presented at MATES VLDB Workshop, Sept. 1st, 2017

Exploratory spatio-temporal queries in evolving information

Chiara Francalanci, Barbara Pernici, and Gabriele Scalia

Politecnico di Milano - DEIB
Piazza Leonardo da Vinci 32, Milano, Italy
{chiara.francalanci,barbara.pernici,gabriele.scalia}@polimi.it

Abstract. Using evolving information within rapid mapping activities in the response phase in emergency situations poses a number of questions related to the quality of information being provided. In this paper, we focus on image extraction from social networks, in particular Twitter, in case of emergencies. In this case issues arise about the temporal and spatial location of images, which can be refined over time as information about the event is being collected and (automatically) analyzed. The paper describes a scenario for rapid mapping in an emergency event and how information quality can evolve over time. A model for managing and analyzing the evolving information is proposed to be used as a basis for analyzing the images quality for mapping purposes.

Keywords: imprecise spatio-temporal information, exploratory queries, evolving information

1 Introduction

Information extracted from social media has proven very useful and informative in many crisis situations [7, 19]. In particular, information extracted from Twitter has been studied for its immediacy in making information about the events available.

One of the issues being studied is how to make use of this information within the emergency response activities being activated during an emergency. In particular, we focus on exploiting this information to support rapid mapping activities. Rapid mapping has the goal of providing rescue teams and operators with information about the current situation of the area being interested by the emergency. In the project "Evolution of Emergency Copernicus services" (E^2mC[1] the goal is to extend the support the activities of the existing Copernicus Emergency Mapping Service (EMS)[2] providing the rapid mapping operators with additional information derived from social networks and crowdsourcing, to integrate and complement satellite information available for the service.

Information derived from social media, and in particular images posted by eyewitnesses, is useful for rapid mapping activities only if associated with an

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

[2] http://emergency.copernicus.eu/mapping/ems/service-overview

adequate specification of the spatial (geolocalization) and temporal information about the images. However, as we discuss further in this paper, such information may not be available in connection to tweets: only a very small percentage of tweets is geolocalized, the images in the tweets may refer to other areas rather than the geolocation of twitterer, the time and place the image has been taken may not be the same as the time and place of its posting.

In emergency management, often the information related to events becomes clearer only over time, when additional information is collected. For instance, the initial location of the event might be only approximately known initially, and refined over time, or the context of a tweet become known only analyzing a number of associated tweets or other information, which help disambiguating associated information. Also in terms of time information for images, there is a need to distinguish between the time of image posting, the time when the image is taken, the relation to specific events (e.g., in an earthquake tremors sequence, the same building might be in different states after the initial tremor and subsequent tremor events).

The aim of the paper is to discuss the characteristics of information whose spatio-temporal attributes evolve in time. In particular, we analyze how analysis based on imprecise spatial and temporal information can be performed, proposing a meta-model as the basis for the retrieving spatio-temporal information, and discussing how an exploratory approach can support formulating queries and presenting relevant information to operators.

The paper is structured as follows. After discussing relevant state of the art, we present a scenario for illustrating our work in Section 3. Then in Section 4 we discuss some issues concerning data with imprecise spatial and temporal information. In Section 5, we propose a model to support exploratory queries and delineate types of exploration. Finally, we adopt the presented model to discuss a case study.

2 Related work

Data quality issues and techniques have been discussed in depth in [4]. Spatio-temporal data quality problem can be found in particular within movement data quality, as discussed in [2], which identifies three quality issues: missing data, accuracy errors, and precision errors. Temporal and spatial resolution, spatial precision, accuracy of positions (such as the accuracy of GPS data or that of the position of a GPS-enabled camera manually set by the operator) affect the quality of information associated to given spatio-temporal coordinates.

Information extraction is the task of extracting structured information starting from unstructured and noisy tweets. For example, in [18] it has been addressed applying conditional random field (CRF), a statistical model which predicts the class of a text token based on its context in the sentence. In general, the named entity recognition (NER) task aims to extract those n-grams which refer to entities of various kinds (people, locations, companies, etc.), and the nature of tweets, very different from traditional texts, poses specific challenges

[25]. The recognition task is usually followed by the disambiguation task (named entity linking, NEL) where an n-gram is linked to the exact and unambiguous entry it refers to in an external database.

