Focused Data Mining for decision support in Emergency Response Scenarios

Sam Chapman and Fabio Ciravegna

Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello Street,
Sheffield, S1 4DP, United Kingdom
{s.chapman, f.ciravegna}@dcs.shef.ac.uk

Abstract. This paper introduces the growing emergency response domain where there is a strong need to collate structured information to aid in decision-making. Semantic Web and natural language technologies can aid this process by providing the ability to perform focused mining of unstructured data to create a timely structured data repository. A focused use-case is detailed along with a working system (Armadillo e-Response) demonstrating how this can be employed in a real world application.

Keywords: E-Response, Decision Support, Semantic Web, Data Mining, Geopositioning, Natural Language Processing, Human Language Technologies, Web Intelligence, Focused Information Retrieval.

1 Introduction

Resilience has become a hot topic with recent western political climates: The term 'resilience' has previously described individuals with the ability to withstand or recover easily and quickly from illness or hardship, but this has been expanded to encompass organizations and even societies. In order to build resilience into society it has become necessary to formalize plans, systems, training and flexible architectures for the management of unknown catastrophic emergency events.

Many governments and even corporate bodies plan for resilience from differing perspectives in order to ensure that the country or corporate entity they represent can handle and recover quickly from any emergency, either previously perceived or entirely unknown, such as a major flood, terrorist attack or industrial accident.

Regardless of the cause of any emergency, available resources (employees, police, medical, fire, military, experts, volunteers, equipment etc) must be used with the maximum efficiency to deal with arising issues. In order to coordinate such a response it is hugely important to have a timely supply of relevant knowledge, delivered to a central point of hierarchical management. Currently generic information such as geographic plans, lists of potential contacts and available medical facilities are collated manually in advance of any event to be used later when required for managing any emergency response. Manually collated knowledge such as this is typically high level and coarse grained due to the cost of its capture and organisation.

There is however a strong demand from decision makers to be able to view precise information at specific levels of detail to facilitate the real needs of an emergency response when focusing upon an incident.

This paper attempts to outline methods in which semantic technologies, automated web mining and human language technologies can help to provide timely, versatile and searchable knowledge for this need. Armadillo e-Response is introduced, as a system that collates information in real time to aid responding to emergencies. Further details about the

architecture and the interface of the system are then presented.

2 Background work

Most of the existing work in the emergency response field has been focused on producing structured feeds and repositories such as list of potential mortuary facilities or definition of marshalling zones. All these structured feeds are usually pre-compiled, trying to get an overview of possible relevant information in order for the emergency forces to be prepared. As an example londonprepared[1] is a website that contains publications about various emergency measures and planning. These resources can be useful if they really match the disaster, but in case of an unplanned emergency they cannot provide adequate information and the cost of manually gathering such information and keeping it up to date is high.
When considering building automatic systems for aiding the emergency response sector, it is important to take into account previous work and research conducted in similar or complementary areas, such as automatic systems that use NLP technologies to generate maps extracting content from news feeds [1].
Other systems focus on associating text indexing with spatial indexing methods to categorise web documents with respect to the geographic location [2]: although this is a sophisticated approach is very time consuming and I focuses on the disambiguation problem more than on collecting quickly information about a disaster area. When a disaster occurs, it is usually in a limited geographical area, where the problems of disambiguation are less a concern.
Automated methods have been previously investigated to construct structured data, such as UK Ordinance Survey that uses photographs to automatically extract information about buildings [3]. This system relies upon satellite telemetry and on the availability of postcode data. The structured data produced is limited and highly focused, but is not integrated with information available from the web.
The Armadillo e-response research project is part of a larger consortium (AKT) that aims to develop a set of tools for providing timely and relevant information for improving disaster management [4, 5, 6, 7].

3 Use Case

Throughout the world centralised planning exists to handle emergencies with largely similar organisation: in the United Kingdom a number of centralised bodies perform this role. In London, emergency events are strategically coordinated by the Joint Emergency Services Control Centre (JESCC). This is a hierarchical command structure which in times of emergencies strategically commands the Police, Fire Brigade and Ambulance Service Control/Command Units, together with the public utilities and local authority(s) within which any event occurred.
Within strategically important capital cities organisations such as JESCC can be maintained in a state of readiness; however time is still needed to move to a relevant base for operation and switch management to the command centre. Initially the JESCC and the police define three cordons which are managed by the police in cooperation with other authorities, whilst the JESCC sets up its central control point.

