CrisisTracker: Crowdsourced Social Media Curation for Disaster Awareness

Authors
Jakob Rogstadius, Maja Vukovic, Claudio Teixeira, Vassilis Kostakos, Evangelos Karapanos, Jim Laredo

Abstract
Victims, volunteers and relief organizations are increasingly using social media to report and act upon large-scale events, as witnessed in the extensive coverage of the 2010-2012 Arab Spring uprisings and 2011 Japanese tsunami and nuclear disasters. Twitter feeds consist of short messages, often in nonstandard local language, requiring novel techniques to extract relevant situation awareness data. Existing approaches to mining social media are aimed at searching for specific information, or identifying aggregate trends, rather than providing narratives.

We present CrisisTracker, an online system that in real-time efficiently captures distributed situation awareness reports based on social media activity during large-scale events, such as natural disasters. CrisisTracker automatically tracks sets of keywords on Twitter, and constructs stories by clustering related tweets based on their lexical similarity. It integrates crowdsourcing techniques enabling users to verify and analyze stories. We report our experiences from an eight-day CrisisTracker pilot deployment during 2012 focused on the Syrian civil war, processing on average 446000 tweets daily and reducing them to consumable stories through analytics and crowdsourcing. We discuss CrisisTracker’s effectiveness based on the usage and feedback from 48 domain experts and volunteer curators.
Introduction

Social media refers to on-line communication between users, often in the form of blogs, content communities, and social networking sites [1], and is becoming a prevalent mechanism of interaction [2]. Advances in social media and the adoption of mobile devices are transforming how we experience and share news.

Twitter [3] is a microblogging service that enables its users to send and receive short messages, called tweets, of up to 140 characters through Web, instant messaging or SMS interfaces. Over 140 million people use Twitter [4] as part of socializing, activism, and other non-crisis activities [5]. Characteristic to Twitter is a high level of redundancy of the content, both due to users who re-share and optionally modify posts authored by others (called re-tweeting) and to users who independently report the same event. Duplication drives how information spreads through the service and previous research has indicated that 29-47% of crisis-related tweets are retweets and that 60-75% of all tweets are near-duplicates of at least one other tweet [6].

Victims, responders and volunteers increasingly use Twitter to share and access situation awareness reports and to provide help during a wide range of disasters, such as wildfires and floods [7,8,9,10], accidents [11] and terrorist attacks [12]. Furthermore, relief organizations use Twitter to provide emergency contact and recruit volunteers [13]. While Twitter can suffer from strong sample bias on a local scale, larger-scale patterns can accurately match those obtained through traditional data collection methods [14,15].

Twitter is a promising channel to explore due to availability of open APIs that give access to almost all the communication in real time. Unlike traditional communication technologies such as cell phones, Twitter can be monitored and members of the disaster-affected population can be employed as a sensor network. Even in cases where data connectivity is limited to a subset of the affected population, important information spreads locally by word of mouth. Once the information reaches those able to share it online, it can in theory be sensed remotely, as was observed for instance during the devastating Haiti earthquake in 2009 [16].

Despite its popularity during crises, Twitter is very challenging to monitor during large-scale disasters due to very high message volumes and lack of essential meta-data. Although geotagging is supported, only around 1% of tweets are in practice geotagged and tags refer to the user rather than the message subject. Other disaster-related meta-data such as higher-level report categories and named entities are unavailable. The 140 characters are also insufficient for most algorithms that build latent topic models, and the use of local languages with relaxed rules for spelling and grammar reduces performance of standard automation techniques such as named entity extraction. Finally, the data is streaming, and therefore the corpus of data relevant to any single story is constantly changing, which many established techniques for text mining were not designed for. We propose to overcome these challenges by combining scalable automated techniques for event detection with crowdsourced human-based computation for further interpretation [17].

Human-based computation is a technique in which a computational process performs its function by outsourcing certain steps to humans. This is often done in a crowdsourced manner, where the
computation is done by a distributed group of people. The technique is practical for handling computational problems that are easy to solve for humans, but challenging for computers, such as image labeling [18], knowledge management [19] and solving business problems [20]. The performance of crowds largely depends on incentives [21] such as money [22], game mechanics [18,23], social capital [24] and public good [19]. Intrinsically motivated volunteers have been shown to be more likely than paid crowds to produce high-quality results [20]. With computationally challenging data and availability of motivated crowds crowed sourced human computation becomes an attractive approach for disaster information management. Yet existing solutions offer insufficient support for dedicated crowds to cope with the torrent of information from social media during mass disasters [25].

In this paper we present CrisisTracker [6], a Web-based platform that integrates automated real-time analysis with crowdsourcing, to annotate rapid streams of unstructured tweets with metadata and group related content, with the goal of increasing situational awareness during disaster. While previous work has relied on crowdsourcing (e.g. via SMS) [26], or automated analysis [27,28,29], little prior work has considered integrating the two. CrisisTracker aims to support the first two levels in Endsley's model of situation awareness [30], according to which fundamental perception of events is improved by event detection and ranking, and comprehension is improved by relating events to each other.