Focusing on location information, the extraction and disambiguation tasks are also called *geoparsing* and *toponym resolution* respectively. In [14] it has been demonstrated that a state-of-the-art library to perform named entity recognition in a traditional setting (Stanford NER) is not able to deal with bad capitalization, misspellings etc., that are plentiful in microtext. Geoparsing and toponym resolution has been addressed mainly using statistical techniques through trained models [23, 21, 30], also in combination with heuristics [20]. The social networks have been also used to infer the location of a user [8] or to disambiguate it [15] thanks to the additional contextual information. These researches focus mainly in the user home location rather than the locations mentioned in the messages, but recently it has been proved the usefulness of social networks also to overcome the problem of shortness and sparsity of tweet messages analyzing them with respect to other tasks. For example, in [22] interactions and text similarity among tweets are used to improve the topic identification task. Using the social network as additional feature it is possible to refine the extracted information as it grows — that is, as the related event evolves.

Event detection has been performed mainly through clustering [5] and probabilistic models [26], also considering the geographical information [24] and the *spatial density* of the tweets reporting an event [28, 3].

Within emergency management, the potential of using data from social media and multiple sources in crisis situation has been studied by several authors [7, 17]. The goal in the European E^{2}mC project is to identify in social media information useful for rapid mapping activities, providing tools that help improving the production times and quality of obtained maps. In [12], we have discussed emergency mapping requirements for building an integrated service-based architecture for E^{2}mC. Part of the project is focusing on extracting useful information from tweets in the case of earthquakes emergencies with adequate tools. The IMEXT [11] tool environment is a first prototype developed in the project to support Twitter crawling with specific keywords for given event types, geotagging tweets, and extracting images from tweets and from documents linked to the tweets themselves, such as other social media and traditional media.

In recent times, the need is emerging to change the query-answer paradigm common in databases towards exploratory queries, exploratory computing, and exploration systems for big data analysis where the goal is to find interesting patterns or information in large amounts of data (e.g., [9, 29]). In their proposal for Queriosity [29], the authors propose a novel approach for developing data exploration systems, to provide insights to users, based on autonomously ranking the relevance of data, learning from users interactions and observation of the environment.

In the present paper, we discuss open issues about the use of information extracted from tweets, with particular reference to their spatial and temporal characteristics, and how an exploratory approach to data can be beneficial for rapidly identifying useful images.

3 Scenario

Copernicus is an European programme aimed at developing European information services based on satellite Earth Observation and in situ (non space) data. Its Rapid Mapping (RM) service provides on-demand and fast provision (within hours or days) of geospatial information as a emergency management service (EMS).

The products provided by the Copernicus EMS Rapid Mapping are standardized with a set of parameters the users can choose requesting them; different products are characterized by different information provided, information quality and time necessary to receive them. In particular, there are three different map types: *Reference* maps, which provide knowledge on the territory using data prior to the disaster and as close as possible to it, *Delineation* maps, which provide an assessment of the event extent (e.g. earthquake impact area map, flooded area map) and optionally its evolution using post-disaster data, and *Grading* maps, which provide an assessment of the damage grade (affecting population and assets like settlements, transport networks, industry and utilities) and optionally its evolution using post-event data. Moreover, maps can be requested in service level 1 (SL1), provided within some hours after delivery and quality approval of imagery, or service level 5 (SL5), provided typically in five working days.

One of the main challenges faced by Copernicus EMS is related to *timeliness*, since it is not unusual to experience delays up to 72 hours to receive the first information as satellite information can be incomplete (e.g. due to clouds or delays in receiving information due to satellite passages).

The E^{2}mC project tries to fill the gap demonstrating the technical and operational feasibility of the integration of social media analysis and crowd-sourced information in the Copernicus EMS improving the timeliness and accuracy of geo-spatial information, particularly in the first hours after the event. Indeed, social media are a relatively new and increasingly important source of information and one of the main advantages is related to timeliness: immediately after an event large amounts of potentially useful information and media are posted on social media. Processing social media is challenging: it is an example of “big data” with hundreds of millions of posts every day that can be overwhelming and confusing and often include personal impressions rather than useful information. Information on social media is not verified, can be incomplete and partial, even if “some really interesting and important messages do get posted, sometimes providing information that is not available through other channels” [7].

In [13], as a case study, the earthquake in Central Italy of August 2016 has been considered, analyzing the tweets posted just after the event focusing on image extraction of potentially useful and geolocated images. Focusing on image extraction in this context, the goal is to find *useful* images. To consider an image useful, near to an objective usefulness of the image itself, it is necessary to be able to collocate it precisely in time and space. However, as detailed in Section 4, some important issues are related to the *evolving* and *imprecise* information available on social media, which reflects in the *quality* of the information provided. The recognition and disambiguation phases of the locations cited in the text are prone to imprecisions, so their extraction is an open problem (see Section 2). To address the problem, in [27] a new approach has been proposed to improve the recognition and disambiguation performances using both the context provided by the other locations in the same message and the context provided by other messages in the implicit social network related to each message. In the following the locations will be extracted from text using this algorithm.