Inner - provides immediate security of the hazard area and potential crime scene.
Outer - seals off an extensive area around the Inner Cordon.
Traffic - set up at or beyond the Outer Cordon to prevent unauthorised vehicle access to the

[1] http://www.londonprepared.gov.uk/

area surrounding the scene.

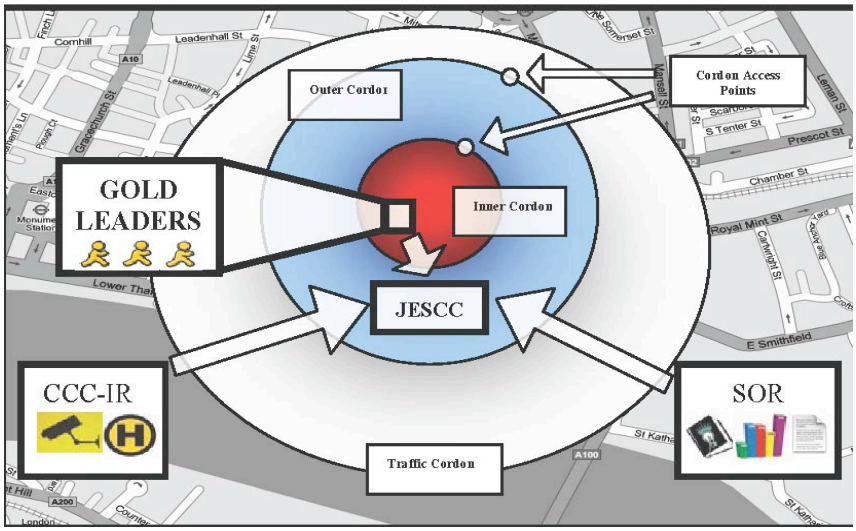

Fig 1 - Details the Cordon and information flow to the JESCC in the event of an emergency

JESCC operates outside of the inner cordon but at a point of ease where communication can be coordinated between the previously described bodies. In the event of an emergency, JESCC typically takes approximately one hour to assume full centralised control of the situation. The running of the JESCC [8] and equivalent bodies require timely information in order to make fast and appropriate decisions. Information is provided by three main means (see Fig.1):

1. Ground reports from **Gold Leaders** (appointed leaders of each emergency service, utility and local authority at the scene).
2. External video feeds provided by the **CCC-IR** (Central Communications Complex – Information Room). The **CCC-IR** provides live video feeds from any relevant security camera or police helicopter.
3. Relevant Intelligence via the **SOR** (Special Operations Room). The **SOR**'s function is to monitor disasters, terrorist incidents, disorder and large demonstrations. It combines information streams specifically for the ambulance, police, fire and military amongst others. It contains various service liaison officers who relay vital information and requests. Although the **SOR** does not control the incident, it provides information and support to the decision makers based in JESCC.

JESCC are increasing reliant upon background information the SOR provides. Currently the SOR is limited to providing structured information that is manually precompiled or manually located at the time of need. For instance the location and student population and term times of schools are structured and easy to retrieve from existing records. Determining the number of potential employees within a corporate building at 8pm is instead harder. The scarcity of specific low level information being readily available has a strong need concerning urgently needed knowledge concerning the inner cordon.
To demonstrate the utility for automatically structuring such low level information an example disaster is discussed in section 4.3. First an outline of the Armadillo e-Response is presented.

4 Armadillo e-Response

Armadillo E-Response is a system that aims to aid the SOR in gaining timely information regarding geographically based information regarding building usage, in order to better support the JESCC. Such a system is not restricted to SOR usage as it could have many external potential uses, i.e. wherever it is required to extract timely information from geographically relevant websites. In particular Armadillo e-Response attempts to focus upon automatic extraction of information from web material referring to the geographical area desired (in this case information within the inner cordon specifically).
When an emergency occurs the system starts automatically mining the web looking for relevant information for that area: it can start either by a user manually inserting the target area's geographical co-ordinates or automatically from a request from another application or process. The knowledge is extracted using an incremental approach, building a knowledge base for the immediate area surrounding the emergency.
The whole system is centred around an ontology which defines the disaster domain; when resources are identified all findings are stored within a shared storage, in this case a triple store of the assertions found. A series of independent services are then fired, each of them contributing to the knowledge harvesting process by investigating the geographical co-ordinates and expanding the radius as time progresses. All the processes act concurrently but some are more intensive (thus taking longer time) and expand slower than more rudimentary processes that provide basic information. This incremental method builds knowledge in a timely manner focusing firstly around the emergency area.
When suitable resources are found, they are locally cached, both to increase the speed of the system and above all to avoid network instability problems; in fact, as soon as an emergency occurs, access to web resources can become unstable and very busy. Related resources are also classified and, if considered relevant, useful information like contact names or telephone numbers are extracted.
The full architecture of the system is now detailed, followed by a brief description of the user interface.

