Crowd-Sensing Meets Situation Awareness
A Research Roadmap for Crisis Management

Andrea Salfinger*, Sylva Girtelschmid†, Birgit Pröll‡, Werner Retschitzegger§, Wieland Schwinger¶
*§¶Dept. of Cooperative Information Systems, †‡Inst. for Application Oriented Knowledge Processing
Johannes Kepler University Linz
4040 Linz, Austria
{andrea.salfinger*, werner.retschitzegger§, wieland.schwinger¶}@cis.jku.at, {sgirtelschmid†,bproell‡}@faw.jku.at

Abstract

When disaster strikes, human lives may depend upon emergency organizations' rapid establishment of Situation Awareness (SAW) to take the appropriate decisions and actions. Recently, systems emerged, enabling humans to act as crowd sensors contributing valuable crisis information via mobile devices through social media channels. This should allow enhancing situational pictures gained through traditional SAW systems, as employed in control center domains. A common understanding about the necessary functionality of such crowd-sensed SAW systems for crisis management, however, is not yet reached nor is a detailed comparison thereof available up to now. This paper makes a first attempt towards this by a reference architecture incorporating crowd-sensed crisis information into SAW systems. Based on that, a systematic catalog of evaluation criteria is derived and used for an in-depth comparison of nine existing systems, thereby highlighting current capabilities and directions for further research.

1. Introduction

Situation Awareness. When disaster strikes and crisis like earthquakes, grassfires or floods occur, human lives may depend upon emergency organizations' rapid establishment of an understanding of the situation at hand, termed *Situation Awareness* (SAW) [10], to take the appropriate decisions and actions in managing such crisis. Systems supporting human operators' SAW by fusing various sensor information have already proven their value in a range of different environmental monitoring applications (e.g., air and road traffic monitoring [3] or maritime surveillance [29]). The situational pictures observed in these domains employing automated situation assessment (SA) mainly base on information gained from traditional "hard" sensors [22]. Situations of interest are thereby either assessed on basis of predefined situation templates [3] or detecting anomalies within the environment [29].

Crisis Management. Situations that need to be managed during crisis, however, are often *unpredictable*, *unique* and of *large-scale dimensions* [41], and therefore pose unprecedented challenges for such SAW systems supporting human operators, as they may not be sufficiently covered by sensors. Furthermore, infrastructural damage may even worsen sensor coverage.

Crowd-Sensing to the Rescue. Recently, chances to mitigate these problems have appeared: The rise of social media (SM) and ubiquitous hand-held devices have enabled humans to act as *crowd sensors* (or *citizen sensors* [35]), contributing valuable crisis information [9], [43] in (near) real-time via SM channels like Twitter. It has also been shown that SM channels can be maintained, or quickly re-established, even in severe disaster situations, such as earthquakes [32], grassfires or floods [40]. However, sharing information through SM entails certain peculiarities such as noise, brevity, specific conversational practices (like hashtags, shortened URLs, improper grammar, lack of context). Besides that, also the *sparseness* of actually relevant information within the plethora of conversations [42] renders exploitation of SM data for SAW a non-trivial task, contrasting existing systems for joint fusion of hard and soft sensor data like [22].

Contribution. Whereas initial approaches like [1] and [30] already visualized crisis information sensed from SM, most of them do not fully leverage automated SA, leaving the actual SA tasks to the human operators. Therefore, this paper makes a first attempt towards bridging the gap between traditional, authority-sensing based SAW systems and crowd-sensed SAW. For this, we propose a reference architecture extending the well-known JDL data fusion model for automated SA [21] with additional system components addressing challenges faced when incorporating crowd-sensed information into SAW systems. Based on this, we derive a systematic catalog of criteria for evaluating crowd-sensing SAW systems and present an in-depth comparison of nine existing systems, thereby highlighting current capabilities and directions for further research.

Structure of the Paper. We discuss related work in Sec. 2, before deriving a reference architecture and a set of evaluation criteria in Sec. 3. Based thereupon, we evaluate approaches for crowd-sensed SAW and elaborate on lessons learned in Sec. 4, before concluding in Sec. 5.

2. Related Work

To the best of our knowledge, there is neither a dedicated architecture, nor a survey of SAW systems for the application domain of crisis management (CM), integrating both, authority-sensed and crowd-sensed information. However, valuable preparatory work can be found in related areas, which we discuss in the following.

Llinas [22] studied sixteen problem-domain-agnostic information fusion (IF) frameworks to identify common principles. Based on lessons learned, he proposes a domain-agnostic, robust and adaptive IF framework covering all IF levels. This already comprises integration of hard sensors and soft sensors, and thus represents a foundation for our architecture. However, its generic and domain-agnostic nature is not fully suited for elaborating on the specific challenges encountered with sensing untargeted SM content for SAW systems.

Salfinger et al. [33] surveyed SAW systems for supporting environmental monitoring tasks, with focus on systems' coping with evolving situations. Whereas several of their criteria are also relevant for our work, they do not encompass crowd-sensing approaches. Nevertheless, [33] can be seen as a complimentary survey to our research work.

Stavrakantonakis et al. [37] presented an evaluation framework for SM monitoring tools, focusing on concepts, user interfaces, and technology. Since investigating general SM monitoring tools with an application focus on market research, they consequently do not consider several aspects relevant for SAW applications for CM. However, we incorporate several of their criteria in our current study where appropriate.

Concluding, although there is already valuable work available, a dedicated reference architecture allowing for a comparative study of current crowd-sensing approaches enhancing SAW in CM is still lacking.

3. A Reference Architecture for Crowd-sensing SAW Systems

In this section, as basis for our comparative study, we propose a first attempt towards a reference architecture for crowd-sensing enhanced SAW systems for CM. These represent a subset of *emergency management information systems* [6] and *Decision Support Systems* (DSS) for emergency management (focusing on the emergency phases of detection, mitigation, response and recovery) [41].