4 Evolving and imprecise information

In this paper we focus on representing and analyzing evolving information, and specifically on spatio-temporal aspects of information. We define as spatio-temporal evolution of information the process of refining the information on a specific event over time. Information becomes more precise as the event evolves, e.g., the location of the event becomes more precise over time, delineating the area of interest, tweets can be geolocated using context information provided from other sources or extending them with information derived from their analysis, external additional information may become available.

4.1 Imprecise information

As mentioned in Section 2, several authors have proposed approaches to exploit tweets as information sources during emergencies and there is a wide literature about quality of spatial and temporal information.

Even if theoretically an event is defined as something which happens at a precise time and in a delimited space (for instance, the EMSR177 activation for the Central Italy earthquake is associated with an Event Time (UTC): 2016-08-24 01:36, has an Area Descriptor: Lazio, Abruzzo and Umbria Regions, and is associated to an Activation Extent Map that provides the polygons for delimiting the areas of the grading maps with their geographical coordinates), the information extracted from social media and the nature of real events themselves bring a series of *imprecisions* with respect to the availability of both spatial and temporal information, which is necessary to take into account and address performing rapid mapping.

Spatial information Geographical information is a rare resource on social media. Only 0.5%-2% of tweets are geotagged [20, 7], and the metadata associated to images cannot be used since images loose all metadata, including their geographical coordinates, when stored by Twitter.

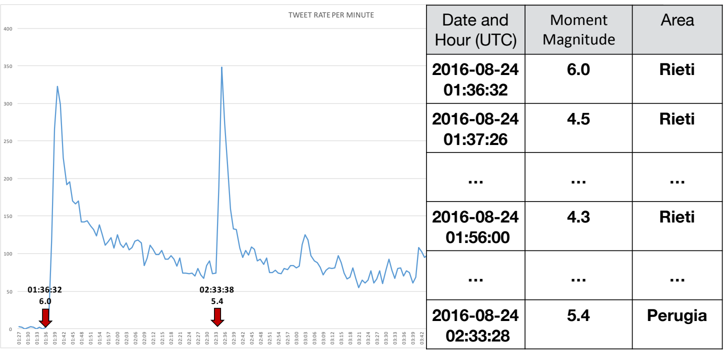
Therefore, *geolocation*, that is the activity of associating a location to the messages using other indicators like the text, the social networks, the URLs contained in the message, etc., becomes crucial. In particular, it has been demonstrated the value of extracting the locations referenced in the text [30, 20], firstly *recognizing* the toponyms mentioned and then *disambiguating* them to the exact locations they refer to. However, this task introduces imprecisions due to ambiguities which exist among location names and common names (*geo/non-geo* ambiguities) and among location names themselves (*geo/geo* ambiguities). Extracting locations from tweets has additional challenges with respect to traditional texts because the short, noisy and decontextualized nature of tweets.

Even if the locations mentioned in tweets are correctly recognized and disambiguated, they are typically imprecise, at some extent, for rapid mapping purposes. Indeed, while the coordinates associated to a geotag attached by Twitter should precisely identify the location where the tweet has been submitted, the locations mentioned in the text could be more general, citing for example a city or a region or could contain multiple references to locations, also with different levels of granularity.

A related challenge comes from the *gazetteer* used, which could not cover equally all the target locations and could contain errors. Indeed, "the output of any geocoding algorithm is only as exact as the knowledge base that underlies it" [30]. For example, GeoNames does not contain many street/road names or point of interests. Moreover, a gazetteer like GeoNames contain only few "alternative names" for each entry and they do not account for any possible way each location could be referred to. For the analysis in this paper, locations are derived using an approach based on geonames and proposed in [27], based on Named Entity Recognition libraries [1] and geonames.

Another imprecision is related to the location of the events themselves. Indeed, it is not always trivial to define precisely the boundaries of an event or establish whether two near events are actually the same one. For example, considering an earthquake, the most affected areas typically are those nearer to the epicenter. However, the actual damages will depend also from other factors, like the state of the affected buildings and the population density, therefore the most affected areas could not be spatially continuous. An earthquake is typically felt also at many kilometers of distance, causing only minor damages in more distant areas, therefore is not easy to establish the boundaries of the event to monitor. An example is given by the earthquake of 24th August in Italy. The epicenter is in Accumoli, a small municipality of 650 inhabitants, but the shake has been felt distinctly also in other near areas, in particular Rome, even if no significant damages were reported there. However, the fact that Rome is the most populated city of Italy, brought a significant number of reports from and about that city, so that, especially in the first hours, simply monitoring Twitter it seems like the main target of the earthquake is indeed Rome instead of Accumoli.