4.1 Architecture

The architecture of the emergency response web extraction system resolves around a number of service components. The use of service components was selected to make the system flexible allowing extensions and customisations to suit the needs of the users. Each service component is distributed over a number of desktop PCs each connected via a high bandwidth connection. The service components are now detailed:

- **EmergencyTriggerService:** In the event of an accident, emergency or event this service is called to begin the focused information-gathering task. This works by triggering a number of services which all interact via semantic data (RDF triples) stored within the repository accessed via the TripleStorageService.
- **TripleStorageService**: This a simple interface to input and query (via SPARQL queries) a semantic triple store, in this case a locally stored version of the 3store[2] specifically for this task.
- **PostcodeDataService**: This service converts basic postcode data into RDF for submission via the TripleStorageService for the location desired (see above). For the purpose of the demo application postcode data is limited to a section of London due to data licensing constraints, in a full model this can apply to anywhere in the UK or even worldwide, given web resources for the area involved.

[2] http://triplestore.aktors.org

- **LocationCentredSearchService**: Provides localised graphs, matching latitude and longitude data obtained from the TripleStorageService with those provided by the PostCodeDataService.
- **External Web Search Services**: Service that performs focused queries on Internet Search Engines, Google, MSN, Yahoo for example.
- **URIFinderService**: service that, using multiple search services (see above), returns a list of URIs for a given piece of knowledge in the graph (in this scenario postcodes are used to find potentially geographically relevant URIs). The found URIs are then stored as provenance for each postcode in the graph.
- **DocumentCacheService**: a service that caches documents from a URI and stores then locally accessible in a fast access cache, as the local look up- address is stored via a submission to the TripleStorageService.
- **URICrawlerService**: service that, using content from the DcoumentCacheService, crawls a URI providing link URIs to a given depth and scope (within this scenario links are explored to a depth of a single link to constrain the search and make it faster. This obtains only pages within the scope of the host URI). All linked URIs are added into the knowledge store using the TripleStorageService.
- **URIClassifierService**: A previously trained classifier which assigns classification to URI content using the DocumentCacheService. The classifications are input into the TripleStorageService, along with a confidence rating (confidence of the classification algorithm). This classifier is independently trained upon classifications of interest within the emergency domain. The text classification uses a standard bag-of-words representation, applying a stoplist and (Porter) stemmer. Indexing terms are selected via a minimum threshold for document frequency. Documents (web-pages) are classified using cosine-vector similarity based on the TF-IDF values [9].
- **FastNameExtractor**: A service that uses rudimentary yet fast NLP techniques for extracting individual names from documents within the DocumentCacheService, storing triples of discoveries using the TripleStorageService. It starts its process by extracting all the potential names from the webpage, using regular expressions to spot capitalised word bigrams or trigrams. Then uses a name gazetteer to act as a validator. This is a weak approach but one designed for the time constraints of the E-Response scenario.
- **FastTelephoneExtractor**: A service that uses rudimentary yet fast NLP techniques for extracting contact telephone numbers from documents within the DcoumentCacheService, storing the found triples using the TripleStorageService. Again, it uses a fast and simple regular expression to extract telephone numbers from within the page text.

A central component of the architecture is the LocationCentredSearchService, this provides the mechanism to centre the search from existing triples held in the TripleStorageService, so that services can centre there search upon the most relevant information for that area, i.e. as close as possible to the centre of the emergency. This approach could be expanded to include additional focusing i.e. using the user focus to drive the extraction to the viewed location.
One result of this centralized geographic approach is that the amount of information retrieved by the different services expands as time progresses, fast expanding out from the centre of the emergency, thus constantly improving the knowledge base and also providing more and more sophisticated information. The services are orchestrated to provide the knowledge accordingly to a timeline (see Fig 2); for a given location they operate independently depending upon available data and processing speed:

1. A triple store is prepared by wrapping ordinance survey data regarding UK postcodes[3]

[3] As provided by Ordinance Survey.