Design Rationale. We base our conceptualization on the well-established *JDL (joint directors of laboratories)* data fusion model [21], representing the common reference architecture for IF systems, and architectures based thereupon [22], [33]. This is since, from a technical viewpoint, SAW systems implement an IF system,

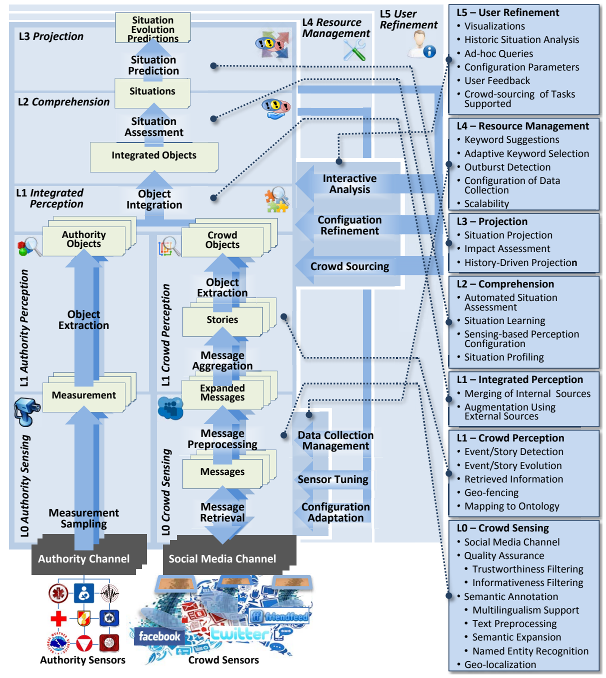

Figure 1. Reference Architecture for crowd-sensing SAW.

as situations can be regarded as sets of interrelated objects which are of interest to the human operators.

Our architecture (cf. Fig. 1) is thus structured according to the six JDL levels comprising the *Sensing Level (L0)*, responsible for observing the environment through various sensors, the *Perception Level (L1)*, fusing these measurements *within* and *across* sensors in order to infer relevant real-world objects, the *Comprehension Level (L2)*, exploiting relationships between objects in order to assess situations, the *Projection Level (L3)*, estimating the impact and likely development of situations, *Resource Management (L4)*[1], providing an adaptive feedback loop for all system components, e.g., sensor tuning, and finally *User Refinement (L5)*, comprising the user's interactions and her potential adaptations at all processing levels. As Fig. 1 shows, levels L0 to L3 base upon each other, by aggregating and interpreting data from the precedent level. Contrastingly, levels L4 and L5 influence other levels, forming feedback loops orthogonal to the functional levels. Both levels allow for an adaptive SAW system (e. g., L4 may comprise feedback loops from SA or object fusion that automatically adapt the sensing level, by actively triggering the sensors to retrieve more information on identified objects or situations), with the user being enabled to interact upon any of the levels' components (e. g., the operator may inspect the sensing output and reconfigure the sensors accordingly).

[1]Despite the term "Resource Management", chosen to be in line with the JDL model, Level 4 comprises a broader understanding including, e.g., also configuration.

The JDL model represents an agreed upon framework [22] denoting the core functional components a full-fledged SAW system should provide. However, whereas it outlines the processing and functionality of SAW systems in principle, its high-level perspective does not allow elaborating on the challenges and requirements faced when including crowd-sensed data in particular. In order to derive more specific criteria for studying crowd-sensing systems in the context of SAW for CM, an extension to this model dwelling on the peculiarities of crowd-sensed SM data is needed.

In the following, we investigate on the novel challenges imposed by crowd-sensing for CM on each of the JDL processing levels. Based on these, we introduce necessary functional components addressing these and derive an evaluation criteria framework structured according to the six JDL levels (cf. Fig. 1). The set of evaluation criteria is achieved by studying related crowd-sensing systems in a bottom-up way. These are supplemented by criteria imposed from the CM domain and requirements on SAW systems in a top-down fashion, thus allowing a systematic evaluation pointing towards directions for further research and ideas for potential technical solutions. These criteria, however, only evaluate the presence of certain abilities at a rather coarse-grained level, while the actual abilities depend on the effective configuration and parametrization and thus may vary in the context of the application at hand.

Due to our focus on information processed through crowd-sensing, we restrict our discussion of the *Sensing* (L0) and *Perception Level* (L1) to challenges induced by crowd-sensed data only. We omit elaborating on processing of data obtained from authority sensors (comprising *Authority Sensing* L0 and *Authority Perception* L1), since this is already extensively covered in literature [22], [33].

A. Crowd Sensing (L0)

The aim of this level is to gain potentially relevant information from diverse SM channels, posted by a *crowd* of human observers (each regarded as a single *crowd sensor* [24], [31]), to complement or enhance the situational picture obtained from *authority sensors*. Whereas crowd-sensing could also be employed for structured, i.e., "hard" data, for instance delivered by mobile devices' sensors (e. g., motion data obtained from smartphones), their processing resembles (hard-sensor based) authority-sensed data. Therefore, we will restrict our focus to crowd-sensed data comprising unstructured, i.e., free-text data requiring dedicated processing.

Consequently, this level's input consists of *SM messages*[2] (e. g., Tweets, Facebook status updates), from which relevant observations of evolving events, objects or situations should be extracted for further processing. The criteria and necessary functional components are motivated by the very unique nature of (soft) crowd sensors, being, in contrast to authority sensors, characterized by (i) *unknown sensing behavior* of crowd sensors, meaning that the SAW system needs to detect when they are becoming *active* [31], i.e., start delivering crisis-specific observations, such as eye-witness reports of a flooding site, (ii) *untargetedness* in that SM users may post about virtually any topic, inducing that only a small and varying subspace of SM will comprise valuable crisis information, and (iii) *unstructuredness*, i.e., most of the meaningful information (potentially containing spatio-temporal-thematic characteristics [23], [24]) is concealed in free-form text in the messages' content together with some structured meta-data depending on the SM channel.

These special characteristics of crowd sensors require preprocessing SM messages in various ways and lead to the following criteria and functional components, finally resulting in so-called *expanded messages* being the input for the next level (L1).

Social Media Channel: For accessing and retrieving content from SM channels, dedicated *adapters* are needed, tailored towards a SM platform (e. g., Twitter, Facebook, Foursquare, Weblogs etc.). These may be implemented as crawlers, or queries over APIs provided by a SM platform (e. g., the Twitter Search or Twitter Streaming API).