Temporal information It is not trivial to precisely assess the duration of an event in terms of its starting and ending time. If a sudden event like an earthquake has at least a clear starting point, other events like floods could start slowly as simple rains. Moreover, an event could be characterized by several subevents [7]: for example several aftershocks of an earthquake could be separated even by hours or days. Consider for example the sequence of shakes related to the earthquake of August, Italy, shown in Fig. 1. It is challenging to precisely distinguish the subevents starting from tweets: for example, after a shake, the posts and the related images could be related to a previous shake, and, even if apparently useful, being indeed outdated as information.



Date and Hour (UTC)	Moment Magnitude	Area
2016-08-24 01:36:32	**6.0**	**Rieti**
2016-08-24 01:37:26	**4.5**	**Rieti**
...	**...**	**...**
2016-08-24 01:56:00	**4.3**	**Rieti**
...	**...**	**...**
2016-08-24 02:33:28	**5.4**	**Perugia**

Fig. 1. Time of events [16]

Therefore, assigning a precise time to an information or a media could be challenging and often the only known information is an upper bound.

Following the temporal database concepts described [10], it is also necessary to distinguish the time of occurrence of events or of the information about the real world from the time in which they are recorded in the system (transaction time). In general, an evolution in time is associated to all information to be analyzed, in particular concerning the precision and accuracy of available information.

5 A model for spatio-temporal imprecise information

In this section, we propose a model for spatio-temporal information related to events that can be used to support queries and exploratory queries on an event. First, we describe the model in Section 5.1, then we propose some directions for exploring the available data.

5.1 Modeling spatio-temporal information

The goal of the model is to support information related to events that can become more precise over time and to be able to associated to it time and space information. The model is centered on the concept of *event*, which may have subevents. We assume an occurrence time and location for the event are specified, as a starting point for the exploration. To each event, related *documents* are associated, composed of *items* (carrying information) which are progressively located in time and space, as related information (*info*) becomes available (or may be automatically derived). The time of the document production is also recorded. This is not necessarily the time to be associated to its items (e.g., an image included in a tweet could have been produced at an earlier time). The documents have an *author*, whose location in time and space may be also known with different degrees of precision and could be also be variable in time. It has to be noted also that the location associated to the profile of the author could be available, but could be also misleading in some cases if it is the home location. Time and location associations are characterized by two attributes, *precision* and *accuracy* as it is customary in the data quality literature (see Section 2).

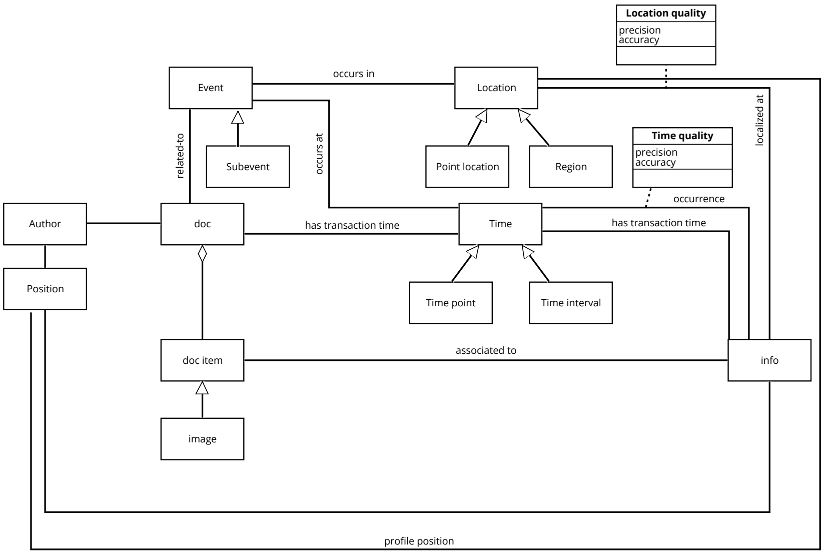


Fig. 2. Event exploration model

5.2 Exploratory queries

The risk of information overload and, on the other hand, of having too little information is present. For instance, in the first hours after the event, the contribution of new images (i.e., not duplicates of previous ones), and related to the event might be limited.

