A Big Data Approach to Support Information Distribution in Crisis Response

Niels Netten, Susan van den Braak, Sunil Choenni
Research and Documentation Centre (WODC)
Ministry of Security and Justice - Turfmarkt 147
2511 DP The Hague, Netherlands
{c.p.m.netten;s.w.vanden.braak;r.choenni}@minvenj.nl

Maarten van Someren
Informatics Institute
University of Amsterdam
Science Park A - Science Park 904 Amsterdam, Netherlands
m.w.vansomeren@uva.nl

ABSTRACT
Crisis response organizations operate in very dynamic environments, in which it is essential for responders to acquire all information critical to their task execution in time. In reality, the responders are often faced with information overload, incomplete information, or a combination of both. This hampers their decision-making process, workflow, situational awareness and, consequently, effective execution of collaborative crisis response. Therefore, getting the right information to the right person at the right time is of crucial importance.

The task of processing all data during crisis response situations and determining for whom at a particular moment the information is relevant is not straightforward. When developing an information system to support this task, some important challenges have to be taken into account. These challenges relate to the structure and truthfulness of the used data, the assessment of information relevance, and the dissemination of relevant information in time. While methods and techniques from big data can be used to collect and integrate data, machine learning can be used to build a model for relevance assessments. An example implementation of such a framework of big data is the TAID software system that collects and integrates data communicated between first responders and may send information to crisis responders that were not addressed in the initial communication. As an example of the impact of TAID on crisis response, we show its effect in a simulated crisis response scenario.

CCS Concepts
• Information systems~Data mining

Keywords
Big Data; Crisis Response for Public Safety; Machine Learning; Relevance Assessments; Information Distribution

1. INTRODUCTION
Adequate and timely processing of information originating from different sources is an important determinant for handling disasters in an effective and efficient way. Disaster studies [7, 10, 11] show that incomplete information and information overload are key factors that determine the success or failure of the operational efforts and the course of crisis response. It is therefore essential for emergency responders to acquire all information critical to their task in time, especially at the beginning of a crisis response situation [12]. Lack of relevant information or too much irrelevant information hampers the emergency responders' decision-making process, workflow and situational awareness.

Crisis response aims to optimise information distribution. In large crisis response operations, an information manager is added to the operational command. This person is responsible for collecting, processing and distributing all information relevant for the common and up-to-date view of the crisis. Despite these efforts to better centralise relevant information, a gap still exists between the information supply and information needs of responders. To narrow this gap, relevant information should be distributed more effectively and efficiently between people. Therefore, organizations involved in disaster management are seeking for opportunities to exploit the large amount of complex data generated during a disaster.

Crisis response can, therefore, benefit from the surge in research and development in the field of big data [3, 4, 18]. In the literature, big data is characterised by at least three aspects: volume, variety and velocity. Volume refers to the large amount of data that needs to be stored, processed, transformed, analysed and presented. Variety refers to the different types of data that should be dealt with, such as natural language text, audio and video. Velocity refers to increasing rate at which data flows are generated. An additional, fourth, characteristic that is commonly used to describe big data is veracity, which indicates that the truthfulness of the data is often less than 100%. All four of these characteristics also apply to the domain of crisis response. The methods and techniques developed in the field of big data are, therefore, also relevant for this domain.

In this paper, we explore the possibilities of big data to improve the disaster management by crisis response organizations. We exploit big data research for supporting decision-making and improving the processing of collected data and the distribution of relevant information between first responders. Since big data can be regarded as an integration of various disciplines, concepts and paradigms, we discuss which components are required to support disaster management. These components will in turn be the building blocks for a framework that may serve as the basis for a computerized system that fulfils the central role of processing and distributing information to professionals involved in crisis response and management in time.

A computer system that fulfils such a role has to meet certain requirements. These requirements relate to 1) processing large amounts of different data, 2) being able to create meaning from this data for the purpose of disseminating relevant information, and 3) overcoming bottlenecks in the way information is communicated. Since crisis response is very dynamic, the system will also need to adapt quickly to someone's particular context. Current information

ICEGOV '15-16, March 01-03, 2016, Montevideo, Uruguay
© 2016 ACM. ISBN 978-1-4503-3640-6/16/03…$15.00
DOI: http://dx.doi.org/10.1145/2910019.2910033

systems for crisis response lack the ability to provide this type of functionality (see, e.g. [21]).

However, we have developed a big data framework that takes all these requirements into account. Subsequently, we have successfully implemented this framework in the software system called TAID (Task-Adaptive Information Distribution). We also applied the TAID system in the context of simulated crisis response. For a detailed description and analysis of the TAID system in the context of crisis response, see [1, 17].

The remainder of this paper is organized as follows: Section 2 addresses the challenges encountered when developing an information system using big data to exploit the large amount of complex data generated during a disaster. In Section 3 we describe our big data framework while in Section 4 we describe TAID. In Section 5 we show the effect TAID's information distributions can have on information sharing and activities of crisis responders in a crisis response scenario. Finally, Section 6 concludes the paper and proposes further work.

2. PROBLEM SETTING

In the beginning of a crisis response situation, responders are faced with much data about the disaster situations and they have to deal with a large number of unknowns. Acquiring relevant information is very important to take the right actions. In the next subsection we elaborate more on these issues.

2.1 The Crisis Response Domain

This subsection describes the complexity of the crisis response domain. We address the hierarchical and adaptive structure of the crisis response organization in Section 2.1.1 and the varying sources of data available in crisis response situations in Section 2.1.2. Both aspects influence the collection and distribution of information, as will be described in Section 2.2.

2.1.1 Structure of a Crisis Response Organisation

A crisis response organisation consists of multidisciplinary units, in which the main actors are firefighters, policemen, and medics. The organisation is set up ad hoc, depending on the type and scale of the emergency situation. These agencies have to complement each other's activities, while collaborating in a temporarily dynamic and high-risk environment. This type of organisation is best described as an "adhocracy" [13], since it is able to dynamically reorganise its own structure and workflow, and by doing so is able to shift responsibilities and adapt to the changing environment. Thus, the organisations' immediate purposes during the emergency shape the organisation.