2. Using the UK postcodes only, Armadillo [10] is used to perform a focused crawl into web resources concerning the geo graphical area.
3. Armadillo provides a list of URLs worthy of investigation.
4. Found URLs are crawled within a domain limited constraint to surrounding URLs to get content other than find us pages.
5. URLs' are classified by passing to a bag of words classifier previously trained to recognise pages of interest within an emergency setting.
6. Fast extraction algorithms are employed to recognise specific features, such as contact names and telephone numbers.

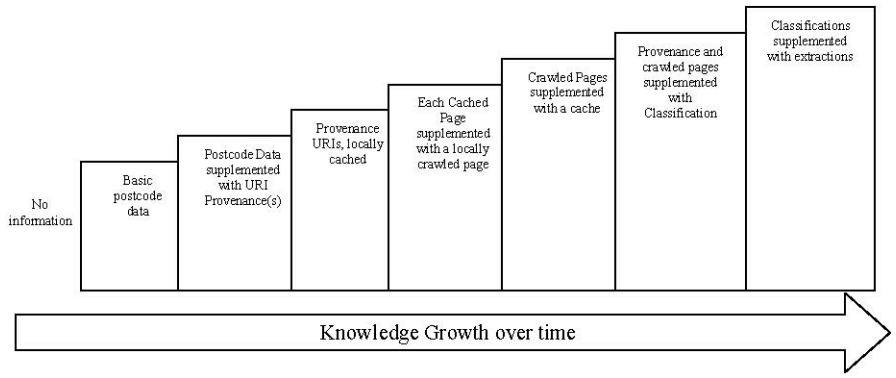

Fig 2 - Knowledge is expanding over time

Degrees of differing knowledge have differing generation costs – i.e. basic postcode data is a fast process that propagates fast as far as data allows, web mining is slower, caching requires bandwidth to the external world, crawling requires minimal processing of each page, classification is fast as far as NLP is concerned but still requires document tokenisation to pass into the classifier a bag of word vector space model to perform its classification.
This means that, to provide knowledge in a timely manner, it is important to focus the more time demanding techniques on the most relevant data, so NLP style extraction is focused upon resources that are classified as important for the domain and are above a certain pre-defined threshold.

4.2 Interface

A map-based interface has been developed for the system, allowing zoomable level of detail and service based queries. The chosen interface follows the *mash-up* paradigm, in which content from different sources is integrated in one interface, to present an immediate overview of a domain.
For Armadillo e-response the choice has been to visualise the information focusing on the geographical dimension. For this reason, a map has been chosen as the main navigation aid for the user, this can zoom and change the level of detail when browsing the knowledge space. Content extracted from different sources is presented in the form of layers overlapping the map. Further information is then visualised in the right hand side of the interface.

The interface has been developed using Google Map API[4].
In Fig 3, an example of search performed with Armadillo e-Response. The user is able to

[4] http://www.google.com/apis/maps/

zoom on the map and change the level of detail, while on the right the information extracted for each area is displayed.

Fig 3 - Armadillo e-response interface

4.3 Real world use case

As an example, consider this scenario of a disaster happening in London. It is now 10:00 a.m.; a cargo aeroplane crashed on the City of London roughly an hour ago. Individual response units, their efforts mostly uncoordinated, have begun to tackle the emergency. The Joint Emergency Services Control Centre (JESCC) is now operational, with communications in place, and is ready to assume strategic control over the incident.
If using a e-response system (shouldn't you find a name to armadillo? Even if is simple armadillo e-response), detailed information could be provide in a timely manner to the JESCC. The user could insert he exact location of the triage centre within the inner cordon of the disaster. The system could immediately pull out contact names and telephone numbers of nearby nightclubs, conference halls, pubs from existing triples.
This would save time for the emergency service in looking for information, allowing them to better coordinate and focus their efforts.
In particular the information should take the example of finding a triage centre within the inner cordon, the system can suggest from existing triples that nightclubs/conference halls etc within the vicinity could be contacted using the obtained contact name and telephone number as soon as the page can load after clicking on the region. This obvious saving of investigation and providing a degree of provenance to double check can alleviate timely information search for a user in a stressful and an unpredictable manner.