Initially, the exploration of available data (which can be massive, see for instance the tweet rates for the event in the scenario, shown in Fig. 1) has the goal of delineating the affected areas in order to facilitate rapid mapping preparation.

Immediately afterwards, the goal of the mapping operator is to find useful images. Sometimes one (or very few) image is enough (e.g., to confirm the forecast of a flood, showing initial damages, or if an image shows clearly an affected area). In other cases, the goal is to identify twitters who are eyewitnesses and therefore giving direct information on the event. The temporal and spatial information associated to documents, document items, and authors can all be useful to support the exploration. When looking for images, the goal is to identify images which are produced after the event, i.e.,
(`image.associated-to.info.occurrence.time>event.occurs-at.time`).
The time of an image precedes the time of the document in which it is inserted
(`image.associated-to.info.occurrence.time<`
`doc.has-transaction-time.time`).
However, it is not necessarily true that
(`image.associated-to.info.has-transaction.time=`
`image.associated-to.info.has-transaction.time`
and therefore `image.info.occurrence.time` could be before the event occurrence, if a tweet about an event shows an inventory image. Similar considerations can be given for spatial information: for instance, it is common to find as twitter images some landmarks about the location of the event (e.g. the Tour Eiffel for event in France, the Coliseum for events in Italy, even if Paris or Rome are not within the event area).

For the above mentioned analyses, time and space comparison operators have to support imprecise information (as described, for instance, in [6]).

Ranking of relevance of images and information is difficult (and multicriteria), and it depends on the operator and his/her area of interest, however also the actions done by the operator can be useful to improve this assessment: the selection of an image, the focus of queries on a specific area, the search for information about the author of documents are all elements that can be exploited to improve the search for useful information during the exploration.

The *query formation* process can be exploratory, navigating available info in time and in space to identify useful images, and the system should suggest interesting information to the users.

Type of queries include:

- spatial to delimit area of interest
- event time and space derivation (through constraints from document information)
- evolution in time of a point /area in space

Information suggestions, not related to a user-generated query, but autonomously provided by the system, might include selection of images by relevance to be proposed to operators, delineation of areas, identification of Twitter influencers eyewitnesses for the event, and so on.

6 Case study

In this section, we discuss some aspects related to the use of the previously described model to explore information about an emergency event.

We illustrate the evolution of tweets in two EMS activations for the earthquake in Central Italy. After the first activation EMSR177 described in Section 3, another activation was started in October (EMSR190), following a major aftershock on October 26.

6.1 Event 1 - August 24, 2016

In this event, using the crawling techniques illustrated in [11], about 150,000 possibly relevant tweets have been extracted from the 48 hours immediately following the event.

The amount of tweets (*documents* in the model described in Section 5) with locations recognized and disambiguated from text increases over time, considering the additional context provided by new tweets posted (see [27]). For example, focusing on the tweets posted in the first 30 minutes after the earthquake, 28.1% of them have locations recognized and disambiguated thanks to the context they provide each other. 48 hours later, among *the same tweets* posted in the first 30 minutes, additional 3% of tweets have locations recognized and disambiguated thanks to the additional context. As comparison, only 0.35% of tweets are geotagged in this dataset.

In addition to the locations associated to individual tweets, it is interesting to evaluate the location related to the event itself. As mentioned in Section 4, considering the cumulative number of the geotagged tweets only, the location related to the epicenter, Accumoli, is not highlighted with respect to, for example, Rome, which is a not damaged but much more populated location. If, instead, the locations referenced in the text are used (extracted as explained in [27]), the situation improves, as shown in the graph in Fig. 3. Indeed, there is an improvement in terms of the *volume* of locations, which are significantly (orders of magnitude) more, and in the *rate of growth* of the reports related to Accumoli with respect to those related to Rome, so that after about 3 hours Accumoli is highlighted as main location target of the event. It is interesting to notice that, as the time passes, the cumulative number of reports related to Accumoli become greater and greater, making the situation more and more clear. However, there exists a "window of uncertainty" in the first three hours where Accumoli does not emerge yet as main location.

The first tweets carrying useful information with images started to arrive early, even if the event occurred at night (see for instance Fig. 4). Several tweets contained the photograph shown in the figure, which was taken about two hours after the event. It has to be noted that the image was therefore available many hours before the EMS activation, which officially started more than 8 hours after the event).

