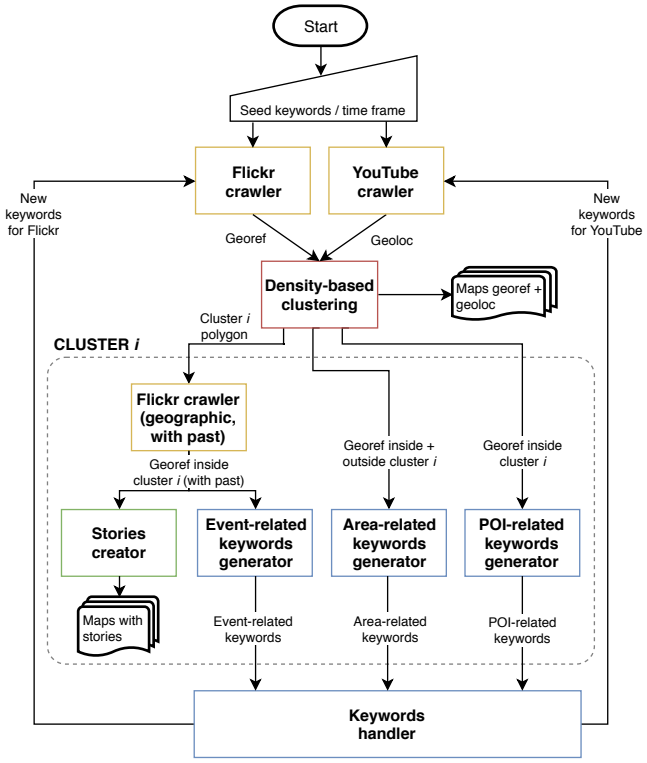


Figure 2: General schema of the methodology

2. *Spatio-temporal clustering.* Spatio-temporal density-based clustering is performed on all the georeferenced Flickr media extracted so far. In this way, the most event-affected areas can be detected. If too few clusters are produced using only georeferenced Flickr media, YouTube geolocated media are also used.

3. *Keyword generation.* For each identified area (cluster), three parallel steps are performed:

   - *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 past months in the target area. This step enables also the creation of "stories", which are "before/after" comparisons of interesting POIs.
   - *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.

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

The two main steps characterizing our approach are described in the following: clustering (Section 4.1) and keyword generation (Section 4.2).

# 4. INCREMENTAL KEYWORDS MINING

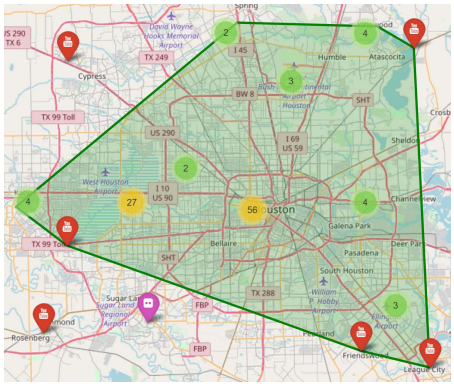

Figure 3: Cluster example

## 4.1 Spatio-temporal clustering

Spatio-temporal clustering identifies sub-areas characterized by a higher posting activity related to the event.

Density-based clustering (DBSCAN [14]) has been chosen for the task. Indeed, a higher density of posts related to domain-specific keywords is typically related to emerging events [13]. In addition, it accounts for affected areas which vary in number and shape.

The clustering is spatial because posts are interpreted spatially (without any normalization) as geographical points (latitude, longitude) on the map, using the haversine distance[5], and it is temporal since posts are clustered at each time frame independently, following the evolution of the event. An example of cluster is shown in Figure 3.

Clustering is done preferentially on georeferenced Flickr media. This should provide an approximate but faithful representation of the event-affected areas in the real world, since, as described in Section 2, georeferences are the most accurate kind 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 not reflecting real world affected areas), since the following keywords generation phase has 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 provide enough clusters (3 in the current implementation — more complex and flexible models could be employed).

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

In this domain *minPts* must be not too low, since some fakes/not relevant media exist and it is necessary to consider this noise. $\epsilon$ must account for the sparsity of posts: ideally, relevant areas should be constituted by close points, but factors like 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 not achievable. Therefore setting $\epsilon$ is a trade-off between being not too low, because there would be a lot

[5] https://en.wikipedia.org/wiki/Haversine_formula

of areas not found due to missing data, and not too high, because the goal is to find affected areas within the event. $\epsilon$ can also depend on the amount of media extracted: less they are, more likely it is that an insufficient number of posts is associated to an affected area, however, choosing $\epsilon$ too high might affect the relevance of the mined keywords for the event, as it will include also not affected areas.

## 4.2 Keywords generation

Three categories of keywords are generated starting from each identified event-affected area:

- *Event-related keywords:* keywords generated considering posts in the target area with respect to posts in the same area in the past. These keywords should refer to the event, because they are the keywords emerging in the present with respect to the past.