In the next section we put CrisisTracker in the context of existing social media applications for disaster awareness. We then describe the system architecture of CrisisTracker and present the deployment of CrisisTracker during the 2012 Syrian civil war, demonstrating the system’s effectiveness and usability.

Related Work

Social media is used throughout the emergency management cycle to detect potential hazards, gain situation awareness, engage and mobilize local and government organizations and to engage volunteers at the disaster recovery stage. Users of social media at disaster time include victims, volunteers, and relief agencies. Existing systems, compared in Table 1, can be loosely grouped into disaster management [31,32], crowd-enabled reporting [26] and automated information extraction [27,28,29,33].

Sahana [31] and VirtualAgility OPS Center (VOC) [32] support the emergency disaster management process with information and inventory management and collaboration support for response organizations. Such systems often integrate raw social media feeds, but lack capabilities for distilling useful reports, and reducing information overload when activity is exceptionally high.

Crowd-reporting systems like Ushahidi [26] enable curation and geo-visualization of manually submitted reports from a wide range of sources. Due to reliance on users in all information-processing stages, Ushahidi’s effectiveness depends entirely on the size, coordination and motivation of crowds. The majority of the most successful deployments have been by the volunteer-based Standby Task Force (SBTF) [34], which has set up dedicated teams for media monitoring, translation, verification, and geolocation. This team structure is further supported by
task management extensions to the platform [35] and adapts well to needs of specific disasters, but is difficult to scale to match information inflow rates during the largest events [24].

TweetTracker [27] is a system that parses a Twitter feed to extract and rank popular hashtags, user mentions and URLs and provides time filters, a map for geotagged tweets, and a word cloud of popular terms. Yin et al. [28] went further by developing a system for use in emergencies that detects trending words and phrases within a Twitter stream to enable drill-down to increasingly detailed content through a series of word clouds. The system also incorporates pre-trained language-specific classifiers to detect messages containing specific information (e.g. reports infrastructure damage) and a map for geotagged tweets. Similarly, Twitcident [29] is a Twitter filtering and analysis system that improves situation awareness during small-scale crisis response, such as music festivals and factory fires. It gathers geotagged tweets only and employs classification algorithms to extract messages about very specific events.

Despite extensive research into automated classifiers for short contextual strings, classification and information extraction has proven to be significantly harder than for well-formed news articles and blog posts. As in [28] and [29], classifiers tend to be language specific and new training data is needed for each new desired label. This restricts their use in the mass disaster space, where report language is not known beforehand and unsupported report types or distinctions between subtypes may be sought in new disasters. For instance, reports of violence and looting may be grouped into one category in a post-earthquake setting, while several violence-related report categories may be of interest while mapping a civil war.

EMM NewsBrief [33,36] automatically mines and clusters mainstream news media from predetermined sources in a wide range of languages, with new summaries produced every ten minutes. It too relies on rule-based classifiers for meta-data, but substantial investment has been made to create such rules over a decade. Despite this great investment, it has not been extended to handle social media.

To our knowledge no currently available system enables any of the mentioned communities to use timely social media as a structured information source during mass disasters. CrisisTracker’s approach to accomplish this is by combining language-independent fast and scalable algorithms for data collection and event detection, with accurate and adaptable crowd curation. Rather than displaying only raw tweets and high-level statistical metrics (e.g. word clouds and line graphs), the system also provides an intermediary level of natural language narratives that retain details within reports. CrisisTracker is intended for use during mass disaster and conflict when responders or observers lack resources to sufficiently monitor events on the ground, or when physical access to local communities is for some reason restricted.

**System Architecture**

This section describes CrisisTracker’s information processing pipeline (Figure 1), consisting of data collection, story detection, crowd curation and information consumption. Crowd curation is made possible by decoupling the information itself (stories) from how it has been shared in the social network (tweets). CrisisTracker is free and open source and available at https://github.com/JakobRogstadius/CrisisTracker.
Data Collection

Tweets are collected through Twitter’s streaming API. This allows a system administrator to define filters in the form of words, geographic bounding boxes and user accounts for which all matching new tweets will be returned as a stream. Twitter’s API returns messages tagged with any geographic region that partially overlaps with specified bounding boxes. As regions can be very large, we perform an additional filtering pass to discard such messages and keep only those explicitly tagged with a coordinate within bounding boxes of interest (assuming no other filter is matched). The tracking filters are constrained by the API and the system cannot yet suggest keywords or user accounts to track. Generally around 1% of all tweets are geotagged and it is infeasible to manually select user accounts to cover truly large-scale events, thus good keyword filters are the primary way to obtain high information recall in the system.

CrisisTracker also discards messages having fewer than two words after stop word removal and a very low sum of global word weights (approximated inverse document frequencies). In practice discarded messages are mostly limited to short geotagged messages without any context (e.g.”@username Thanks!”).

Depending on the filters used to collect tweets, the sample will be more or less focused on actionable reports from a particular geographic region. CrisisTracker does not yet integrate algorithms to classify individual messages’ content or to identify sources local to the event. Instead, the system improves signal through ranking and filtering of stories, as explained below.