The structure of crisis response organizations is hierarchical. A hierarchical structure causes people at different levels to have different information needs, since no one person needs to know everything. Information is sent down and up the organisation through intermediate actors, who are responsible for sharing relevant information with others.

During crisis response information flows via the hierarchical channels of communication. Information does not propel itself as a continuous flow, but spreads as a result of discrete communication events. For example, actors share prescriptive information (i.e. 'do this' or 'stop doing that') which flows downwards in the organisation and descriptive information (i.e. 'status report') flows upward the organisation. The hierarchical structure of the organisation determines how the information is passed between actors.

These types of communication procedures, that involve a number of steps through the hierarchy, errors can be made at each step, may lead to information delay. This delay in relevant information may lead to wrong decisions and as a consequence to a less effective collaboration.

2.1.2 Data Sources and Flow

Crisis response and management in the public safety domain is highly communication-centric. Information during such events is conveyed in different ways. In the first phase, much of the information is shared digitally from the alarm centre (i.e. centralist) to those that need to respond to the call. Responders use (mobile) communication devices and whenever possible communicate face-to-face to convey important information. Radio communication is currently a critical component of emergency response for firefighters working distributed. Each firefighter carries a radio while present at the emergency area. Radio communication is carried out using a shared vocabulary of specific, concise domain terminology. At small incidents only one channel for radio communication is used, but at very large incidents this is quickly scaled-up to multiple channels. We can say that sharing of information in current crisis response and management by responders at the operational level is predominantly orally, either through the walkie-talkie or by face-to-face communication. At the more strategic levels text messaging, emailing, or image sharing are also often used.

Other sources of data outside the professional network may also provide valuable data for the crisis response operation. For example, citizens near to the disaster situation often communicate information about the disaster via social media or news channels report about the event. All these external sources are valuable, certainly at the beginning of the operation.

2.2 Challenges

Due to the described hierarchal structure of crisis response organizations and the variety of data sources available, collecting and integrating large amounts of data in the crisis response domain involves several challenges. In this section we elaborate on some of these challenges.

2.2.1 Structured versus Unstructured Data

In crisis response, available data covers the whole spectrum from structured to unstructured [24, 25]. Structured data refers to information with a high degree of organization, such that inclusion in a relational database or spreadsheet is seamless and readily searchable by simple, straightforward search engine algorithms or other search operations; whereas unstructured data is essentially the opposite [26]. Oral communication, is an example of unstructured data. While the registration of the communication might be arranged by date, time or size, if it were truly fully structured, it would also be arranged by exact topic and content, with no deviation or spread – which is impractical, because people don't generally speak about precisely one topic.

2.2.2 Data Quality

The quality of the data is also a challenge in crisis response, since misinterpretations based on the data can lead to wrong assumptions and the execution of wrong actions [7, 10, 11, 28]. The premise is that data sources, such as professional communication networks have a better veracity of the data then external sources, such as social media data. Validating the truthfulness of information will certainly be necessary [27].

2.2.3 *Information Overload*

The sharing of information is for a large part done by the centralist who is the 'central point' within a particular emergency service that channels the information to the emergency responders. On the one hand, this centralistic approach entails the danger that in situations where multiple responders and centralists are involved, the amount of information becomes so large and complicated that information is sometimes withheld by responders or simply forgotten, in spite of its relevance for other emergency services [6, 7, 10, 11].

On the other hand, sending particular information to all responders (i.e. broadcasting it) is not a solution, as this possibly creates additional processing time for message receivers who already have little time and who often only require part of it for their tasks. Strategic personnel might become confused by irrelevant messages. Too many of these messages leads to information overload (see e.g. [9] for an overview, and [8], chapters 6 and 7 for a detailed analysis in the context of crisis response).

2.2.4 *Dynamics*

In crisis response, responders collaborate on interdependent tasks, of which many are very dynamic in nature. Therefore, responders lack the time to actively search for information that is actually available. They may not know that certain information exists and, therefore, do not search for it. For example, if a team is working at one location and is unaware of another team working nearby, they will not search for the other teams' findings or plans. As a result, information is not always available at the right place at the right time. When this information was also relevant for the first team and reaches the team at a later stadium, it caused an unnecessary delay. Consequently, it may have hampered the teams' decision-making process, workflow and situational awareness. Information is only relevant for a certain period of time depending on the responders' current or near future context.

3. A BIG DATA APPROACH TO CRISIS RESPONSE

Our approach to improving disaster management is to use methods and techniques from big data. It often remains unclear to people what big data actually is. The focus of current paradigms appears to be solely on computational and storage resources or data analytics. Instead, here we address the processing steps required for a big data system approach and propose a framework in which applications can be developed.

The framework presented contains the elements that meet the necessary requirements for supporting the distribution of relevant information in crisis response. Using this framework it is possible to process large amounts of data, give structure to incoming data of different types, and assesses the relevance of the extracted information. In the next section, system requirements for the big data framework in the context of crisis response are described. Subsequently, section 3.2 describes the big data framework for crisis response and the necessary steps within that framework.

3.1 System Requirements

An information system that supports the distribution of relevant information between crisis responders must meet at least certain important design requirements, which will be explained below. Each requirement is related to the four components of the big data framework that is described in the next section.

First, such a system will need to be able to deal with all types of multi-media information (e.g. sound, images, video, and text) that are available. Access to information from informal networks (e.g. social media) can also play a relatively big role in the information supply. Thus, to fully address the task of disseminating relevant information, a system would need to have access to all these sources of information. This requirement relates to the challenges of dealing with large amounts of structured and unstructured data and the truthfulness of the data sources used. This requirement is addressed in the data acquisition component of the proposed framework.

Second, the roles of emergency response should be key in the design of information systems for emergency management [19]. Therefore, in the data acquisition component of the framework, roles of actors are identified and a set of tasks is associated to each role. In the in the framework's data processing and data mining and analytics components this information is also used for relevance assessments and information distribution.