It is interesting to analyze the tweets associated to the image in Fig. 4, to show the related challenges and their contribution to address time and space

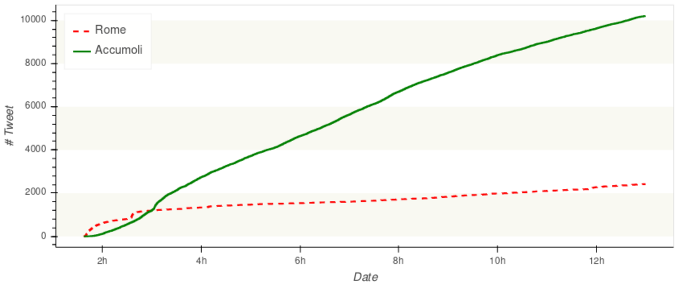


Fig. 3. Cumulative number of tweets about Rome and about Accumoli according to the text-extracted locations.

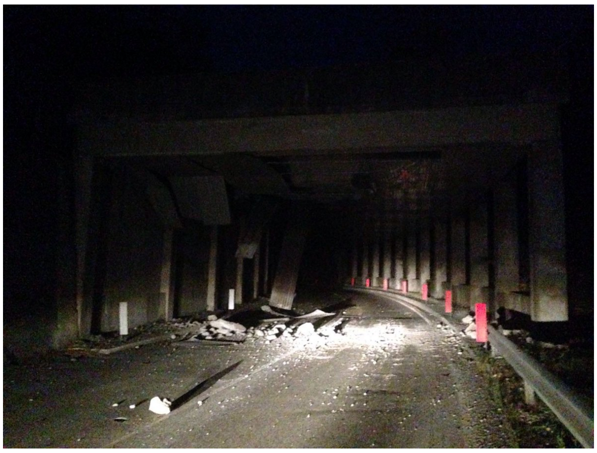


Fig. 4. Image extracted from tweets in the first hours after the event of August, 24.

imprecisions in the proposed model. They are shown, along with their timestamp and the profile location of the authors, in Table 6.1[3].

Analyzing them, several considerations arise. First of all, no tweet is geotagged, so that the positions of the authors submitting the tweets is unknown. Except tweets #3 and #4, they all have different authors. The locations associated to the authors profiles are not useful: they are missing or not real locations (MotoriNews24 & KeepRadio) or too coarse-grained (low precision, e.g., the province) and in any case incoherent. The tweet #1 is posted about two hours after the earthquake, describing the location generically as "the road to Nor-

[3] Notice that they are not the *complete* set of posted tweets with that image. Some tweets could not be retrieved and ≈ 2.7% of the retrieved tweets are not available anymore. Moreover, to find also the same image at different resolutions or with slight modifications a hashing algorithm has been used and it has false negatives.

cia". It is a very imprecise information: there exist many roads to Norcia each one traveling many kilometers. Therefore, one could simply take Norcia itself as location for the image — even if it is more than 275 km^2 — accepting a low precision and without being able to quantify the accuracy. Regarding the time associated to the image, this tweets gives an "upper bound" to its creation, so that the associated time is an interval of about two hours with high precision[4]. The tweet #2 cites Norcia again so it does not add significant information regarding the location or the time of the image. The tweet #3 is particularly interesting because it is posted just after half an hour the first tweet, but it better describes the location. Indeed, it cites the precise road ("Salaria SS4"), which is a road of more than 200 km, citing exactly the point ("at Sigillo"). There are several locations named "Sigillo" in Italy: the tweet cites other locations, and in particular "Posta", indicating that it refers to Sigillo part of the municipality of Posta, which has just 151 inhabitants, and not, for example, Sigillo municipality in the province of Perugia, which is more populated (population 2468). The approach described in [27] (see Section 3) is able to correctly recognize and disambiguate Sigillo, thanks to a spatial correlation found among the different locations mentioned in the tweet. The road SS4 at Sigillo is about 5 km, so using this tweet it is possible to associate a much better precision to the image, even if the accuracy is still unknown. This tweet does not add precision to the time associated to the image since it is posted later the first one. The following tweets do not add precision since they do not add details regarding the space or time of the image.

#	Time (UTC+2)	Text	Profile Location
1	5:41:37	#terremoto Crolli lungo strada per Norcia. Mandateci le vostre foto dall'Umbria a redazione@umbria24.it o sui social	Perugia
2	5:43:32	Strada per #Norcia crollo #terremoto	spello
3	6:11:57	Questa la Salaria SS4 all'altezza di Sigillo ... #terremoto #amatrice #Accumoli #Posta #ArquatadelTronto	MotoriNews24 & KeepRadio
4	6:37:57	RIPETO!!! FONDAMENTALE LASCIARE LIBERA LA SALARIA!!! #terremoto SE NON NECESSARIO NON USATE LA SALARIA SS4 !!!	MotoriNews24 & KeepRadio
5	6:45:26	Terremoto ecco la salaria all'altezza di sigillo #terremoto @SkyTG24 @RaiNews	
6	7:17:34	?????#Terremoto #Reatino invitiamo a non recarsi nelle aree terremotate per agevolare soccorsi #InfoAstral	Lazio Italia
7	7:25:05	La galleria sulla Salaria per Ascoli all'altezza di Sigillo #terremoto	

Table 1. All the retrieved tweets associated to the image in Fig. 4.