Story Detection using Locality Sensitive Hashing

Incoming tweets are compared to previously collected tweets using a bag-of-words approach, meaning that the textual content of tweets is treated as a weighted set of unsorted words and that tweets are more similar the more words they have in common. A cosine similarity metric is then used to group together messages that are highly similar. Clustering is performed using an extended version of an algorithm previously applied to Twitter corpuses [37] based on Locality Sensitive Hashing [38], a probabilistic technique using hash functions that quickly detects near-duplicates in a stream of feature vectors. The time range in which duplicates can be detected for any incoming message depends on the rate at which similar messages have been received, and we refer to [37] for details.

Petrovic et al. [37] used an initial computation pass to calculate global word statistics (inverse document frequencies) in their offline corpus. In an online setting, word frequencies cannot be assumed to be constant over time, e.g. due to local changes in the tracked event and global activity in different time zones. The algorithm was therefore extended in two important ways. First, word statistics are collected based on both the filtered stream and Twitter’s sample stream, a 1% sample of all posted tweets. The word distributions are then approximated with a simple IIR filter with exponential decay and a four-day half-time. Words that have a global frequency greater than 90% of the maximum frequency are labeled as stop words unless they are tracking keywords. Words that have been seen less than three times are ignored. The second extension is to replace the oldest hash function hourly with a new function created from the current dictionary. The new hash table is then populated with the items from the removed table.
Furthermore, the original thread-based clustering algorithm offers high precision [37] such that clusters typically contain only highly similar tweets (precision). However, after implementation we found that the algorithm’s recall was relatively low, such that the set of tweets that discuss a particular event was often split across several clusters. In CrisisTracker, all new clusters are therefore compared with the current clusters to check for overlap. We refer to such a cluster of clusters as a story and as the next section explains, this method also enables human intervention in the clustering process. Initial informal evaluation suggests that our approach greatly improves recall without substantially reducing precision, but accurate measurement of cluster recall for Twitter-scale corpuses is a research problem in itself. We therefore leave the specific cluster evaluation for future work and instead focus in this paper on evaluating CrisisTracker’s general ability to improve situation awareness and support decision making.

A limitation of any bag-of-words based event detection technique is that clusters do not necessarily correspond to events, as tweets can potentially have high textual similarity and be grouped together without discussing the same event. We have observed this with sensor-based feeds that publish regular updates about weather and earthquakes. Additionally, as the system would quickly run out of storage space if all content was kept, increasingly larger stories and all their content are deleted with increasing age, unless they have been tagged by a human. Stories consisting of a single tweet are kept for approximately one day.

**Crowd Curation and Meta-Data Creation**

The purpose of clustering the tweet stream into stories is to facilitate crowd curation. De-duplication (ideally) eliminates redundant work, directly reduces the number of items to process per time unit, enables size-based ranking of stories, and groups together reports that mention the same event but contain different details necessary for piecing together a complete narrative.

Search and filtering requires meta-data for stories. Some of this meta-data is extracted automatically, i.e. time of the event (timestamp of first tweet), keywords, popular versions of the report, and number of unique users who mention the story (it’s “size”). Story size enables CrisisTracker to estimate how important the message is to the community that has shared it [8,25]. Users of the system can rank stories by their size among all Twitter users, or among the 5000 users most frequently tweeting about the disaster. In our experience, the top 5000 option better brings out stories with detailed incremental updates to the situation, while the full rank more frequently includes summary articles, jokes and opinions. Since meta-data is assigned per-story, it also covers future tweets in the same story.

Curators are directed towards recent and extensively shared stories, but can self-select which stories to work on. The first curation step is to further improve the clustering, by optionally merging the story with possible duplicate stories that are textually similar but fall below the threshold for automated merging. Miss-classified content can also be removed from stories, which are then annotated (Figure 2) with geographic location, deployment-specific report categories (e.g. infrastructure damage or violence) and named entities. Stories deemed irrelevant (e.g. a recipe named after a location) can be hidden, which prevents them from showing up in search results. Only a Twitter account is required to volunteer as a curator and as demonstrated by the Standby Task Force, curators can be recruited globally on a volunteer basis, or locally if post-disaster infrastructure permits.
Information Consumption

Disaster responders and others interested in the information can filter stories by time, location, report category and named entities. An information management officer at UN-OCHA explained in a private conversation that these are the basic dimensions along which information is structured in the disaster space, and that they match how responsibilities are typically assigned within the responder command structure, i.e. by location and/or type of event or intervention. Figure 2 presents the interfaces for exploring stories and for reading and curating a single story. The interface for curators to select work items is not shown.

CrisisTracker Field Trial

To evaluate how CrisisTracker supports its users, a field trial was conducted focusing on the 2012 civil war in Syria. Syria has a population of 21 million, high Twitter adoption and tweets in both English and Arabic, thus the conflict includes many of the characteristic challenges of social media information management during international crises. The evaluation had two main purposes: to measure curators’ processing capacity and to understand how real-time access to structured social media reports can support analysts and decision makers in disaster management organizations.

We recruited 44 unpaid volunteer curators to participate in the eight-day field trial. Half participated for a single day only, three participated every day and on average 13 unique participants were active each day. Volunteers were recruited through the Standby Task Force or independently. Approximately two thirds had previous experience from online disaster volunteering and one third regularly participated in onsite disaster response. Curators were given detailed instructions on how to use the system to organize and annotate stories, and could interact with each other and with the researchers through a dedicated persistent text chat.