Third, the system must quickly adapt to the responders' changing information needs due to the dynamics of the environment, for example when responders change roles, take on new tasks, and abandon old ones. This requirement relates to the dynamic nature of crisis response and is addressed in the framework's data processing and data mining and analytics components.

Fourth, relating to the challenge of information overload, broadcasting all information is not optimal in crisis response. Therefore, an information system should support situational awareness and optimize the information flow using a selective push (i.e. 'narrow' casting) of relevant information. This kind of 'push' will also minimize the seeking time for relevant information. An information system could also be used to filter out irrelevant information but experience in other studies shows that users and organizations do not easily accept filtering [2]. This requirement is implemented in the data mining and analytics component..

Fifth, the information system requires some knowledge about which information is relevant for whom at a certain moment in time. To assess the relevance of information, some degree of understanding of the meaning of the information is required. This is implemented in the data mining and analytics component.

Sixth, as described in section 2.1.1, to be more effective in crisis response there is a need for flexible, non-hierarchical organisation and communication. This is especially the case in the early stages of a crisis when relevant information is scarce and communication channels may be damaged [7]. In those cases, an open communication flow is desirable. This requirement is built into the information distribution component of the framework.

Seventh, the system should present the extracted, relevant information in a comprehensible manner to the users. Responders should be able to process the presented information easily. This requirement relates to the challenge of dealing with dynamics and the lack of time to seek for relevant information and is part of the information distribution component.

In the next section, we show how these requirements form the building blocks for a big data framework that may serve as the basis for a computerized system.

3.2 A Big Data Framework for Crisis Response

As explained above, the characteristics of big data also apply to the domain of crisis response. Therefore, the methods and techniques of big data can be used to address the issue of distributing relevant information during crisis response. The overall process to exploit big data and gain new insights can be broken down in multiple steps, according to [29,30], which form the two main sub-processes of big data: data management and data analytics. Data management refers to the processes of data acquisition and storing data in order to prepare and retrieve it for analysis. Data analytics, refers to techniques used to analyse and acquire insights from big data.

In Figure 2 we show the proposed big data framework for crisis response with the necessary components: data acquisition, data storage and processing, data mining and analytics, and information distribution. Each step will be described separately in the next subsections. Note that compared to other big data frameworks like [29,30], in this domain an additional processing step is required after data management and analysis: information distribution.

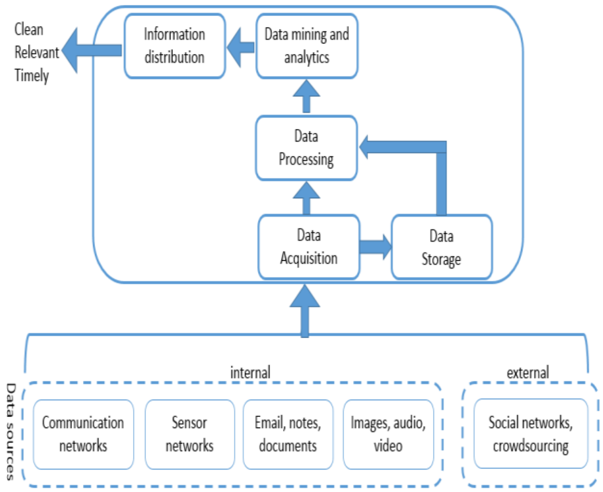

Figure 2 A Big Data framework for crisis response

3.2.1 *Data Acquisition*

The various types of relevant raw data sources and meta-data should be accessible and collected for further integration. Some data sources have to be collected once, while others have to be monitored real-time to acquire the most recent data. Data may be already structured (e.g. sensor data) or unstructured (email, documents, or oral communication). The data can be acquired internally from the crisis response organization itself or from external sources. External data, acquired from sources outside the crisis response organization, for example, from social media, usually have less veracity than the data acquired internally.

A valuable source for data are mobile devices which are used for various purposes. Besides voice communications, mobile devices are used to share information through applications, such as WhatsApp and Facebook. Recent technological developments of mobile devices have made it easier to involve citizens in collecting data, also known as 'crowdsourcing' [23]. By collecting data by means of crowdsourcing, crisis response can connect with the common mass by acquiring information quickly and learning the issues that affect citizens during disaster situations.

3.2.2 *Data Storage and Processing*

The next step in the big data framework is processing the raw data. Some pre-processing steps can be undertaken at this point before the structured and unstructured data is stored. Whereas even a few years ago a terabyte was seen as a large amount of data, today individual applications can generate petabytes of data per second. Since a lot of information in crisis response is only relevant for a certain period it has to be processed very fast, in real-time.

Once the data is collected, the usability of that data for data mining and analysis often needs to be evaluated by professionals. Annotations or additions may be necessary. In some cases this is as simple as classifying the records or adding meta tags, while in other cases it requires data to be converted into compatible formats. An example of this human involvement is that each communication record needs to be assigned to a certain role in crisis response. This must be done in an offline status of the system, a training phase in which models are built for the data mining and analysis phase.

To make the data ready for data mining purposes, some transformations are required to make them readable. In a sense, unstructured data is given structure to be useable. For example, prediction variables for data mining algorithms have to be selected before any analysis may take place.

3.2.3 *Data Mining and Analytics*

After collecting and integrating various types of data sources, the data have to be analysed to get valuable insights from the data [20]. Data mining and analytics can be used so that the data is leveraged for this purpose. In the crisis response domain, data mining is used to assess the relevance of information.

The underlying prediction model of the information system contains the rules that indicate when certain information is relevant or not for a particular actor in a particular context. These high-level rules do not include all possible factors that influence relevance for actors in these dynamic situations, but they should be sufficient to adequately distribute relevant information.