- *Area-related keywords:* keywords generated starting from posts in the target area with respect to all the other posts (for the same time frame). 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 subset of relevant POIs inside each area. They are essentially the POI names extracted from the gazetteer (OSM) in the target area for the considered time frame.

On each time frame several iterations are performed to refine the identified clusters, 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 allows crawling new media, which allow to refine the clusters, which in turn could allow to compute new keywords in a cycle.

Theoretically, the iterative generation of keywords can be done until a fixed point is reached. However, for practical reasons, a limit has to be set[6].

In the following details about the generation of each category of keywords are provided.

### 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 time frame 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 in the following:

[6]In the current implementation at most 3 cycles are accomplished.

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 [17]:

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

   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 cluster, in the past |
   | $C(t)$ | 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 [17]:

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

   where:

   | | |
   |---|---|
   | $x = \frac{\|C(t)\|+\|P(t)\|}{\|C(t)\|}$ | is the *coverage ratio* |
   | $\gamma$ | is a hyper-parameter to be empirically set[7]. |

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

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

   At 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. They are tokenized and for each token $w$ its quality score is produced by using the aforementioned method for tags.

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

3. *Tag quality scores.* Each tag's temporary quality score obtained at Step 1 is refined by the token score of the

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

same candidate keyword (if any) found in titles and descriptions, thus obtaining a *final quality score* for each tag:

$$tagQS(t) = tagQS^*(t) \cdot (1 + tokenQS(t)) \quad (5)$$

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.* At 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 cameras names like "nikon", "canon", etc. and social media names like "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 which bring to posts too sparsely distributed 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 considered time frame, the centroid (coordinates) is computed. This centroid is an approximation of the event centre. Then, 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 media extracted with new keywords with respect to the estimated centre of the event is less 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, because no georeferenced media is extracted from them. 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 for event-related keywords presented above. The notable differences are:

- 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. Using geolocated media (from not-georeferenced Flickr posts or YouTube posts) would lower the confidence of the result, due to potential inaccuracies of the geolocation algorithm.
- Data used: in Equations (1) and (2) the two different sets of posts to use are no longer $C$ and $P$ but, respectively, $In$: set of posts which were taken in the current time frame and are located *inside* the target cluster and $Out$: set of posts which were taken in the current time frame and are located *outside* the target cluster.

#### 4.2.3 Points Of Interest (POI) keywords generation

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

POIs belonging to each cluster are retrieved from OpenStreetMap (OSM), through the Overpass API[8], querying for places tagged as specific types, e.g., *railway=station*, *bridge=yes* and *place=square*[9]. Then, each post in the cluster (for the current time frame) is associated to the possible POIs it refers to.

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

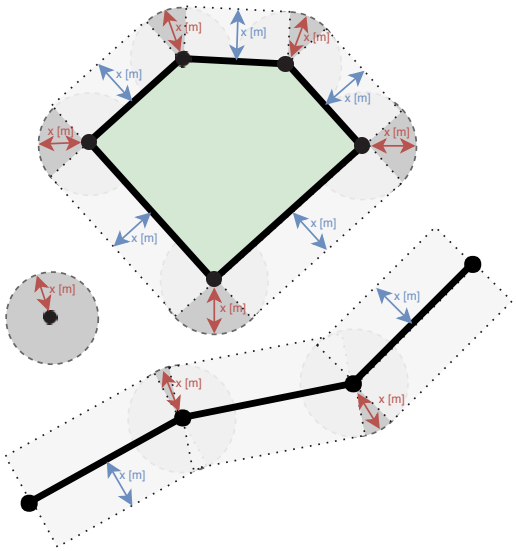


Figure 4: Media association strategies to POIs

#### 4.2.4 Generated keywords handling

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 evolution of the event. The handling algorithm assigns keywords a score for their reuse in subsequent 24-hour time frames.

As described in Section 3, keywords are generated iteratively at each time frame and, given the fact that the event is the same, the new set contains both already existing keywords and new generated keywords. The goal of this phase is to introduce a *memory* between time frames with normalization purposes. Intuitively, if the same keyword has been extracted for $n$ subsequent time frames, considered valid for the event, but not at the $n+1$th one, it is anyway a good idea to keep it at least one time frame more to account for

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

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

"holes" in the detection and removing it only if the keyword keeps not being extracted in the subsequent time frames.

To do this, a scoring mechanism is employed. Any new keyword has a score, which starts with 1 for new keywords, and represents its validity. Each time a keyword is re-extracted this score is increased by 1 (until a maximum) and when an existing keyword is not extracted anymore its score is reduced by 1. A keyword is effectively removed only when its score reaches 0. The maximum has been fixed at 2 in the current implementation.