At the end of the trial, 22 volunteer curators responded to an open-ended questionnaire regarding their experiences with the platform. Further semi-structured interviews were held with five of these curators. Interview questions followed up on the free-text answers from the questionnaire and focused on usability, workflows, motivation, perceived value of the work, and platform extensions.

Seven domain experts with extensive experience from the disaster management domain were also recruited to explore the curated stories in CrisisTracker during the field trial week and assess the value of the system. Their expertise included disaster management at different levels, GIS analysis, information management and media monitoring. Two of the experts were actively following the Syrian conflict beyond the scope of our study and three experts also participated as volunteer curators. Semi-structured interviews were held with all seven experts with questions relating to 1) their past use of social media reports during crisis, 2) their knowledge of the Syrian conflict, 3) their experiences using CrisisTracker, 4) potential applications of CrisisTracker during past events, and 5) its ability to support different users and use cases. Transcripts from interviews, plus chat logs, survey answers, emails and other communication were content analyzed through bottom-up coding, clustering and interpretation by one researcher [39].
To collect tweets, we defined a bounding box covering Syria and together with an analyst familiar with the conflict we selected 50 keywords (syria, assad, hama, damascus, aleppo, homs, bashar, #fsa, daraa, #siria, #syriarevo, #syrie, #syrian, #syrias, idlib, #realsyria, lattakia, houla, latakia, rastan, deraa, #yr, hamah, tartus, harasta, zabadanî, darayya, baniyas, babaamr, darayaa, raqqah, yayladagi, basharcrimes, suweida, latakiah, #دمشق, #الجيش_السوري_الحر, #الجيش_الحر, جيش_تويتر, #حلب, #حماه, #الجيش_السوري_الحر, #بشار, #اللاذقية, #الثورة_السورية, #سوريا_تتحرر, #إدلب, #طرطوس, #سوريا, معركة_دمشق_الكبرى) in English and Arabic to track. The words referred mainly to major place names and popular hashtags. We also worked with the same analyst to define 14 relevant report categories to be applied by the human curators, e.g. “demonstration”, “eyewitness report” and “people movement”. About 3.5 million tweets were collected during the evaluation week, but the system had been up and running for several months before the volunteers began working and older stories could be retrieved through searches.

Only a few participants spoke Arabic and curators were encouraged to use machine-translation features built-in to some Web browsers to read stories and linked articles, and add meta-data based on the translated content. The platform supports inclusion of manually translated summaries to stories, which native language speakers in some cases used. An option to hide Arabic stories also existed for users who found the Arabic content too tiring or difficult to work with.

Results

Clustering Reduces Workload

Clustering the incoming tweets into unique stories greatly reduced the rate at which new items needed to be processed. The system collected on average 446000 tweets per day (min 400783, max 560193). 70% of tweets were immediately discarded, almost exclusively because the geographic bounding box covering Syria overlapped with parts of Turkey, causing geotagged tweets from all over Turkey to be returned by the API. The remaining messages were then clustered into about 33000 daily stories, of which 1200 stories contained tweets from at least five users and 246 stories from at least 50 users.

Although size-based story ranking has been shown to generally improve signal-to-noise ratio [6,8], high levels of spam were initially found among the top stories. These spam stories mainly originated from a group of 58 highly active spammer accounts. Once these accounts were blocked, the number of daily stories shared by at least 50 users dropped to 141. Focusing on these top stories thus reduced the workload by three orders of magnitude, with only a handful of new top stories per hour. Figure 3a summarizes how each processing step contributes to making the workload manageable.

Scalability of Crowd Curation

On its own CrisisTracker collects and clusters tweets into ranked stories, and supports search and filtering based on time and keywords. Human curation is required for location-, category- and named entity-based filtering, which primarily helps find less impactful (and thus less mentioned) events. Curators spent on average 4.5 minutes per story, with a heavy skew towards shorter times (median 2.3 minutes). Each work session lasted on average 28.5 minutes (median 20.8), with
work sessions defined as periods of user activity separated by at least 15 minutes of idle time. A total of 3600 tags were added to 820 stories (1775 before merging), and together the curated stories contained 616009 tweets.

We consider this total volunteer output substantial when compared to most Ushahidi deployments. Data provided by CrowdGlobe [40] shows that of 871 Ushahidi [41] instances that were ever active (containing ten or more reports), only 67 (7.7%) contained over 820 reports. Furthermore, 14 of 15 public instances with 500 to 1100 reports received their reports over more than two months, compared to our eight-day effort. As curation tasks are similar in the two platforms, we mainly attribute the higher productivity of CrisisTracker curators to automated information gathering.

As is common in crowdsourcing, the work effort was unevenly distributed among the participants and 25% of the curators did 75% of the work. Among the 22 people who curated at least 10 stories each, the average time spent per person per story ranged between 1 and 13 minutes. Only in a handful of cases did curators work on the same stories or remove others’ tags.