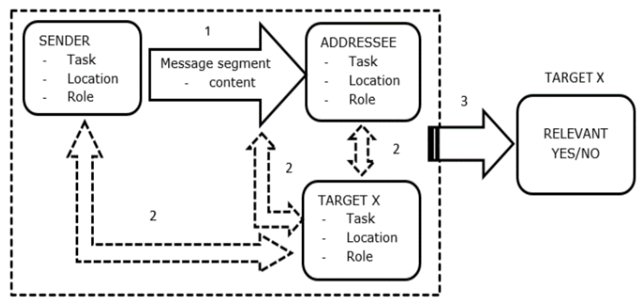

Figure 3 Conceptual model of TAIDs relevance assessment task

We take into account that the relevance assessments by data mining methods and techniques will not be full-proof. The reason for this margin of error is that crises are considered a chaotic system [16]. The term "chaos" does not mean that the system is in total disorder or will fail; rather it is a way of describing a system that cannot be predicted with full certainty (similar to e.g. weather prediction). The reason for this is that the prediction outcomes depend on the initial conditions of the system.

3.2.4 *Information Distribution*

After the data analytics part of the framework, in the crisis response domain an additional processing step is required. The data that are interpreted relevant for other crisis responders have to find their way to them. Many responders in the field will not have time to search for this information. Therefore, an information distribution functionality should be in place to locate the target device and send the information to the (mobile) information device of those responders that need the information.

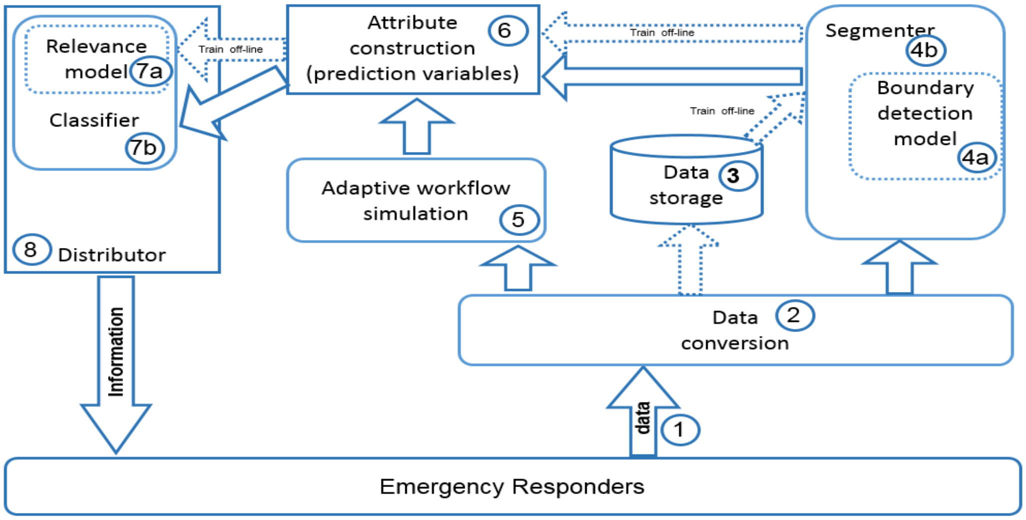

Figure 4 TAID Architecture

4. TAID

In this section we describe our software system called TAID (Task-Adaptive Information Distribution). TAID is an application of the big data framework for crisis response described in Section 3.2. In Section 4.1 we describe the specific task of the TAID system for real-time relevance assessments in crisis response. In Section 4.2 we give an overview of TAID's architecture and functions.

4.1 Relevance Assessments

The goal of the TAID system is not to prevent distribution of irrelevant information but to actively distribute relevant information. The system has to learn when communications between crisis responders are relevant for other responders who did not take part in the initial communication. To do so, it combines different types of information, including the information communicated and information about the situation.

In general, a message entails the transfer of information of a sender to an addressee. In the case of a dialogue, multiple (transcribed) speech utterance are exchanged. Within our conceptual model, these utterances are grouped together automatically, based on their topic, into one large message (i.e. message segment). The relevance of this message segment for other responders within the organisation is then assessed based on their current task, role, and location. In case of multiple addressees for a certain message, we create a message for each addressee, so that there is only one sender and one addressee for each message.

In Figure 3 we show the conceptual model of TAIDs relevance assessments task. The different steps in the model are:

1. A message has a sender and an addressee, who both have a role, task, and location.
2. At the same time that the message is sent, other involved crisis responders, who have their own roles, tasks and locations become potential targets. TAID assesses the relevance for a responder Target X by measuring similarities between Target X's current task description and that of the sender and the addressee. Geographical distances are measured from Target X to the sender and the addressee. In addition, the similarity between the description of the task of Target X is compared with the message sent.
3. Based on the similarity between Target X and those who communicated the original message (i.e. sender and addressee), the relevance or irrelevance of the message is determined for Target X.

In the next section we present the architecture of TAID that implements this relevance assessment model.

4.2 Architecture

The TAID system has multiple functions. The main functionalities of the TAID system are:

a) information distribution,
b) automatically generating (a part of) the distribution system by learning from examples during a separate training phase,
c) simulating the behaviour of (a) + actors + their environment to determine (current) situational information of actors.

TAID consists of a online and offline mode. In the online mode TAID actively monitors the communications between crisis responders, assesses if information shared is relevant for others, and if so, disseminates this information immediately to those crisis responders. To assess relevance it uses real-time information about the involved crisis response actors and their context.

However, before TAID's information distribution functionality can operate online, it must first automatically generate a model for relevance assessments that will be part of its information distribution function. To build this model the system must be trained with data examples that have labels, usually added by domain experts, to indicate which information is relevant for whom. These examples have to constructed beforehand from the collected data based on attribute selection methods and techniques. Based on a set of these examples a '*relevance model*' is build (see Figure 4 <7a>]). This phase is TAID's offline phase.

TAID is setup as an open software system that can easily be adapted to run as an application on most modern communication systems. This is because the system needs to be able to function on any modern cell phone and other communication systems used by professional crisis responders.