Quality Assurance. The following criteria study the means for quality-related filtering.

Trustworthiness Filtering: SM content comprising *misinformation* and *disinformation* may confound instead of enhance crisis SAW [26]. An extensive literature review on assessing the trustworthiness of SM for CM has been conducted in [13]. In the light of this, trustworthiness filtering, which, on this level, considers each message in isolation, can be based on different information: Firstly, on analyzing the SM messages' meta-data (e. g., the poster's reputation by using spam-detection modules), or, secondly, textual content itself (e. g., scanning for specific words or character sequences, e.g., "oooooh", indicating emotional Tweets which may comprise lower information content than content produced by *High Yield Twitterers* [40]). Thirdly, especially w.r.t. crisis situations, the number of retweets can indicate trustworthy information [40], allowing for specific quality metrics [38].

Informativeness Filtering: Not all messages retrieved will be informative w.r.t. the current information need. For example, when retrieving Tweets matching the keyword "earthquake", messages like "Yesterday I attended the earthquake conference" will be delivered as well [31]. Informativeness filters may be based on

[2]The term "message" refers to SM content posted by a specific SM user at a specific point in time.

machine learning classifiers (such as SVMs) trained (offline) on manually labeled data sets (e. g., [17], [31]), which operate on word vectors obtained and discard negatively classified Tweets. Therefore, this criterion investigates if methods for a filtering of messages w.r.t. their information content are provided.

Semantic Annotation. Whereas the previous filtering approaches may operate solely on the basis of metadata or lexical properties, the criteria category semantic annotation identifies abilities w.r.t. the *semantics* of SM content based on techniques such as Information Extraction (IE) or Natural Language Processing (NLP), ideally by considering the specific context [15].

Multilingualism Support: As only 50% of Twitter messages are in English [14], being an issue in emergency situations like the Haiti earthquake [9], it is crucial to overcome language barriers in terms of multilingualism. Thus, this criterion tests if automated translation facilities, as in [19], are exploited.

Text Preprocessing: Prior to semantic analysis, the textual content needs to be suitably preprocessed, which potentially encompasses classical NLP steps such as tokenization, word cleaning, stop-word elimination, stemming (e. g., reducing "flooded" and "flooding" to "flood"), and possibly dictionary lookups (e. g., WordNet) to only retain meaningful words.

Semantic Expansion: Due to ambiguity of natural language, SM users may employ different terminology to describe the same. This criterion thus studies whether means for semantic expansion are provided by including *syntagmatic relationships* between words (e. g., accounting for synonyms [25]).

Named Entity Recognition (NER): NER refers to the process of annotating textual entities w.r.t. the *semantic categories* they refer to (e. g., "Steve Jobs" refers to a *person*, "Apple" may refer to *food* or *company*). Whereas automated NER methods have been extensively studied on documents, such as news articles, *entity span identification* and *entity class annotation* on SM content have proven to be difficult. Even human annotators vary considerably in their annotation results [7], especially in case of location entities being, however, quite important for crisis management [23]. We therefore discuss this in the separate criterion *Geo-localization*. Furthermore, SM content requires the decoding of *"unmarked"* information, which represents information tacitly understood by humans in a certain context [40]. For instance, during Colorado flooding, people simply mentioned "the River", referring to the Red River [40]. Whereas these aspects represent severe challenges, recently promising IE techniques have been developed for SM, such as Part-of-speech tagging approaches for Twitter (e. g., [12]), dedicated NER approaches (e. g., [20]), and even full-fledged IE processing pipelines (e. g., TwitIE [5]).

Geo-localization: Geo-localization refers to locating a SM message and/or its content to real-world geographical reference points. Geo-localization can be performed on basis of GPS coordinates embodied in message-specific metadata (User's Current Location [16]) or by employing user-specific meta-data (e. g., Twitter-users can specify their location, i.e., a User's Location Profile [16]). These approaches are, however, less favorable, since only 0.7% [25] of Tweets are geolocated, and it just represents the location of the user, which does not necessarily correspond to the location referring to in the crisis report. The analysis of the message's *textual content* itself is more promising. This, however, requires a NER approach, and a mapping of the extracted location entities to coordinates. Locations in text entail further challenges [16], requiring *toponym resolution* (e. g., "Obama" may refer to a president of the US or a city in Japan), on basis of further contextual information.

B. Crowd Perception (L1)

The aim of this level is to aggregate relevant SM messages which were identified at L0 to so-called *stories* [24], [30], each referring to a specific, cohesive topic (e.g., flooding of a certain section of a street). This is another crucial task, since case studies have shown that SM data are predominantly of use in crisis situations if they are on an aggregate level [9]. Furthermore, this level is responsible for conceptualizing the inferred stories by extracting so-called *crisis objects* (e. g., infrastructural objects, like bridges and buildings, or crisis types, like heavy storms, flooding, casualties). Conceptualization can be achieved by an ontology available across all levels, allowing for machine-based processing and reasoning in subsequent SAW system levels. Finally, the derived crisis objects need to be fused, both, *within* a certain SM channel (i.e., fusion of observations over time reconstructing the objects' evolution) and *across* (i.e., integrating observations stemming from various SM channels), before integrated with object-level information obtained from authority sensors at the next level, *Integrated Perception* (L1).

Event/Story Detection: Similar to conventional sensor-based systems, real-world events will likely be sensed by multiple crowd-sensors (i.e., multiple people report on the same event). These multiple observations need to be fused to a single, coherent description of the underlying real-world event, which can be further on processed by the SAW system. Dedicated *Event/Story Detection* methods aim at inferring such underlying real-world events from analyzing and aggregating related SM content. As opposed to *Trustworthiness Filtering* performed on L0, which treats each message in isolation, these methods analyze multiple messages in conjunction. Thereby, the confidence in reported events is increased. If a multitude of observations reports a specific state of affairs, but isolated messages report contradictory information (thus potentially comprising *misinformation* or *disinformation*), these are dealt with thereby.