Based on the reported usage statistics and the rates at which information was collected, we estimate that about 15 curators each active for 30 minutes per day would be enough to have full meta-data for all the main events during a humanitarian crisis of this type and magnitude. Approximately 150 curators would be able to build and maintain a very detailed database of almost all reported events, below city-level resolution. Workforce size can be greatly reduced if curators can be encouraged to spend more than 30 minutes per day. Feedback from volunteers also suggests that per-curator effort would have been higher if the deployment had been a request from a humanitarian organization, rather than an academic study.

**Real-Time Overview**

One of the domain expert participants, an anonymous English and Arabic speaking representative of Syria Tracker [42], was actively monitoring ongoing events in Syria in parallel with using CrisisTracker. Syria Tracker had been monitoring the crisis for 18 months and had set up infrastructure to automatically mine Arabic news media. They also received daily eyewitness reports and manually monitored online social media. This participant was therefore in a unique position where he could in real time compare the information made available through CrisisTracker with that independently collected.

Twitter was from the start of this conflict an active communication channel used mainly by the opposition, and more recently also by government supporters. Links to relevant images, videos and news articles are also often posted on Twitter almost immediately following publication of the resource. Syria Tracker was well aware of this rich source, but had not found any way to reliably monitor it in real-time. After using the system, the expert explained that CrisisTracker was the first of many tools they had tried that provided a subjective sense of real-time overview of this “information storm”, in both English and Arabic.

**Timely and Rich Reports**

One of the concrete ways in which Syria Tracker used the system was to take trusted but brief and single-sided eyewitness reports submitted via email and in CrisisTracker quickly seek out
complementary pictures and videos as well as reports from the opposing side. This augmentation of eyewitness reports enriched their understanding of ongoing events hours and sometimes even days before mainstream media picked up the news.

According to Syria Tracker, the tool improved sensitivity of information collection by helping them detect several important events (e.g. massacres, explosions and gunfire) before they were reported by other sources. CrisisTracker also improved specificity, by quickly picking up links to evidence such as images and videos. Early detection of events enabled focused manual monitoring of other sources to corroborate evidence and make early assessments of the truthfulness or severity of new claims. For instance, videos of missiles being fired with claims regarding time and place led to searches for corresponding impacts. During ongoing events such as urban skirmishes, CrisisTracker would also provide real-time updates whereas eye-witness reports would arrive as high-quality summaries later during the day. Syria Tracker also valued the historical record of social media communication provided by the tool, as Twitter does not support searches more than a few days back in time.

Six of the seven interviewed experts in some way mentioned timeliness as a primary reason for using the system. An incident commander who actively used the platform over several days said “I feel very confidently that those reports will come out ahead of CNN and BBC and that they will have the central nuggets of who, what, when, where, why. For an incident commander, it is the difference between learning something in 2-3 hours versus learning it in 6-8”. A data analyst with a background in disaster response further added that the timeliness and being able to see how stories are evolving is “huge”.

Figure 3b shows how quickly stories of different magnitude would be detected when monitoring all stories above a size threshold. Of 9207 stories each eventually containing tweets from 150 or more users, 10% would be detected within four minutes, 50% within 31 minutes, and 90% within 3.5 hours of the first report, with a size threshold set at 50 users (141 stories per hour).

On its own, Twitter is in fact so timely that some participants described how they had felt in the past that “if you’re not on the Twitter stream you quickly lose sight of the history. It becomes very much a snapshot, a moment-in-time assessment of what’s happening”. Since CrisisTracker keeps a historical archive of all larger stories as the main event progresses, it becomes possible to go back in time and analyze both short-term and long-term trends.

A GIS expert who was also one of the most active curators described how seeing real-time commentary on what was happening in Moscow, next to reports of events on the ground and how many people were killed gave her a sense of how the whole world was connected. She further described how “you can see over a period of time where people are moving, how that relates to conflict areas. Water shortage, or food, you can almost anticipate where needs are going to be.” As discussed later, reaching this level of understanding requires a significant time investment. However, we consider it promising that such insights can at all be gained from data largely derived from automated processing of freely available social media.
Discussion

Usage Barriers

Arabic content proved frustrating for many non-Arabic speakers, who simply hid it and focused on stories in English. Others were concerned about the quality of machine translations, but it was noted that “obviously some reports seem just fine with little room for error, while others are just unusable.” Another volunteer said that “using Google Translate was very easy and by being able to curate stories, especially those in Arabic, I felt more of a ‘direct’ connection with what was going on.” One of the greatest obstacles interestingly proved to be geotagging of English reports, as translated location names often could not be found on the map or through search. Overall, our impression is that machine translation is tiring but relatively safe to use for curators, as the impact of mistranslation is limited to tagging errors that reduce the quality of search and filtering.

Many experts and volunteers were impressed with the platform’s overall intuitiveness and ease of use. A few users explicitly compared CrisisTracker with their experience using the Ushahidi platform, which they considered to be more tedious to work with and less intuitive. Participants with experience from both platforms particularly valued CrisisTracker’s ability to relate stories to each other and its method of handling duplicate reports. However, the evaluation also identified several minor user interface issues that lead to breakdowns in user interaction and which should be resolved before the platform can be considered mature enough for public use.