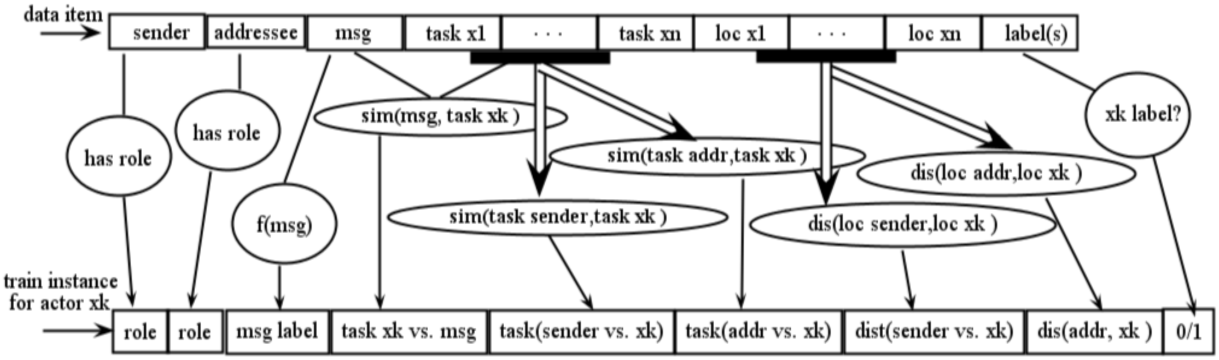

Figure 5 Integrated data item to instance for machine learning

In Figure 4 we show the (software) architecture of TAID. The architecture consists of several components. Each of these components relate to the components of the big data framework presented in Section 3.2 and will be described in the next subsections.

4.2.1 Data Acquisition

For TAID the focus is largely on unstructured data such as oral and textual data (like email or texts) communicated between crisis responders. The available situational information consists of task descriptions and geographical locations. Task descriptions are textual unstructured data, whereas the geographical information involves sensor data.

The data conversion component (see Figure 4 <2>) converts the raw communication data (see Figure 4 <1>) of the responders into a useable format. This component transcribes oral communications into text. It receives input from a central communication system. This can be a central communication server, but can also be a distributed peer-to-peer system with proxy servers that transmit information exchanged locally to a central system. In this way, also the senders and receivers of communications can be identified. TAID also assumes that information about the situation is available, in particular the locations of responders and their current tasks.

4.2.2 Data Storage and Processing

All the input data (i.e. all communications between responders and all situational information) is logged and stored in the data storage component (see Figure 4 <3>). In the offline phase of the system this is the place where valuable meta-data is added to the message. These include labels indicating the (ir)relevance of information for other responders at a particular time. These labels are added by domain experts.

In the next subsections we describe TAIDs approach in processing two different (real-time) streams of incoming data. This includes the communications between responders and their fast-changing situational information.

4.2.2.1 Communications

Recognizing relevant information in dialogues involves a combination of segmentation, i.e., identifying coherent segments, and classification of these message segments as relevant or not for a responder. These dialogues are continuous streams that can involve multiple changing participants introducing various topics.

The basic segments of dialogue discourse are utterances. In a dialogue turn a responder will utter some words. In crisis response these utterances in a dialogue turn are often very short. Grouping coherent utterances creates a larger segment that contains more information. A larger segment was used to improve the ability of the relevance model to better determine relevance of the information, since there is more to assess relevance.

The segmenter component includes an offline phase for learning a model to detect boundaries in a stream of dialogue communications (see Figure 4, <4a>). During training, labelled communication data collected from the crisis response domain is used. This input is at the level of utterances in a dialogue turn. For a more detailed description of the boundary detection approach see [5, 17]).

Online, the segmenter uses the build boundary detection model to detect boundaries (see Figure 4,<4b>). The incoming dialogue input is passed on by the data conversion component (online) or from the data repository (offline). When it identifies a boundary, it groups all the preceding dialogue utterances to a larger segment (i.e. information units) and passes the segment on to the attribute selection and construction component (see Figure 4 <6>).

4.2.2.2 Situational Information

Workflow information (i.e. about tasks and locations) is important for the TAID system to assess the relevance of information and to adapt this to the fast-changing information needs of responders. The TAID framework has a separate component (see Figure 4, <5>) that exploits this information. This component implements (adaptive) workflow simulation. It provides the necessary situational information to the responder based on the adaptive workplan, which is deduced by the modelled adaptive workflow simulator (AWS). The AWS is fed with the extracted data (see Figure 4, 2) and is used as the enabling condition that triggers certain activities of simulated agents (i.e. simulated workflow of emergency response actors) within the AWS. The full range of functionalities and usage of such an adaptive workflow simulator for emergency response is described in [10].

4.2.3 Data Mining and Analytics

As it is nearly impossible to define and maintain all relevance assessment rules for general crisis response by hand, TAID uses automatic techniques from the field of machine learning. Machine learning methods are capable of solving these types of complex tasks by coming up with their own program (i.e. model) based on examples. TAID builds a relevance model that was set up as a classification task in which input information and knowledge about (ir)relevance of information was used.

Before machine learning techniques can be applied the data has to be transformed into a more structured format. This requires attribute selection and construction as will be described below.

4.2.3.1 Attribute Selection and Construction

The attribute selection and construction component (Figure 4, <6>) transforms a segment, which it receives from the segmenter component, into classifier attributes (i.e. features). These attribute are used for learning and assessing relevance. The basic attributes adopted from a segment are the name of the sender and receiver(s). The most informative attributes come from the message content and are selected using pre-processing methods. Additionally, derived attributes pertaining to the task descriptions and locations of responders are added. For example, the value for the distance attribute is derived by using the location of the different responders.

The attribute selection and construction component is used in both the training (offline) and operational (online) mode. In Figure 5 the transformation from data items to machine learning instances is shown. Similarities between unstructured text (sim[x,y]) and geographical distances (dis[x,y]) as well as a function (f[msg]) and relations (has role) are used to derive the attributes. In this step, data items are transformed to attributes that abstract from the location-specific data, to make it applicable to different crisis response situations.

4.2.3.2 Relevance Classifier

The classifier is the component that actually assesses the relevance of crisis response communication. To assess relevance a supervised classification task is used (see, e.g. [22]). As input, the classifier receives data from the attribute selection and construction component. In the training (i.e. offline) phase the classifier learns to build a model for relevance (Figure 4, <7a>). The model learns the relevance of a data instance for one or more actors A1, A2 . . . An, who are used as the class labels. In the online phase (Figure 4, <7b>) the classifier uses the learned relevance model for assessing relevance for newly received data instances (i.e. automatically assigns a 'yes', or 'no' label to the data instance).