Many approaches for *Event/Story Detection* base on *clustering* of SM content (e. g., by computing the cosine similarity between the messages' TFIDF vectors [43]). The resulting spatio-temporal-thematic clusters are assumed to describe the real-world event discussed in this cluster, often termed a *story* [24], [30]. By extracting suitable spatial, temporal and thematic *descriptors* from these clusters [24], the spatial (i.e., location), temporal and thematic (i.e., the type of the event) properties of the underlying real-world event can be estimated.

Another broad class of event detection and tracking techniques bases on probabilistic approaches, which compare the actually encountered message density w.r.t. specific topics against the *expected density* [32], [36]. This allows incorporating *temporal context* (e. g., day-and-night and weekday-specific SM usage patterns [32], [36]) and *spatial context* (e. g., different usage patterns in rural and urban areas). However, one needs to consider *concept drift*, therefore these methods ideally should be able to detect when underlying (contextual) reference patterns change over time.

Event/Story Evolution: Since CM stretches over the entire lifecycle of a crisis, continuous tracking of and updates on the crisis situation are required. Therefore, crowd-sensing approaches need to be capable of detecting and tracking the evolution of the extracted events or stories (i.e., *storylines*), in order to deliver adequate update information (e. g., [24], [31]).

Retrieved Information: This criterion examines which types of information are derived from events inferred from SM, in terms of *spatial*, *temporal*, and *thematic* information, and further, *sentiment* dimensions, which may provide indication on the severity and evolutionary phase of the disaster.

Geo-fencing: Whereas the previous criterion *Geo-localization* addressed the geographic positioning of single SM messages, we term the *geo-localization* of *entire stories* or *events* as *geo-fencing*, i.e., the determination of their location, geographical extent, and possibly the magnitude of SM content that contributed to those (e. g., [39]). This allows operators to assess the extent of the crowd-sensed crisis event, which may exhibit dynamically varying geographical boundaries over time (analogously to *situation fencing*, i.e., geo-fencing w.r.t. user-centric situations [27]). Furthermore, by aggregating over the multitude of reported (or user) locations, the effect of outliers can be mitigated, thereby raising the overall quality of extracted information.

Mapping to Ontology: A large fraction of SAW systems utilizes semantic reasoning techniques for SA, predominantly those that operate on heterogeneous domains comprising a large variety of encountered object and situation types [33], as in CM [18]. Therefore, these approaches base on ontological descriptions codifying domain specific knowledge throughout the different levels of the monitored domain, on which their reasoning functionality is based. This implies that crowd-sensed crisis events need to be mapped to object types of the employed ontology, in order to be processable by SAW systems. Therefore, this criterion studies whether such an ontology mapping is performed by the studied system.

C. Integrated Perception (L1)

In order to obtain a coherent perspective on the monitored environment, being the prerequisite for SA, the object observations retrieved from crowd-sensed data (which we term *crowd objects*) need to be fused with observations obtained from the authority sensors (i.e., *authority objects*), cf. Fig. 1. This integration step thus requires appropriate merging and conflict resolution strategies in case the different sources report adversarial information.

Merging of Internal Sources: This criterion analyzes whether a system provides means to merge information obtained from a specific SM channel with data obtained from other sensors, such as authority sensors or other SM channels.

Augmentation Using External Sources: Furthermore, the system may actively seek for gathering additional information by retrieving data from complimentary sources (e. g., ESA collects photos from Flickr that match a detected incident [43]).

D. Comprehension (L2)

The incorporation of crowd-sensing components adds additional challenges, but also novel possibilities, to the *Comprehension* level. Different evolutionary phases of a crisis imply and require different actions and thus need to be adequately detected and classified by the SAW system. For reasons of intuitivity, we include criteria reflecting feedback loops triggered by the *Comprehension* level in this subsection, and not in the *Resource Management* level.

Automated Situation Assessment: First of all, the SAW system needs to be capable of detecting *evolving situations* and highlighting the current evolutionary phase of the encountered crisis situation(s), and to adequately incorporate *situational update* information that can be sensed from SM [40].

Situation Learning: Since crisis situations are often *unexpected* and *unique* [41], SAW systems need to be capable of reacting to previously "unseen", i.e., novel, situations. However, by sensing and tracking the situational descriptions from the crowd, the system could aim at dynamically *learning* the type of encountered crisis situation, which could be incorporated in a knowledge base to improve future detections. Thus, this criterion analyzes whether means for *learning situation types from the crowd* are provided.

Situation-based Perception Configuration: The evolution of a crisis, comprising the phases *warning*, *preparation*, *climax*, and *recovery*, also determines which types of information can be retrieved from SM [40]. This allows an intelligent feedback loop from the *Comprehension* level to the *Crowd-Sensing* level: Once a crisis' evolutionary phase is detected, the semantic annotation modules on the crowd-sensing level could be fine-tuned w.r.t. the expected types of information that should be extracted.

Situation Profiling: Based on a partially assessed situation, the SAW system could actively trigger the crowd-sensing adapters to gather missing information pieces. In the Twitcident system [1], for example, based on an initial *incident profile* obtained from emergency broadcasting services, an incident profile (comprising initially populated spatio-temporal-thematic information slots) is created. The system seeks to retrieve further information matching this profile from SM, which continually refines this incident profile.

E. L3 Projection (L3)

The projection level aims at forecasting the crisis situation's evolution and impact, thereby providing a foundation for emergency managers' and operators' decision making on adequate actions to mitigate the situation.

Situation Projection: This criterion evaluates if the studied system provides any predictions w.r.t. the likely development of the detected situations [2]. The incorporation of crowd-sensed SM content opens up novel means for such predictions, as SM users may provide forecasts or indicate where needs are emerging. However, extracting and exploiting this information presents a remarkable challenge on NLP modules and the SAW system's reasoning facilities. Furthermore, emerging trends could be used for forecasts (similar to market predictions based on SM trending topics).

Impact Assessment: Determining the potential impact of a situation represents a related challenge, however, focuses on recognizing the affected area and the consequences. In CM, impact assessment is mainly concerned with assessing infrastructural damage (such as flooded bridges and roads). Since this represents a remarkable requirement from emergency managers and operators, as outlined in [43], this criterion therefore evaluates whether automated impact assessment functionalities are provided (e. g., ESA [43]).