In rapid onset disasters, ease of deployment (effectively deployment time) is a critical aspect of an information management system. While anyone can download the CrisisTracker source code and set up an instance, further work is needed to simplify and automate the deployment process. Ideally this would be done through a system similar to the Crowdmap website [34], where new instances of the Ushahidi platform can be deployed on cloud servers through a web interface.

Many participants raised concerns that the system’s complete openness combined with lack of “undo” functionality leads to high sensitivity to vandalism, in the form of purposeful incorrect tagging, merging or hiding of stories. This is particularly an issue during conflict settings and extensions need to be made to handle this, either by adding mechanisms for screening of curators, or by implementing a version control and community moderation system similar to those on Open Street Map [43] and Wikipedia [44].

Using CrisisTracker to Support Decision Making

Our evaluation suggests that the greatest value of CrisisTracker during complex and constantly changing large-scale events is improved real-time situation awareness. This result is very similar to an evaluation study [45] of the Ushahidi deployment for the 2010 Haiti earthquake, which showed that improved situation awareness at an aggregate level was the greatest contribution of the crowdsourced map, in particular during the early days of the crisis when the situation on the ground was still unclear. Content analysis of tweets collected during natural disasters [46] indicates great availability of response-phase related reports: hazards, interventions, fatalities, personal status and damage. This agrees with our general perception of the content distribution for the Syrian conflict. Though one disaster manager participant in our study speculated that
CrisisTracker would be of great value also in the recovery phase, the content analysis indicates that recovery-related tweets are scarce.

CrisisTracker also addresses two important limitations identified by the Ushahidi evaluation. First, CrisisTracker taps into existing social media to access the voices of affected populations, rather than relying on independently submitted or manually collected reports. Second, while the vast majority of reports go uncurated in both CrisisTracker and Ushahidi, CrisisTracker is able to direct curation efforts towards those reports that are discussed most in the core community and thus most likely to improve situation awareness.

Despite general praise for CrisisTracker’s good usability, intuitiveness and ability to extract rich and timely information of important events, many participants made clear that gaining an accurate understanding of the information remains a time-consuming task. A high-level manager in a humanitarian organization explained that while much useful information is collected by the system, key points need to be distilled from the stories to make that information useful in time-pressed situations.

The system is therefore not yet ready to be used directly as a decision support tool by decision makers who have very limited time to sit down, read and analyze information. Rather, CrisisTracker is suitable for use by analysts and others who already work on filtering and aggregating information from different sources to produce maps and reports tailored for the organization’s decision makers. The analysts we interviewed were very enthusiastic about the tool and only requested export functionality to be implemented, to enable comparison of social media reports with data from other sources.

Several participants, both curators and experts, expressed that they wanted CrisisTracker to help them assess the trustworthiness of stories in the platform. Discussions focused on that assessment should be done on a source level, for instance by inferring how credible a first report is based on the past record of that account, or by highlighting the most trustworthy account that has shared a story. Others noted that different sources have different authority for different information, which greatly adds complexity to such calculations. For instance, a personal account with strong political bias may contribute little to international news, but still be authoritative regarding events in the particular suburb where they live. In other cases an untrusted source may post a link to a highly trusted government website, or a trusted government account may post outdated information that is contradicted by citizen generated video footage. Potentially feasible approaches include letting curators mark single accounts as having extremely high or low credibility, and deriving source ratings from sharing patterns of past stories.

One participant noted that human curation itself may be misinterpreted as validation, which if true would be a serious risk in decision making and could be an entry barrier for volunteer curators. This is something that future development will need to keep in mind and find ways to avoid.

Several of the more experienced disaster managers we interviewed, or have spoken to at other times, have pointed out how accuracy and timeliness will always remain a tradeoff. Verification requires both time and expertise, and a system that can deliver very quick reports in a
consumable format does not need to always be correct to be valuable. One disaster management consultant estimated that even her trusted personal sources that she uses for verification may only be right in 80% of the cases. Several participants also reported that they felt the system’s grouping of complementary reports into stories helped build reliability and supported validation. We strongly believe that the rich and timely but unverified reports in CrisisTracker are most valuable when combined with other sources, for instance to enrich brief but trusted eyewitness reports as demonstrated by Syria Tracker. We also believe the system is capable of providing rich historic narratives around specific points in time and space that can help explain interesting features in other datasets.

Managing a CrisisTracker Deployment

When integrating crowdsourced human-based computation into disaster information management, leadership has a direct impact on performance metrics such as precision, recall and processing speed. Based on experiences from our evaluation, we propose that humanitarian organizations that want to deploy this class of systems assign a person the new role of crowd director.

First, the crowd director is responsible for recruiting curators, e.g. the organization’s own registered disaster volunteers, or through an independent volunteer organization such as the Standby Task Force. As volunteering competes with other tasks in people’s lives, it is of great importance during recruitment to motivate potential curators by transparently explaining how the work will create benefit and help a population in need. Volunteers will also get up to speed quicker if they receive basic background information regarding the disaster. Basic training materials regarding how to use CrisisTracker are already in place.