The multi-label classification task was transformed into multiple binary relevance classifiers, one for each responder role. Naive Bayes was used as a learning algorithm, since it scored best in terms of predictive performance on the available data sets.

4.2.4 Information Distribution

The function of the information distribution component is to decide if a segment of information is to be passed on to responders for whom it is assessed as relevant (see Figure 4, <8)>). It can be used in different modes and it can easily be incorporated in most communication systems as an additional function. No training or special working procedures are required. Only if an organization wants to impose a particular policy on information distribution, then this must be imposed by introducing human control, for example by an information manager.

Our approach was to let the distributor pass on a segment (i.e. a coherent message) automatically. However, user-controlled options for passing on the information are also feasible. The output of the distributor is a segment of dialogue in unstructured text format or optionally, in the case of voice communication data, the original speech exchanges.

5. THE IMPACT OF TAID ON CRISIS RESPONSE

In this section we show the impact of information distributions performed by TAID on crisis response. To measure the impact of TAID some experiments were conducted in which TAID was applied to a simulated crisis response operation. This crisis scenario was based on real events from a large crisis response exercise and was modelled using a modelling and simulation tool.

Our approach to evaluating the effect of TAID's 'relevant' information distributions focusses on the time spent by the crisis responders on their response activities and communication of information. In Section 5.1 we describe how we simulated crisis response and how this was used to measure the effects of TAID. The measurements include timely information access and efficient and effective response activities of simulated responders. In Section 5.2 we present some of the results. For a detailed description and analysis of the effects of TAID system in the context of crisis response see [17].

5.1 Simulated Crisis Response

Our approach requires a simulation environment in which we are able to model specific behaviours of individuals that interact with each other and their environment explicitly, and to analyse their emergent behaviour. This environment needs to have detailed modelling capabilities of agent activities and communications, and the 'relevance' of information for a crisis responder, including the expected behaviour after acquiring such information. The simulation must provide insight into the execution and duration of agent activities, agent decisions and communications during the scenario. The software tool Brahms meets these modelling and simulation requirements [14] and was, therefore, used for our experiments [15].

5.1.1 Scenario Modelling

To measure the impact of TAID, we simulated parts of a large crisis response exercise. In the exercise, multidisciplinary responders had to deal with a large train accident near a densely populated city centre. A passenger train collided with a truck at a railroad crossing adjacent to a large construction site. The collision caused the train to derail and fall into a large construction pit. In the simulation, we remained as close as possible to the activities of the responders, the real communications, and the events that occurred during the exercise. The responders were modelled as software agents with tasks and they communicated to share information. The information communicated between agents and their work context was shared real-time with TAID. This simulation was used to measure the effect of TAID on some time variables that are explained below.

5.1.2 Measurements

To get a precise insight into the effects of TAID, we measured the time it took for agents to start and execute their tasks and the time it takes the agents to finish all their work. The first measurement indicates whether actors started activities earlier or later than in the original scenario, latter whether the activity of the actor ended quicker or later. This second measurement is needed, because sometimes an actor may start an activity earlier but because of the interdependence with the task of another actor the task is still completed at the same time.

In short, in the simulated experiment we measured:

- I. The quality of the relevance predictions;
- II. The time from the first mention (FM) of a problem until it reaches the target agent(s);
- III. The time from the first mention (FM) until the problem is solved;
- IV. Duration of responder agent activities;
- V. The effect on the amount of time actors take to finish their work.

For points II and III, the agents are faced with an immediate problem, for example confrontation with a large fire. To solve the problem (i.e. extinguish the fire) the agents have to execute their tasks. The execution of the tasks, and hence the resolution of the problem. This can be accelerated by means of other agents, who because of receiving information from the TAID about this problem can act sooner. Four of these situations occurring in the simulation were identified and monitored (M1 t/m M4), from which we follow the first mention of the message. Next, for IV and V we compared the start and finish time of activities in both simulation runs for the involved agents. The TAID relevance distributions can lead to earlier execution of a task, which may affect the duration of the whole emergency response scenario.

5.1.3 Evaluation Methodology

The quality of the relevance distributions is measured using precision (i.e. if TAID makes correct relevance decisions and does not send too much irrelevant information). For each message segment, assessed by TAID, we analysed whether TAID's prediction was correct. This was done through matching the information and the relevance for the responder agent at that exact moment. The messages that were assessed as relevant by TAID (i.e. true positives) were compared to the messages that were sent, but appeared to be irrelevant (i.e. false positives). This was done manually.

To analyse how long it takes for information from the first mention to reach the target agent, we view the communication structure as a graph. The nodes of the graph represent the agents who take action and the edges the communications between the agents with a probability distribution. The probability at an edge indicates the probability that the information is communicated to the next agent. For simplicity we assume that all actors in our simulation always pass the information on to the other actors.

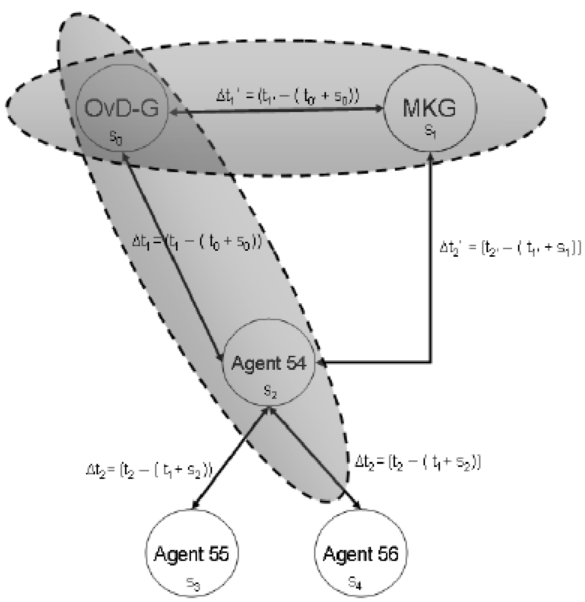

Figure 6 Communication graph

Figure 6 shows that each edge represents the direction in which information is passed along to other actors. Each edge has a certain duration (Δt). The calculate the for value Δt we distinguish between two durations: 1) the time the information needs to transfer from one actor to another (t_i) and 2) the time the information stays at some intermediate actor (actor state time s_i) before it is passed along. In the equation, we add the latter duration s_i to the last moment of receiving the information t_{i-1}. The beginning values, s_0 and t_0, are both zero. The path the information follows accumulates to a total duration. This is the duration from the first mention to reach target actor F_n, where n is the actor who acts upon the information. Thus, F_1 is the time it takes from the first mention of the information to reach the addressee 1.