This solution allows managing data scarcity and noise, and also monitoring ongoing (even minor) effects of previously extracted major occurrences.

# 5. EXPERIMENTAL EVALUATION

## 5.1 Case studies and experimental setting

The considered studies include 1) Hurricane Harvey[10], which lasted from 17th August 2017 to 2nd September 2017 and has been analyzed for the period 28-30th August 2017 and 2) the 2013-2014 United Kingdom winter floods[11], which lasted from around 5th December 2013 to 25th February 2014 and has been analyzed for the period from 10-12th February 2014.

Seed keywords are: 'flood', 'inundation' for the UK floods and 'hurricane', 'flood', 'inundation', 'huracán', 'inundar', 'inundación' (to account also for Spanish posts) for Hurrican Harvey. Each time frame is 24 hours. DBSCAN parameters have been set as: $minPts = 10$ and $\epsilon = 6km$.

For each case study we apply the algorithm, then we evaluate it considering as a validation parameter, the relevance of the extracted images for the crisis raping mapping purpose.

## 5.2 Results

Tables 2 and 3 summarize the results for the two cases.

For each analyzed 24-hours time frame the amount of unique media extracted thanks to seed and generated keywords is shown, together with the relative increment 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 sake of simplicity, in these summary results we consider only the first day of crawling of each time frame, i.e., we do not show delayed extractions, but only media immediately available.

The manually-annotated relevance of media extracted thanks to seed keywords and generated keywords is also shown for selected cases, together with the variation (+/-) of relevance of "generated" with respect to "seed". The first two days for each event event have been validated for YouTube, while just the day with higher % increase for each event has been validated for Flickr, given the high amount of media.

Results show that generated keywords can increment crawled media up to three times with respect to only seed keywords in YouTube. The increment is smaller in Flickr, but it is anyway significant ranging from 6% to 54%. 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 have a comparable relevance (5-8% difference). The results reinforce the need of cross-social triangulation: the keywords, obtained mainly starting from Flickr, bring to an higher improvement of precision and recall when used on YouTube.

Notice that UK winter floods did not require the activation of the geolocation module, while Hurrican Harvey did. Therefore, results show that the methodology is able to give results in both cases.

Further analyses need more detailed results as those shown in Figure 4 (only for the 29th August for the Hurrican Harvey). Here, 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 which come from previous time frames are denoted also by an asterisk. In this table $x/y$ 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 comprises both places, 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

| Day | Extr. method | New unique | | Relevance [%] | |
|---|---|---|---|---|---|
| | | Flickr | YouTube | Flickr | YouTube |
| 10/02 | Seed | 332 | 55 | 89.02% | 52.63% |
| | Generated | 43 | 157 | 83.72% | 67.35% |
| | % Var. | **+13%** | **+285%** | **-5.3%** | **+14.72%** |
| 11/02 | Seed | 282 | 48 | - | 54.17% |
| | Generated | 28 | 119 | - | 79.07% |
| | % Var. | **+10%** | **+248%** | - | **+24.9%** |
| 12/02 | Seed | 192 | 46 | - | - |
| | Generated | 12 | 41 | - | - |
| | % Var. | **+6%** | **+89%** | - | - |

Table 3: UK winter floods (10-12th February 2014): summary of crawling results

# 6. CONCLUDING REMARKS

The contribution of this paper is towards increasing recall and precision of social media crawling with the goal of extracting pictures and videos from Flickr and YouTube, dynamically mining search keywords during an emergency. The keywords mining is mainly based on spatial and temporal features, and language dependent functionalities are

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

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

| 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 4: Some more detailed results about Hurricane Harvey (29th August 2017)

limited solely to (optional) filtering. The proposed methodology could be extended also to other social media, used in the various phases depending on the information they provide, and ultimately to increase the number of posts crawled. Starting from intermediate results of the methodology is also possible to create stories, which are before-after image comparisons of interesting POIs [4]. Future work will further analyze the impact of the different parameters on the results, in particular different time frames, seed keywords sets and thresholds. Ongoing work aims at analyzing also images to improve relevance and automatically compare and match media, further increasing precision and recall.

## Acknowledgments

This work has been partially 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 during this work and Nicole Gervasoni for her support in ground truth analysis and annotations.

## 7. 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, 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.

[7] 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.

[8] 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.

[9] 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.

[10] B. Pernici, C. Francalanci, G. Scalia, M. Corsi, D. Grandoni, and M. A. Biscardi. Geolocating social media posts for emergency mapping. *arXiv preprint arXiv:1801.06861*, 2018.

[11] S. Pezanowski, A. MacEachren, A. Savelyev, and A. Robinson. Senseplace3: a geovisual framework to analyze place−time−attribute information in social media. *Cartography and Geographic Information Science*, 2017.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.