Once work begins, the crowd director must continuously communicate decision makers’ information requirements to the crowd. This will enable the organization to steer the crowd towards work that is of particularly high value. For instance, instructions can be to focus on reports relating to a particular town or report category, or to spend more or less time on tracking down precise locations. Unlike information management systems that are completely automated, interactive crowd management enables CrisisTracker to effectively function differently depending on the particular needs of each deployment. Rather than having to fine tune or develop new processing algorithms for each new use case and content language, organizations can themselves maximize performance by giving direction and recruiting volunteers with relevant skills and experience.

Finally, insecure or inexperienced volunteers will ask for affirmation that their work is correctly carried out. The crowd director must remain accessible to provide such feedback, which can be the difference between having a volunteer who only curates a single story and then stops, and one who confidently comes back day after day during most of their spare time. Some intervention may also be required to improve accuracy of the assigned labels, by prompting volunteers who despite their good intentions make mistakes or too rushed classifications (e.g. placing geotags in the middle of cities when more specific information is available).
Conclusion

Situation awareness is the main precursor to appropriate decision making [30]. Social media, in particular Twitter, has emerged as a new source of citizen-generated reports that can potentially offer a detailed real-time view of the situation on the ground during large-scale complex disasters. As physical access to affected areas can be restricted and no response organization has resources to be everywhere, such cheap distributed sensing mechanisms are highly attractive during mass disaster and conflict.

During crises affecting millions of people, it is not uncommon to see hundreds of thousands of social media messages being generated every hour and information management tools are needed to effectively extract and organize relevant information in real time. However, social media messages have proven difficult to process using traditional natural language processing algorithms and most success stories in the disaster space have relied on organized crowds of volunteers who process content manually. As this approach on its own has scalability issues, information overload has remained a serious barrier that obstructs integration of social media in decision making during mass disasters.

In this study we demonstrate how combining crowd curation with automated data collection and language-independent real-time text clustering can achieve scalability, accuracy, timeliness and flexibility. Based on feedback and usage statistics collected from 48 participants during an 8-day evaluation deployment focused on the 2012 Syrian civil war, we show that our proposed architecture produces the first system that enables comprehensive real-time overview of Twitter during mass disasters. Stories (tweet clusters) in CrisisTracker contain rich event descriptions, images and video and can in most cases reliably be detected around 30 minutes after major events, placing them between immediate eyewitness reports and traditional media coverage. While clustering helps assess the credibility of reports, the traditional disaster response practice of relying on multiple independent sources remains strongly recommended to avoid false leads.

Both expert and volunteer participants found the system to be intuitive and easy to use. Analysis is however time consuming and we conclude that the current system is ready for use by analysts, but not directly by decision makers. We also recommend that a new role, the crowd director, is assigned to bridge the gap between volunteer curators and organizations that wish to deploy a system like CrisisTracker.

Future Work

Three critical extensions are needed to make CrisisTracker ready for live use by humanitarian organizations around the world. Functionality needs to be added to export data from the system into other products currently in use by relief organizations and volunteers. Mechanisms also need to be added that help prevent and reduce impact of vandalism. Finally, as time is a scarce resource in particular during rapid onset disasters, the system needs to be made easier to deploy.

Future work will look at improving the visualizations to support quicker and more accurate interpretation of real-time data, as well as deeper analysis of the relationships between locations, sources, named entities and events. We also plan to integrate semi-supervised classification algorithms, which can learn from human curators and partly automate meta-data creation.
Furthermore, we hope to add a high-level layer of collaborative interpretation and analysis that goes beyond current functionality for organizing and labeling content.

Acknowledgements

We express our sincere gratitude to all the curators and disaster response experts who participated in this study. We thank Patrick Meier for promoting the evaluation as a Standby Task Force deployment and for his guidance in crowd management. We also thank Syria Tracker for their support in making the deployment as realistic and safe, and for actively using the system in their daily monitoring of the conflict. We thank Rebecca Curzon and Vicki Kraeling for promoting the study with disaster practitioners. Finally, we thank Ko-Hsun Huang for key guidance in gathering, analysis and interpretation of data.

This work is supported by an IBM Open Collaboration Research Award and by the Portuguese Foundation for Science and Technology (FCT) grant CMU-PT/SE/0028/2008 (Web Security and Privacy).

References

[3] Twitter, Online: http://twitter.com
[25] P. Meier, "Some Thoughts on Real-Time Awareness for Tech@State", Online: http://irevolution.net/2012/02/02/real-time-awareness/
[26] Ushahidi, Online: http://ushahidi.org
[31] Sahana Software Foundation, Online: http://sahanafoundation.org/
[32] VirtualAgility WorkCenter, Online: http://www.virtualagility.com/
[34] "Introducing the Standby Task Force", Online: http://blog.standbytaskforce.com/about/introducing-the-standby-task-force/
[36] EMM NewsBrief, Online: http://press.jrc.it
[40] CrowdGlobe, Online: http://crowdglobe.net/
[41] Crowdmap, Online: https://crowdmap.com/
[42] Syria Tracker, Online: https://syriatracker.crowdmap.com/
[43] Open Street Map, Online: http://www.openstreetmap.org/
Table 1: Feature comparison of related state of the art systems.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Full-featured crisis management tool</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Easily handles new event types and languages</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Full 4W metadata (who, what, when, where)</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>User-controlled information management</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>De-duplication of redundant reports</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Scalable real-time overview of social media</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
</tbody>
</table>