5.2 Results and Effects of TAID

In this section, we present the results of the experiments for measures I through V. In total there were 32 grouped message segments in the simulation that were assessed by TAID for information relevance, for each of the 15 emergency responder agents involved in the scenario. For detailed results of all 15 responders see [17], Chapter 8.

5.2.1 Quality of the Relevance Predictions

The number of messages predicted as relevant by TAID was broken down into the number of messages sent correctly (i.e. true positives) and sent wrongly (i.e. false positives). This involved a manual analysis of the content of the sent messages in relation to the relevance of this information for the receiver.

Our results in Table 1 present (1) the relevant and (2) irrelevant predicted message segments assessed by TAID for 4 simulation agents. The number of relevant predicted messages was broken down into (3) number of messages sent correctly (i.e. true positives) and those (4) sent wrongly (i.e. false positives). This is shown under the header 'True/False positives'. This was a manual analysis of the content of the sent messages in relation to the relevance of this information to the receiver.

Table 1 Qualiy of predictions

	TAID		True/False positives	
Agents	(1) +	(2) -	(3) +	(4) -
OvD-B	12	20	1	11
OvD-G	21	11	12	9
OvD-P	11	21	5	6
ProRail	19	13	12	7

We observed that the dialogue communications grouped by the segmenter are reasonably good. The coherent topic in a dialogue was in many cases captured in the message. However, there were also some useless messages. These messages consist of only a dialogue ending, such as 'Thank you' or 'we will do'. The segmenter treated them as separate messages. The most common cause of such messages was the included time limit in the boundary detection model. When the elapsed time between two consecutive dialogue communication takes longer than 6 seconds, the segmenter considered the new dialogue communication not to belong to the previous one. Despite the lack of any information in such a message, the relevance classifier determined these messages relevant for some of the actors, probably because of the other properties taken into account, such as the sender role. In the simulation, these messages did not contribute any real information to the target agent and are by default not relevant for them. They should be detected and filtered out.

5.2.2 Time Reductions

We measured all the time differences of the TAID scenario compared to the standard scenario. We observed mainly local time reductions. There was a minimal speed up of the information reaching the target agents, with just a few of seconds time reduction. Only OvD-ProRail got the information (M1) much sooner, with 108 seconds saved. The resulting speed up for FD-111 (firefighter) was unfortunately diminished because of the evacuation task, which had a higher priority. In case of M2 agent MD-55 and OvD-G (officer on duty medics): showed no time difference. Agent MD-56 (medic), on the other hand, who was not notified about the information in M2 in the original scenario, now received this information indirectly via the OvD-G, who was presented with the information by TAID.

For IV and V OvD-ProRail and MD-56 are again the outliers who finished their tasks much sooner, with 141 and respectively 540 saved seconds. The other agents used approximately the same amount of time.

Unfortunately, TAID was not able to solve all four major information flow errors in the scenario. For example, information about a backpack threat was not detected as relevant for the medical service agents. Although it was ultimately a false alarm, these actors were still exposed to a risky situation because of the lack of this information. On the other hand, TAID did not hamper their task execution as it did with the firefighters that were aware of the threat. The stabilisation with scaffolds and information about the casualty collection points got corrected thanks to the TAID.

6. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

In this paper we have shown how big data techniques and methods can be applied to the crisis response domain to collect, integrate and exploit vast amounts of disparate data at the right speed. The described Task-Adaptive Information Distribution (TAID) system is an application of big data in the context of crisis response. TAID focuses on assessing the relevance of information in real-time and disseminating it to the right people in time. To adapt to the fast-changing information needs of crisis responders it uses information on their task and location.

In a simulated crisis response scenario, TAID showed that it is capable of providing a small positive effect on the operational activities of the responders, with little costs. It also mitigated the impact of some of the major information flow errors present in the original scenario (without using TAID). Although TAID's predictions were not yet optimal, it was able to detect and efficiently forward several relevant messages to other agents not taking part in the dialogue discussions. In this way, it was responsible for raised situational awareness for some responders in the scenario, who decided to take other actions than in the original scenario. In some cases this resulted in more effective and efficient operational activities. Consequently, in some parts of the scenario TAID led to less damage and risk, and less severe injuries for those involved.

One of the social implications of using a big data application, like TAID, is that people will need to share (more) data about themselves to these applications. In case of TAID, what they are doing and what they are communicating to others. In turn, TAID can support better insights into the dissemination of relevant data and also prevent people from making mistakes based on either insufficient data or information overload.

However, even if TAID provides superb communications, there is no guarantee that no mistakes will be made. The real impact of the distributed messages by TAID depends on how responders deal with the new information. Thus, the decision making process of responders plays an important role. High priority messages usually have an immediate effect on task execution, and the sooner in the mitigation process the message is received, the higher the impact on the course of actions is.

Currently, TAID focuses on textual information and text converted from speech. However, relevance of information in another format, for example, images, could also be used by the TAID system. Based on features extracted from an image, taken at the incident scene, TAID could learn to classify images for team members for whom they are determined relevant. The decision process of responders would then not only be supported by language, but also by visuals.

The technology of TAID is also suitable for applications outside the domain of crisis response, for example, in military operations and crowd control. With the growth of data from smart devices TAID could also be used for example in the context of 'smart cities'[23]. TAID could also be applied to social network applications like Facebook or LinkedIn. These applications know where other people from an individual's network are and what they are doing. Thus, relevant communications between two persons in this social network could easily be disseminated to others that are dealing with a similar situation.