History-driven Projection: Projections on a crisis situation's likely evolution and its impact could be enhanced by exploiting experiences from previously encountered, similar crises. Thus, in order to learn from historic situations, this criterion assesses if means to build up and exploit a *crisis memory* [41] are provided.

F. Resource Management (L4)

Due to the "untargetedness" of SM, the addition of crowd-sensors to a SAW system introduces novel challenges to the *Resource Management* layer. Whereas authority sensors are dedicated to deliver information w.r.t. the monitoring tasks, SM are neither focused on a specific topic nor audience, as already mentioned. The actual *information of relevance* needs to be identified, requiring an appropriate configuration and management of sensors seeking to retrieve this information from SM.

Keyword Suggestions: Many crowd-sensing approaches base their gathering of SM content on keywords reflecting crisis information of interest, thereby requiring adequate keyword specification. Therefore, this criterion studies if assistance regarding the definition of suitable keywords is provided (e. g., [24] identifies an initial set of keywords through *Google Insights for Search*).

Adaptive Keyword Selection: This criterion (similar to *listening grid adjustment* [37]) analyzes if the keyword set is adaptively refined over time w.r.t. extracted information and the evolution of encountered events [24].

Outburst Detection: Outburst detection represents a different means for managing the collection of crisis-related information from the crowd, which aims at detecting sudden *outbursts* on specific topics. These may indicate large-scale events concerning a considerable fraction of SM users. Such techniques may base on *language models* encoding the expected word frequencies [43], and thus may be subject to *Concept Drift*, i.e., need to detect when underlying distribution changes over time.

Configuration of Data Collection: This criterion assesses if means are provided to adequately configure data collection from SM channels, such as specifying and adapting its *duration* and *intervals*. This is required to account for distinct *spatial and temporal granularity* of different event types, such as slowly developing long-term events like a drought crisis, requiring long-term monitoring, but exhibiting less-frequent updates, and short-term, quickly developing crises such as a terrorist attack, which necessitate fine-grained SM sensing intervals in order not to miss the frequent short-term update information [24].

Scalability: The high volume, velocity, and variety of SM data [8] pose significant demands on systems exploiting these [34]. Therefore, this criterion evaluates if means to scale to *Big Data* demands are provided, such as *distributed processing* (e. g., SensePlace2 [23] allows for multiple *semantic annotation* module instantiations), *indexing* schemes for quickly retrieving relevant data (e. g., SensePlace2 [23] utilizes sophisticated text search engines), or storing the gathered SM content in NoSQL databases (e. g., [34]).

G. User Refinement (L5)

User Refinement encompasses both, how the SAW system's output is presented to the operator, and thus refines her mental models, as well as the operator's interaction with the system, for instance by dynamically configuring its components to suit her current information needs [4].

Visualizations: CM applications typically involve *map-based visualizations* for conveying the observed crisis situations to operators, who need to assess the geographical location and extent of these. Visual analytics features, such as *drill-down* functionalities (e. g., [24]), *highlighting* of the most relevant Tweets w.r.t. specific criteria (e. g., SensePlace2 [23]), or *timelines* (e. g., SensePlace2 [23], CIACM [36]), outlining the development of the encountered crisis situation, provide further means for enhancing the operator's SAW by interactively exploring the situation and its constituents.

Historic Situation Analysis: This criterion studies whether users are enabled to perform an analysis of historic crisis situations, e. g., by providing a kind of "replay" functionality allowing to assess the development of historic crisis situations (e. g., ESA [43]), in order to benefit from past experience.

Ad-hoc Queries: To satisfy their current information need, operators should be enabled of issuing queries to seek additional information (e. g., HADRian [39]).

Configuration Parameters: This criterion evaluates if the operator can configure the system by specifying parameters (e. g., Twitris [24]), determining whether the system's focus w.r.t. topic tracking should be on global vs. local and on recent vs. prolonged SM trends.

User Feedback: This criterion studies if feedback provided by operators is incorporated by the system (e. g., by refining its story generation thereupon). This could be realized by providing operators with dedicated *rating metrics* for judging the quality of information and recommendations provided by the system, which should trigger the system's adaption components to optimize its performance w.r.t. these metrics [11].

Crowd-sourcing of Tasks Supported: Crowd-sourcing means that volunteers take over tasks that prove to be simple for humans, but cannot be solved well by machines, thus perfectly complementing the machines capability w.r.t. computational throughput on processing humongous data sets with humans' strengths w.r.t. reasoning on implicit and incomplete information. This criterion evaluates if the crowd-*sensing* SAW system also allows that specific tasks are crowd-*sourced*.

4. Lessons Learned

Our comparative evaluation is set on contrasting different state-of-the-art systems that make use of crowd-sensing in order to derive situational pictures in (near) real-time, for the application domains of CM, disaster relief and emergency response. Consequently, we left aside approaches which addressed only specific levels or aspects. Thus, we regarded the following approaches: seven crowd-sensing approaches for the application domain of Humanitarian Aid and Disaster Relief (HA/DR), namely HADRian [39], ESA ("Emergency Situation Awareness") [43], Twitris [24], [28], Twitcident [1], SensePlace2 [23], CrisisTracker [30] and TweetTracker [19]. We included two more widely related systems to broaden our evaluation set w.r.t. approaches basing on probabilistic *Event/Story Detection*, i.e, Toretter [31] and CIACM ("Canary in a Coalmine") [36]. These focus on detection of natural disasters and chemical incidents, respectively, thereby focusing on specific types of disasters. Both systems can be configured towards other types of disasters, thereby still meeting our application domain of CM.

Table 1, condensing our evaluation, reveals interesting concepts, but also open issues which are not focused in current systems, thereby indicating directions for further research, which we will summarize w.r.t. our architectural levels.

Moderate Quality Assurance. Although some approaches provide means for informativeness filtering, assessment of trustworthiness is only performed by two approaches, Twitcident [1] and CrisisTracker [30], which filter out messages that are considered to be too short to convey substantial information, or correspond to emotionally focused Tweets.