(● supported, ○ partly supported, ☑ not supported)
Figure 1: The information processing pipeline in CrisisTracker, consisting of three modules. Word frequencies from the data collection module are used for similarity calculations in both the story detection and crowd curation modules.
Figure 2: Left: User interface for exploring stories, with filters for category (1), keywords (2), named entities (3), time (4) and location (5), with matching stories below (6). Right: A single story, with title (7), first tweet (8), grouped alternate versions (9) and human-curated tags (10).
Figure 3

Figure 3a: Reduction in information inflow rate from each processing step. “N+ users” refers to stories containing tweets from at least N unique users. Numbers in parentheses indicate additional items produced by the 58 spammer accounts. 3b: Growth rate of stories. If stories are filtered by size, impactful events can realistically be detected around 30 minutes after the first tweet.
Biographical sketches

Jakob Rogstadius. Madeira Interactive Technologies Institute, Caminho da Penteada, 9020-105 Funchal, Portugal (jakob@m-iti.org). Mr. Rogstadius is a Ph.D. candidate at Madeira Interactive Technologies Institute, Research Associate at Qatar Computing Research Institute and lead developer of CrisisTracker. He received a M.Sc. in media technology and engineering from Linköping University in 2009 and has previously worked with information visualization research and decision support systems development in industry and academia.

Maja Vukovic. IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (maja@us.ibm.com). Dr. Vukovic is a Research Staff Member in the Services Delivery Department at the Thomas J. Watson Research Center. She received a dual B.Sc. degree in mathematics and computer science from University of Auckland in 1998, M.Sc. degree in computer science from International University in Germany in 2000, and Ph.D. degree in computer science from the University of Cambridge in 2006. She worked as a Research Scientist at MercedesBenz Research Lab in Palo Alto on remote vehicle customization and telematics systems. At IBM, Dr. Vukovic is working on enterprise crowdsourcing and identity access management systems. She is an author or coauthor of over 50 technical papers, and has organized several workshops on crowdsourcing and social media for disaster management. Dr. Vukovic is a senior member of IEEE (Institute of Electrical and Electronics Engineers).

Claudio A. Teixeira. Madeira Interactive Technologies Institute, Caminho da Penteada, 9020-105 Funchal, Portugal (claudio.teixeira@m-iti.org). Mr. Teixeira is a Research Assistant at the Madeira Interactive Technologies Institute. He received a B.Sc. degree in informatics engineering from University of Madeira in 2010, and is currently finishing his M.S. degree in informatics engineering from the University of Madeira.

Vassilis Kostakos. Department of Computer Science and Engineering, University of Oulu, Oulu 90570, Finland (vassilis@ee.oulu.fi). Dr. Kostakos is Professor of Computer Engineering in the Department of Computer Science and Engineering at the University of Oulu. He received a B.Sc. and Ph.D. in computer science from the University of Bath in 2001 and 2004 respectively. In 2010 he became a fellow of Academy of Finland Distinguished Professor Program, and in 2012 he received a Marie Currie Career Integration Grant. He conducts research on human computer interaction, ubiquitous computing, and social media. He is a member of ACM (Association for Computing Machinery).

Evangelos Karapanos. Madeira Interactive Technologies Institute, Caminho da Penteada, 9020-105 Funchal, Portugal (ekarapanos@uma.pt). Dr. Karapanos is an Assistant Professor at Madeira Interactive Technologies Institute where he also acts as the academic director of the Carnegie Mellon | Portugal Professional Masters of Human-Computer Interaction. Dr. Karapanos holds a Ph.D. (cum laude) from Eindhoven University of Technology on methods and tools for user experience evaluation. His work has been published among others at the ACM SIGCHI Conference on Human Factors in Computing and the International Journal of Human-Computer Studies. He also published the book Modeling Users' Experiences with Interactive Systems (Springer) in 2012. He has been a visiting researcher at Philips Research, Philips Consumer Lifestyle and the Human-Computer Interaction Institute at Carnegie Mellon University. Dr. Karapanos is a member of ACM (Association for Computing Machinery).

Jim A. Laredo. IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (laredoj@us.ibm.com). Mr. Laredo has over 20 years experience in the Information Technology Industry, including the areas of Transaction Processing (Transarc Corporation), Business Process Management (Vitria Technology), Software as a Service, Cloud Computing and Systems Management (IBM). He has been building systems and services for customers in Financial Services, Telecommunications, Insurance Services and Manufacturing. Jim has spent the last 10 years at IBM, 8 of which at the T.J Watson Research Center in NY, and the last 3 years in particular working in the area of Enterprise Crowdsourcing, looking for innovative techniques that leverage people's knowledge to scale business process management across the organizations of the enterprise. Mr. Laredo has an Engineering degree from Universidad Simon Bolivar, Caracas Venezuela, and a M.Sc. in Computer Science from University of Toronto.