7. ACKNOWLEDGMENTS

The research related to the TAID system was supported by the Dutch Ministry of Economic Affairs under grant MMI04006.

8. REFERENCES

[1] Netten, N. and Someren, van, M. (2011). Improving communication in crisis management by evaluating the relevance of messages. Journal of Contingencies and Crisis Management, 19 (2):75–85.

[2] Cramer, H., Evers, V.,Someren, van, M. and Wielinga, B. (2009) Awareness, Training and Trust in Interaction with Adaptive Spam Filters, CHI'09, Boston, USA.

[3] Agrawal, D., Das, S., and El Abbadi, A. (2011). Big data and cloud computing: Current state and future opportunities. In Proceedings of the 14th International Conference on Extending Database Technology, EDBT/ICDT '11, pages 530–533, New York, NY, USA. ACM.

[4] Lohr, S. (2012). The age of big data. New York Times, 11.

[5] Netten, N. and Someren, van, M. (2008), Identifying and Routing Emergency Response Dialogue Segments, Proceedings of the 5th International ISCRAM Conference, George Washington University, Washington, DC.

[6] Reddy, M.C., Paul, S.A., Abraham, J., McNeese, M., DeFlitch, C. and Yes, J. (2009), 'Challenges to Effective Crisis Management: Using Information and Communication Technologies to Coordinate Emergency Medical Services and Emergency Department Teams', International Journal of Medical Informatics, Volume 78, pp. 259–269.

[7] Helsloot, I. (2005), 'Bordering on Reality: Findings on the Bonfire Crisis Management Simulation', Journal of Contingencies and Crisis Management, Volume 13, pp. 159–169.

[8] Bruinsma, G. (2010), Adaptive Workflow Simulation of Emergency Response, PhD Thesis, University of Twente.

[9] Eppler, M. and Mengis, J. (2004). The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. The Information Society, 20:325–344.

[10] Scholtens, A. and Drent, P. (2004). Fire in King's church in Haarlem the Netherlands. (In Dutch: Brand in de Koningkerk te Haarlem) Technical report, Inspectie Openbare Orde en Veiligheid (IOOV).

[11] Vollenhoven, van, P., Castricum, F., Elias, J., Brouwer-Korf, A. and Dragt, D. (1996). Hercules disaster Eindhoven, The Netherlands. Technical report. Council of Transportation and Security.

[12] Kapucu, N. (2006). Interagency communication networks during emergencies: Boundary spanners in multiagency coordination. American Review of Public Administration, 36(2):207–225.

[13] Mintzberg, H. (1983). Structure in fives: Designing effective organizations. Patience hall, Englewood Cliffs. N.J.

[14] Sierhuis, M. (2001). Modeling and Simulating Work Practice. PhD thesis, Universiteit van Amsterdam.

[15] Bruinsma, G. and Hoog, de, R. (2006). Exploring protocols for multidisciplinary disaster response using adaptive workflow simulation. In Walle, B. v. d. and Turoff, M., editors, Proceedings of the Third International Conference on Information System for Crisis Response and Management (ISCRAM), pages 0–0, Newark, New Jersey. ISCRAM.

[16] Ditto, W. and Munakata, T. (1995). Principles and applications of chaotic systems. Communications of the ACM, 38 (11): 96 – 102.

[17] Netten, C.P.M. (2015) Machine Learning for Relevance of Information in Crisis Response, PhD Thesis, University of Amsterdam.

[18] Mayer-Schönberger, V., and Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

[19] Turoff, M., M. Chumer, Walle Van de, B., and Yao, X. (2004). The design of a dynamic emergency response management information system (dermis). Journal of Information Technology Theory and Application, 5(4):1–35.

[20] Witten, I. and Frank, A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco, 3rd edition.

[21] Otten, J., Heijningen van, B., and Lafortune, J. (2004). The virtual crisis management center: an ict implementation to canalize information! In Proceedings of Information Systems for Crisis Response and Management (ISCRAM) Brussels, Belgium.

[22] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1–47.

[23] Choenni, R., Bargh, M.S., Roepan, C. and Meijer R., (2015) Privacy and security in smart data collection by citizens, in Smarter as the New Urban Agenda: a Comprehensive View of the 21st Century City, edited by J.R. Gil-Garcia, T.A. Pardo and T. Nam, Springer New York.

[24] Malan, D., Fulford-Jones, T., Welsh, M., and Moulton, S. (2004). Code blue: An ad hoc sensor network infrastructure for emergency medical care. International Workshop on Wearable and Implantable Body Sensor Networks.

[25] Kim, S., Jang, Y., Mellema, A., Ebert, D. and Collins, T. (2007). Visual analytics on mobile devices for emergency response. In Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on, pages 35–42.

[26] Braak, van den, S., Choenni, R, Meijer,R. and Zuiderwijk, A. (2012) Trusted third parties for secure and privacy-preserving data integration and sharing in the public sector, Proc. DG.O 2012, 13th Annual Int. Conf. on Digital Government Research, Maryland, US, June 4-7, ACM Press, pp. 135-144.

[27] Droege, S. B., Anderson, J. R., and Bowler, M. T. (2003). Trust and organizational information flow. Journal of Business and Management, 9(1):45–61.

[28] Abbink, H., Dijk, R. V., Dobos, T., Hoogendoorn, M., Jonker, C., Maanen, van, P., Popova, V., Sharpanskykh, A., Tooren, P. V., Valk, J., Xu, L., and Yolum, P. (2004). Automated support for adaptive incident management. In Proceedings of the First International Workshop on Information Systems for Crisis Response and Management.

[29] Labrinidis, A., Jagadish, H.V. (2012) Challenges and opportunities with big data. Proceedings of the VLDB Endowment, 5 (12), pp. 2032–2033.

[30] Mohanty, H.,Bhuyan, P. and Chenthati, D. (2015). Big Data: A Primer. Springer Publishing Company, Incorporated.