Limited Semantic Annotations (L0). Whereas several of the evaluated approaches employ NER-techniques, current solutions have not yet fully elaborated on extracting and exploiting the semantics comprised in a message's content. Approaches employing NER mostly focus on extracting *location* entities, followed by *person* and *organization* entities (e. g., in Twitcident). Only Twitris provides a semantic analysis in terms of incorporating *relationships*, by computing a *contextually enhanced thematic score* taking into account strong associations between extracted descriptors, that are, however, computed on the aggregate (i.e., story) level.

Support for Crowd Perception. Most of the approaches provide techniques for the detection of events/stories, mostly based on clustering methods, and also means for story evolution. Further on, almost all support different kinds of retrieved information in terms of spatio-temporal-thematic dimensions. Except sentiment analysis and geo-fencing are supported by a few approaches only.

Lack of Integrated Perception (L1). Current crowd-sensing approaches rarely allow for integrated perception, although integration with authority-sensed data would be crucial for crowd-sensing SAW systems.

No Support for Comprehension (L2). Whereas most of the evaluated systems explicitly state their aim of supporting an operator's SAW, we found that attempts

Criterion			HADRian	ESA	Twitris	Twitcident	SensePlace2	CrisisTracker	TweetTracker	Toretter	CIACM
L0 - Crowd Sensing	Social Media Channel	Twitter Streaming API	✗	✓	✗	✓	✗	✗	✗	✗	✗
		Twitter Search API	✓	✓	✓	✓	✓	✓	✓	✓	✓
		Other	✓	✗	✗	✗	✗	✗	✗	✗	✗
	Quality Assurance	Trustworthiness Filtering	✗	✗	✗	✓	✗	✓	✗	✗	✗
		Informativeness Filtering	✓	✓	✗	✓	✗	✗	✗	✓	✗
	Semantic Annotation	Multilingualism Support	✗	✗	✗	✓	✗	✗	✓	✗	✗
		Text Preprocessing	✗	✓	✓	✓	✗	✓	✗	✓	✗
		Semantic Expansion	✗	✗	✓	✗	✗	✗	✗	✗	✗
		NER	✓	✓	✓	✓	✓	✗	✗	✗	✗
	Geo-localization	Message Metadata	✗	✓	✗	✗	✓	✓	✓	✓	✓
		User's Location Profile	✗	✓	✓	✗	✗	✓	✗	✓	✓
		Message Content	✓	✗	✗	✓	✓	✗	✗	✗	✗
L1 - Crowd Perception	Event/Story Detection	Clustering	✗	✓	✓	✓	✗	✓	✗	✗	✗
		Statistical	✗	✗	✓	✗	✗	✗	✗	✓	✓
	Event/Story Evolution		✗	✓	✓	✓	✗	~	✗	✓	~
	Retrieved Information	Spatial	✓	✓	✓	✓	✓	~	~	✓	✓
		Temporal	✓	✓	✓	✓	✓	✓	✓	✓	✓
		Thematic	✓	✓	✓	✓	✓	✓	✓	✗	✓
		Sentiment	✗	✗	✓	✗	✗	✗	✗	✗	✗
	Geo-fencing		✓	✗	✗	✓	✗	✗	✗	✗	✗
	Mapping to Ontology		✓	✗	~	✓	✗	✗	✗	✗	✗
L1 - Integrated Perception	Merging of Internal Sources		~	✗	✗	✗	✗	✗	✗	✗	✗
	Augmentation Using External Sources		✗	✓	✓	✓	✗	✗	✗	✗	✗
L2 - Comprehension	Automated Situation Assessment		✓	✗	✗	✗	✗	✗	✗	✗	✗
	Situation Learning		✗	✗	✗	✗	✗	✗	✗	✗	✗
	Situation -based Perception Config.		✗	✗	✗	✗	✗	✗	✗	✗	✗
	Situation Profiling		✗	✗	✗	✓	✗	✗	✗	✗	✗
L3 - Projection	Situation Projection		✗	✗	✗	✗	✗	✗	✗	✗	✗
	Impact Assessment		✗	~	✗	✗	✗	✗	✗	✗	✗
	History-Driven Projection		✗	✗	✗	✗	✗	✗	✗	✗	✗
L4 - Resource Management	Keyword Suggestions		✗	✗	✓	✗	✗	✗	✗	✗	✗
	Adaptive Keyword Selection		✗	✗	✓	✓	✗	✗	✗	✗	✗
	Outburst Detection		✗	✓	✗	✗	✗	✗	✗	✗	✗
	Configuration of Data Collection		✗	✗	✓	✓	✗	✗	✗	✗	✗
	Scalability	Distributed Processing	✓	✗	✗	✓	✓	✗	✗	✗	✗
		Indexing Schemes	✗	✗	✓	✗	✓	✗	✗	✗	✗
L5 - User Refinement	Visualizations	Maps	✓	✓	✓	✗	✓	✓	✓	✗	✓
		Timeline	✗	✗	✗	✗	✓	✗	✗	✗	✓
	Historic Situation Analysis		~	✓	✓	✓	✓	✓	✓	✗	✗
	Ad-hoc Queries		✓	✓	✓	✓	✓	✗	✗	✗	✗
	Configuration Parameters		✗	✗	✓	✗	✗	✗	✗	✗	✗
	User Feedback		✗	✗	✗	~	✗	✗	✗	✗	✗
	Crowd-Sourcing of Tasks Supported		✗	✗	✗	✗	✗	✓	✗	✗	✗
Application/ Evaluation Domain	1 Humitarian Aid Desaster Relief		1	1		1	1	1	1		
	2 social media monitoring				2						
	3 earthquake alerting									3	
	4 air quality monitoring										4

Legend: fully supported ✓ | not supported ✗ | partially supported ~

Table 1. Comparative evaluation of systems for establishing crowd-sensed crisis SAW.

towards automated SA are rarely provided, i.e., these systems lack systematic *Comprehension (L2)* functionality. W.r.t. establishing SAW, most systems focus on map-based visualizations of their extracted events or stories. Thereby, SA is left to the human user by means of interacting and adequately interpreting these visualizations. However, one exception proved that SM content can indeed be input for more elaborate reasoning: HADRian [39] performs a mapping to an ontology (*Mapping to Ontology*) and further provides full-fledged reasoning functionalities based thereupon. Operators can issue arbitrary queries in terms of this ontology, which the system attempts to answer (*Ad-hoc Queries*). Future work reports plans on supporting queries formulated in natural language, which should then be automatically compiled to the ontology. Whereas HADRian thus provides means to satisfy a human operator's *ad-hoc information need*, it does not provide a continuous, autonomous situation monitoring functionality, such as detecting situations conforming to a-priori specified situation templates, and alerting the operator as soon as these are encountered. Therefore, its functionality largely depends on the human operator's intuition and ability to formulate adequate queries.