[4] This is true if it would be possible to confirm that the image comes from the target event; in general the starting point of the interval is unknown or, equivalently, the precision associated to the interval of two hours is low.

Concerning the images associated to the tweets, as illustrated in [11], only about 19% of the images extracted from geolocated tweets have been considered potentially useful, i.e., showing damages. As a document component we also considered videos. The analysis of videos showed that, among the around 1200 videos, very few of them were interesting for mapping purposes, as most of them were showing interiors. However, it has to be noted that some of these videos were very informative, showing areas from aerial images (mainly from drones) from official sources which reported also the exact location.

This analysis reinforces the need to support an exploratory search for data: the goal is not of retrieving a large number of images or videos, but of identifying the areas for which such images are needed for mapping, and in selecting those images which are more informative for the operators of the mapping activity. An analysis is ongoing for supporting the relevance ranking with statistical and context data: number of retweets, type of source (official sources, local sources), finding them through exploratory methods supported by automatic analysis of the sources and by following operators decisions (selection of an image for mapping activities, focus on a given area, and so on).

6.2 Event 2 - October 26, 2017

A second activation for the same area followed the first one, after a major aftershock. From this activation we have derived a negative result: only a very small (20) number of geolocated images where extracted, many of them not useful. The reasons for this result are probably due to the fact that at that moment the interested areas had been evacuated because of damages to buildings. It has to be noted also that even if another major aftershock was recorded on October 30, no new EMS activation was started. This fact supports the fact that in this special case the need for rapid mapping in already damaged areas becomes less urgent, while other types of maps need to be produced (assessment maps which are produced after the events), which are not the subject of this study.

7 Concluding remarks and future work

The need for facilitating the retrieval of useful information from large amounts of data has been discussed in the paper, focusing on spatial and temporal information extracted from Tweets. As discussed in the paper, while there is a wide margin for improvement of the automatic geolocation of information, the focus of the discussion is about exploiting already available information, which might evolve over time, applying an exploratory approach to data analysis to find useful information for the tasks to be performed.

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 Paolo Ravanelli for his support in creating event-specific crawlers, and Paolo Gugliemino and Matteo Montalcini for their work on the case study.

References

1. Al-Rfou, R., Kulkarni, V., Perozzi, B., Skiena, S.: Polyglot-NER: Massive multilingual named entity recognition. Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada, April 30 - May 2, 2015 (April 2015)
2. Andrienko, G.L., Andrienko, N.V., Fuchs, G.: Understanding movement data quality. J. Location Based Services 10(1), 31–46 (2016), `http://dx.doi.org/10.1080/17489725.2016.1169322`
3. Andrienko, N., Andrienko, G., Fuchs, G., Rinzivillo, S., Betz, H.D.: Detection, tracking, and visualization of spatial event clusters for real time monitoring. In: Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on. pp. 1–10. IEEE (2015)
4. Batini, C., Scannapieco, M.: Data and Information Quality - Dimensions, Principles and Techniques. Data-Centric Systems and Applications, Springer (2016)
5. Becker, H., Naaman, M., Gravano, L.: Beyond trending topics: Real-world event identification on Twitter. ICWSM 11(2011), 438–441 (2011)
6. Brusoni, V., Console, L., Terenziani, P., Pernici, B.: Qualitative and quantitative temporal constraints and relational databases: Theory, architecture, and applications. IEEE Trans. Knowl. Data Eng. 11(6), 948–968 (1999), `http://dx.doi.org/10.1109/69.824613`
7. Castillo, C.: Big crisis data: Social media in disasters and time-critical situations (2016)
8. Davis Jr, C.A., Pappa, G.L., de Oliveira, D.R.R., de L Arcanjo, F.: Inferring the location of Twitter messages based on user relationships. Transactions in GIS 15(6), 735–751 (2011)
9. Di Blas, N., Mazuran, M., Paolini, P., Quintarelli, E., Tanca, L.: Exploratory computing: a comprehensive approach to data sensemaking. International Journal of Data Science and Analytics 3(1), 61–77 (2017), `http://dx.doi.org/10.1007/s41060-016-0039-5`
10. Dyreson, C., Grandi, F., Käfer, W., Kline, N., Lorentzos, N., Mitsopoulos, Y., Montanari, A., Nonen, D., Peressi, E., Pernici, B., et al.: A consensus glossary of temporal database concepts. ACM Sigmod Record 23(1), 52–64 (1994)
11. Francalanci, C., Guglielmino, P., Montalcini, M., Scalia, G., Pernici, B.: IMEXT: a method and system to extract geolocated images from tweets analysis of a case study. In: Proc. RCIS'17, Brighton, UK (May 2017)
12. Francalanci, C., Pernici, B.: Data integration and quality requirements in emergency services. In: Advances in Service-Oriented and Cloud Computing. Springer (in press)
13. Francalanci, C., Guglielmino, P., Montalcini, M., Scalia, G., Pernici, B.: Imext: a method and system to extract geolocated images from tweets analysis of a case study. In: Research Challenges in Information Science (RCIS), 2017 IEEE Eleventh International Conference on Research Challenges in Information Science. IEEE (2017)