4.4 Evaluation

Evaluating an emergency response system in a quantitative manner is a complicated task, as at the moment there are no gold standards to measure findings. A full subjective evaluation will take place in the next months, along with an evaluation of the related tools, as Armadillo e-Response is part of a larger piece of work that aims to apply semantic web technologies to the Emergency response problem.

5 Conclusions and future work

Armadillo e-Response uses Human Language and Semantic Web Technologies to produce an expanding knowledge base from existing unstructured web content; to aid emergency response decisions

In this domain, the main requirement is timeliness of knowledge provided: it is better to have less information but quickly than having to wait hours before some information can be available. In this perspective Armadillo e-Response has been built orchestrating different services that work independently to mine the web and extract useful information in an incremental manner. When the first data becomes available, the system is still working in background to increase the knowledge space that will be provided to the user. Language technologies structure and extract useful information, making possible for a user to quickly get hold of useful contacts in the disaster area, without having to browse web pages that may not contain relevant information or may not be available at the time of disaster, due to technical problems (as often happens in emergency situations).

Armadillo e-Response is part of a much larger work carried out by the AKT consortium, generally involving the usage of semantic web technologies to aid the timely management of resources and information within an emergency scenario. This project will address issues such as simulation and resource monitoring of service units (as an example, fire engines) to facilitate the decision making and information tracking of emergency response. All importantly such technology will provide a trace in time of decisions made and information available to provide better system and planning for the future of emergency response.

Acknowledgements

This research was funded by the Advanced Knowledge Technologies (AKT) Interdisciplinary Research Collaboration (IRC). AKT is sponsored by EPSRC, grant number GR/N15764/01.

References

1 Leidner, J., Sinclair, G., Webber, B.: Grounding spatial named entities for information extraction and question answering, Proceeding of Analysis of Geographical References Workshop, NAACL 2003, Edmonton, Alberta.

2 Vaid, S., Jones, C.B., Joho, H., Sanderson, M.: Spatio-Textual Indexing for Geographical Search on the Web, Proceedings of the 9th International Symposium on Spatial and Temporal Databases, 2005

3 Holland, D. A. and Tompkinson, W.: Improving the update of geospatial information databases from imagery using semi-automated user-guidance techniques. Geocomputation. Southampton, 2003.

4 Siebra, C. (2005) Planning Requirements for Hierarchical Coalitions in Disaster Relief Domains. Selected Papers from AI-2003/4 Poster Session, Expert Update Vol. 8, No. 1, pp. 20-24, Summer 2005, The Specialist Group on Artificial Intelligence, British Computer Society (BCS-SGAI).

5 Berry, D., Usmani, A., Torero, J., Tate, A., McLaughlin, S., Trew, A., Baxter, R., Bull, M. and Atkinson, M. (2005) FireGrid: Integrated emergency response and fire safety engineering for the future built environment, invited talk, Workshop on Ubiquitous Computing and e-Research, National eScience Centre, Edinburgh, UK 18-19 May 2005.

6 Potter, S., Tate, A. and Wickler, G. (2006) Using I-X Process Panels as Intelligent To-Do Lists for Agent Coordination in Emergency Response, Proceedings of the Information Systems for Crisis Response and Management 2006 (ISCRAM2006), Special Session on "Multiagent Systems for Disaster Management and Response" , Newark, New Jersey, USA, May 15-17, 2006.

7 Tate, A., Dalton, J., Bradshaw, J.M. and Uszok, A. (2006) Coalition Search and Rescue - Task Support: Intelligent Task Achieving Agents on the Semantic Web. Interim Technical Report (Final DAML Program Technical Report), January 2003-December 2004, AIAI

8 LESLP Major Incident Procedure Manual – 6th Edition July 2004 - Published by the UK Directorate of Public Affairs, Metropolitan Police Service on behalf of the London Emergency Services Liaison Panel (LESLP). Enquiries about copies of this Manual should be made to the Metropolitan Police Service on 020 7230 3337 or www.leslp.gov.uk

9 Ciravegna, F., Chapman, S., Dingli, A., Wilks, Y.: Learning to Harvest Information for the Semantic Web. Proceedings of the 1st European Semantic Web Symposium (ESWS-2004), Heraklion, Greece, May 10-12, 2004
10 Hepple M, Ireson N, Allegrini P, Marchi S, Montemagni S & Gomez Hidalgo JM: NLP-enhanced Content Filtering within the POESIA Project, LREC 2004.