Missing Support for Projection Level (L3). ESA represents the only system providing functionality towards supporting L3, which employs dedicated classifiers for identifying Tweets referring to infrastructural damage or expressing a need for help, which, however, are not further aggregated nor interpreted. It further demonstrates the high potential of NER approaches for gaining crowd-sensed SAW, as the Stanford NER is employed to extract names of people, organizations, locations, times and dates (*Impact Assessment*). Furthermore, related images and videos that are linked within the Tweets are extracted (*Augmentation From External Sources*). However, ESA's functionality is centered around extracting, and clustering these information on a map. No semantically enriched summary is provided, for instance w.r.t. a *Geo-fencing* of the assessed infrastructural impact, or deriving an interpreted crisis situation. The actual assessment task is thus deferred to the human operator, who needs to interpret the map-located Tweet, entity and image clusters.

Lack of Resource Management (L4). Similarly, many potentially highly beneficial *feedback loops* between the different system components remain unexploited, such as *Situation-based Perception Configuration* (L2/L4) or *Adaptive Keywords* (L4).

Adequate User Refinement Support (L5). User Refinement w.r.t. providing interactive map-based visualizations proved to be in general a well-supported criterion, which may be supplemented with tag clouds (highlighting the relative frequency of extracted keywords), as in Twitris [24], timelines (CIACM [36]), and interactive visual analytics facilities (SensePlace2 [23]). Whereas many systems supported queries and faceted search (SensePlace2) in order to modify the map-based display result, however, only HADRian provided fully semantic, automated question answering capabilities.

Lack of Learning (L1-L5). In general, the systems provide little *learning* capabilities, neither w.r.t. building up and exploiting a *crisis memory* to refine predictions (*History-driven Projection*, L3), adapting towards *User Feedback* (L5), or *Situation Learning* (L2) from the crowd, which indicates a need for research to leverage more *intelligent* crowd-sensing systems.

5. Conclusion

This paper made a first attempt towards a reference architecture addressing challenges faced when incorporating crowd-sensed information into SAW systems for crisis management. Such systems need to be capable of identifying *trustworthy*, *high-quality*, *informative* crisis-relevant SM messages by overcoming the *sparseness* of SM w.r.t. those, extracting relevant pieces of information from these (e.g., location, emerging needs, forecasts), fusing the multitude of observations to a coherent description of the event, inferring the reported crisis objects and analyzing their interrelations to form a comprehensive picture of the crisis situation, anticipating its likely development, while flexibly adapting towards the development of the crisis situation and operators' varying information need. Based on these challenges, we systematized a set of criteria, which we applied to evaluate a series of systems to identify existing coverage and open research in this field.

Whereas we could not identify a single crowd-sensing system that proved to be the most mature w.r.t. all assessed criteria, we encountered a plethora of interesting concepts realized in distinct systems: *HADRian* [39] demonstrates the potential of utilizing ontologies and semantic reasoning facilities, allowing operators to issue ad-hoc queries, based upon which the system provides answers to their current information need. *ESA*'s [43] impact classifier provides a first attempt towards *Projection* level support, by identifying Tweets reporting infrastructural damage and requests for help. *Twitris* [24], [28] shows how L4 *Resource Management* functionality can be realized w.r.t. adaptive crowd-sensing upon continuously refined keyword sets derived from the tracked storylines, and provides sophisticated algorithmic and explorative means to study the detected events' evolution over time. *Twitcident* [1] exemplifies that social media can be employed to gather further information based on an initial incident profile (*Situation Profiling*). *SensePlace2* [23] literally illustrates how *Visual Analytics* functionalities can support L5 *User Refinement*.

Thus, highly promising concepts to target the diverse challenges have been proposed, but require a comprehensive integration in a SAW system combining these, for which a component-based implementation offers a promising approach. Furthermore, whereas current crowd-sensing systems primarily focus on delivering map-based output, our discussion in terms of a holistic SAW system featuring all JDL levels revealed that embedding these approaches in full-fledged information fusion architectures presents novel means to enhance future crisis response.

References

[1] F. Abel *et al.*, "Semantics + Filtering + Search = Twitcident. Exploring Information in Social Web Streams," in *23rd ACM Conf. on Hypertext and Social Media*, 2012.

[2] N. Baumgartner, W. Gottesheim, S. Mitsch, W. Retschitzegger, and W. Schwinger, "Situation Prediction Nets," in *Conceptual Modeling – ER 2010*, ser. LNCS. Springer, 2010, vol. 6412.

[3] N. Baumgartner, S. Mitsch, W. Retschitzegger, W. Schwinger *et al.*, "A tour of BeAware – A situation awareness framework for control centers," *Information Fusion*, vol. 20, no. 0, 2014.

[4] E. P. Blasch and S. Plano, "Level 5: user refinement to aid the fusion process," *Proc. SPIE*, vol. 5099, 2003.

[5] K. Bontcheva *et al.*, "TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text," in *Int. Conf. on Recent Advances in Natural Language Processing*, 2013.

[6] L. Carver and M. Turoff, "Human-computer Interaction: The Human and Computer As a Team in Emergency Management Information Systems," *Commun. ACM*, vol. 50, no. 3, 2007.

[7] W. J. Corvey *et al.*, "Twitter in Mass Emergency: What NLP Techniques Can Contribute," in *NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, 2010.

[8] L. Derczynski and K. Bontcheva, "Spatio-temporal grounding of claims made on the web, in PHEME," in *10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, 2014.