14. Gelernter, J., Mushegian, N.: Geo-parsing messages from microtext. Transactions in GIS 15(6), 753–773 (2011)
15. Ghufran, M., Quercini, G., Bennacer, N.: Toponym disambiguation in online social network profiles. In: Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems. p. 6. ACM (2015)
16. Guglielmino, P., Montalcini, M.: Extracting relevant content from social media for emergency management contexts, Master Thesis, Politecnico di Milano (2016)
17. Imran, M., Castillo, C., Diaz, F., Vieweg, S.: Processing social media messages in mass emergency: A survey. ACM Computing Surveys (CSUR) 47(4), 67 (2015)
18. Imran, M., Elbassuoni, S.M., Castillo, C., Diaz, F., Meier, P.: Extracting information nuggets from disaster-related messages in social media. Proc. of ISCRAM, Baden-Baden, Germany (2013)
19. Imran, M., Mitra, P., Castillo, C.: Twitter as a lifeline: Human-annotated Twitter corpora for NLP of crisis-related messages. arXiv preprint arXiv:1605.05894 (2016)
20. Inkpen, D., Liu, J., Farzindar, A., Kazemi, F., Ghazi, D.: Detecting and disambiguating locations mentioned in twitter messages. In: International Conference on Intelligent Text Processing and Computational Linguistics. pp. 321–332. Springer (2015)
21. Liu, F., Vasardani, M., Baldwin, T.: Automatic identification of locative expressions from social media text: A comparative analysis. In: Proceedings of the 4th International Workshop on Location and the Web. pp. 9–16. ACM (2014)
22. Nugroho, R., Yang, J., Zhao, W., Paris, C., Nepal, S.: What and with whom? identifying topics in Twitter through both interactions and text. IEEE Transactions on Services Computing PP(99), 1–1 (2017)
23. Paradesi, S.M.: Geotagging tweets using their content. In: FLAIRS conference (2011)
24. Reuter, T., Cimiano, P.: Event-based classification of social media streams. In: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval. p. 22. ACM (2012)
25. Ritter, A., Clark, S., Etzioni, O., et al.: Named entity recognition in tweets: an experimental study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1524–1534. Association for Computational Linguistics (2011)
26. Sakaki, T., Okazaki, M., Matsuo, Y.: Tweet analysis for real-time event detection and earthquake reporting system development. IEEE Trans. Knowl. Data Eng. 25(4), 919–931 (2013), `http://dx.doi.org/10.1109/TKDE.2012.29`
27. Scalia, G.: Network-based content geolocation on social media for emergency management, Master Thesis, Politecnico di Milano (Apr 2017)
28. Tamura, K., Ichimura, T.: Density-based spatiotemporal clustering algorithm for extracting bursty areas from georeferenced documents. In: Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on. pp. 2079–2084. IEEE (2013)
29. Wasay, A., Athanassoulis, M., Idreos, S.: Queriosity: Automated data exploration. In: Carminati, B., Khan, L. (eds.) 2015 IEEE International Congress on Big Data, New York City, NY, USA, June 27 - July 2, 2015. pp. 716–719. IEEE (2015), `http://dx.doi.org/10.1109/BigDataCongress.2015.116`
30. Zhang, W., Gelernter, J.: Geocoding location expressions in Twitter messages: A preference learning method. Journal of Spatial Information Science 2014(9), 37–70 (2014)