[9] J. Dugdale *et al.*, "Social Media and SMS in the Haiti Earthquake," in *21st Int. Conf. Companion on World Wide Web*. ACM, 2012.

[10] M. R. Endsley, "Toward a Theory of Situation Awareness in Dynamic Systems," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 37, no. 1, 1995.

[11] C. Gena, "Methods and techniques for the evaluation of user-adaptive systems," *Knowl. Eng. Rev.*, vol. 20, 2005.

[12] K. Gimpel *et al.*, "Part-of-speech Tagging for Twitter: Annotation, Features, and Experiments," in *49th Annual Meeting of the Assoc. for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, 2011.

[13] S. R. Hiltz and J. J. Gonzalez, "Assessing and Improving the Trustworthiness of Social Media for Emergency Management: A Literature Review," *NISK*, 2012.

[14] L. Hong, G. Convertino, and E. Chi, "Language Matters In Twitter: A Large Scale Study," in *Fifth Int. AAAI Conf. on Weblogs and Social Media*, 2011.

[15] X. Hu *et al.*, "Multidimensional context-aware social network architecture for mobile crowdsensing," *IEEE Communications Magazine*, vol. 52, no. 6, 2014.

[16] Y. Ikawa *et al.*, "Location-based Insights from the Social Web," in *22nd Int. Conf. on World Wide Web Companion*, 2013.

[17] M. Imran *et al.*, "AIDR: Artificial Intelligence for Disaster Response," in *Companion Publication of the 23rd Int. Conf. on World Wide Web Companion*, 2014.

[18] G. Jakobson *et al.*, "Situation-Aware Multi-Agent System for Disaster Relief Operations Management," in *ISCRAM 2006*.

[19] S. Kumar *et al.*, "TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief," in *Fifth Int. AAAI Conf. on Weblogs and Social Media*, 2011.

[20] C. Li *et al.*, "TwiNER: Named Entity Recognition in Targeted Twitter Stream," in *35th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2012.

[21] J. Llinas *et al.*, "Revisiting the JDL Data Fusion Model II," in *7th Int. Conf. on Information Fusion*, 2004.

[22] J. Llinas, "A Survey and Analysis of Frameworks and Framework Issues for Information Fusion Applications," in *Hybrid Artificial Intelligence Systems*, ser. LNCS. Springer, 2010.

[23] A. M. MacEachren *et al.*, "SensePlace2: GeoTwitter analytics support for situational awareness," in *2011 IEEE Conf. on Visual Analytics Science and Technology (VAST)*, 2011.

[24] M. Nagarajan *et al.*, "Spatio-Temporal-Thematic Analysis of Citizen Sensor Data: Challenges and Experiences," in *10th Int. Conf. on Web Information Systems Engineering*, 2009.

[25] O. Ozdikis *et al.*, "Semantic Expansion of Tweet Contents for Enhanced Event Detection in Twitter," in *2012 Int. Conf. on Advances in SNs Analysis and Mining (ASONAM)*.

[26] L. Palen *et al.*, "Supporting "Everyday Analysts" in Safety- and Time-Critical Situations," *The Information Society*, vol. 27, no. 1, 2010.

[27] S. Pongpaichet *et al.*, "Situation Fencing: Making Geo-fencing Personal and Dynamic," in *1st ACM Int. Workshop on Personal Data Meets Distributed Multimedia*, 2013.

[28] H. Purohit and A. Sheth, "Twitris v3: From Citizen Sensing to Analysis, Coordination and Action," in *7th Int. AAAI Conference on Weblogs and Social Media, ICWSM 2013*.

[29] B. Rhodes *et al.*, "SeeCoast: Automated Port Scene Understanding Facilitated by Normalcy Learning," in *IEEE MILCOM 2006*.

[30] J. Rogstadius *et al.*, "CrisisTracker: Crowdsourced Social Media Curation for Disaster Awareness," *IBM Journal of Research and Development*, 2013.

[31] T. Sakaki *et al.*, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors," in *19th Int. Conf. on World Wide Web*, 2010.

[32] ——, "Tweet Trend Analysis in an Emergency Situation," in *Special Workshop on Internet and Disasters*. ACM, 2011.

[33] A. Salfinger, W. Retschitzegger, and W. Schwinger, "Maintaining Situation Awareness Over Time - A Survey on the Evolution Support of Situation Awareness Systems," in *TAAI 2013*.

[34] A. Schram and K. M. Anderson, "MySQL to NoSQL: Data Modeling Challenges in Supporting Scalability," in *3rd Annual Conf. on Systems, Programming, and Applications: Software for Humanity*, 2012.

[35] A. Sheth, "Citizen Sensing, Social Signals, and Enriching Human Experience," *IEEE Internet Computing*, vol. 13, 2009.

[36] H. Smid *et al.*, "Canary in a Coal Mine: Monitoring Air Quality and Detecting Environmental Incidents by Harvesting Twitter," in *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '11. ACM, 2011, pp. 1855–1860.

[37] I. Stavrakantonakis *et al.*, "An approach for evaluation of social media monitoring tools," in *Common Value Management Workshop CVM, co-located with ESWC2012*.

[38] B. Ulicny and M. Kokar, "Toward formal reasoning with epistemic policies about information quality in the twittersphere," in *14th Int. Conf. on Information Fusion*, 2011.

[39] B. Ulicny *et al.*, "Situational Awareness from Social Media," in *Eighth Conf. on Semantic Technologies for Intelligence, Defense, and Security, Fairfax VA, USA*, 2013.

[40] S. Vieweg *et al.*, "Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness," in *SIGCHI Conf. on Human Factors in Computing Systems*, ser. CHI '10. ACM, 2010.

[41] B. Walle and M. Turoff, "Decision support for emergency situations," *Information Systems and e-Business Management*, vol. 6, no. 3, pp. 295–316, 2008.

[42] J.-M. Xu *et al.*, "Socioscope: Spatio-temporal Signal Recovery from Social Media," in *Machine Learning and Knowledge Discovery in Databases*, ser. LNCS. Springer, 2012.

[43] J. Yin *et al.*, "Using Social Media to Enhance Emergency Situation Awareness," *IEEE Intelligent Systems*, vol. 27, 2012.