cannot contribute to our understanding of the process employed by decision-makers to convert reasoning to text.

Nevertheless, the heuristic can mimic the process sufficiently well to be useful in the task of supporting document drafting in a complex domain where there will always be a need for human interaction and refinement. Future research in this direction is aimed at formalising the heuristics and engaging in more rigorous evaluations. An evaluation should include comparisons of documents generated with our heuristics with actual documents in addition to studies that measure the readability or coherence of documents generated with our heuristics.

5. Acknowledgements

This research was supported by the Refugee Review Tribunal, Australia and the Australian Research Council.

6. References

[1] Branting, L., K., Callaway, C, B., Mott, B., W., and Lester, J., C, 1999. Integrating Discourse and Domain Knowledge for Document Drafting. *Proceedings of Seventh International Conference on Artificial Intelligence and Law* ***ICAIL'99.*** ACM Press. pp214-220.

[2] Dick, J. P. 1991. *A conceptual, case-relation representation of text for intelligent retrieval*. Ph.D Thesis. University of Toronto. 1991. Canada.

[3] Dietz, J. L. G., and Widdershoven, G. A. M., 1992. A comparison of the linguistic theories of Searle and Habermas as a basis for communication supporting systems in van de Richt, R. P. and Meersman, R. A. (Eds)., 1992. *Linguistic Instruments in Knowledge Engineering*. Elsevier Science Publications. Pp121-130.

[4] Habermas, J., 1987. The theory of communicative action. (tr Thomas McCarthy). Boston : Beacon Press

[5] Grosz, B. J, and Sidner, C. L. 1986. Attention, Intentions and the structure of discourse. *Journal of Computational Linguistics*. Vol 12, No 3. Ppl75-204.

[6] Mann, W, C. and Thompson, S. A., 1988. Rhetorical Structure Theory: Toward a functional theory of text organisation. *Text* Vol 8, No. 3. pp 243-281.

[7] Marcu D. 1997. The Rhetorical Parsing, Summarisation, and Generation of Natural Language Texts. Ph D Thesis, Department of Computer Science, University of Toronto.

[8] Searle, J., 1969. *Speech Acts: An Essay in the philosophy of language*. Cambridge University Press. Cambridge.

[9] Stranieri, A., and Zeleznikow, J. 1999. A survey of argumentation structures for intelligent decision support. *Information Systems and Decision Support Systems* ***ISDSS'99*** Melbourne July 1999 Monash University Press.

[10] Toulmin, S. 1958. *The Uses of Arguments*. Cambridge University Press. Cambridge

Automatic document metadata extraction and manipulation: a working system for the Intelligence Analyst

Mark Burnett

DSTO
Dept of Defence
Fern Hill Park, Bruce, ACT

Richard Jones

Lloyd-Jones Consulting
PO Box 6155
Philip ACT 2605

mark.burnett@dsto.defence.gov.au jonesrl@compuserve.com

Abstract

This paper discusses the design and implementation of an operational system to aid health intelligence analysts. The HINTS system provides automated support to undertake tasks such as specific health-related research and report writing in the face on an ever-growing body of electronic information, available on the web, and on local file systems. Our approach is to provide automated support for document analysis and discovery from technologies that support ad-hoc searching, consistent filtering for specific pieces of information such as hospital facilities, diseases and locations, and that provide document summarisation and keywording. Document metadata is stored in XML in a data structure that allows a variety of searches and views of the document space to be ***performed.*** *The user interfaces to the system by web browser and a map-based geospatial application.*

Keywords
Document Analysis, Document Databases
Information Retrieval, XML,
Information Extraction

1 Introduction

Intelligence analysts typically operate in two modes. In the first mode, they scan, on a regular basis, a wide range of documents from a range of sources, in case they contain something that might be useful. A selection is made on some general criteria and those documents are put to one side for a rainy day, organised in some way so that they can be found later. They move into the second mode of working when called upon to prepare a specific brief. They obtain the source material for the brief by accessing this repository of information, and combining this with the results of specific searches from other sources. Invariably the information that is key is a side issue in a document that may be discussing some other topic. This information extraction requirement is, of course, the focus of the MUC series of experiments [1].

This paper describes a repository system to support both modes of operation described above, but with more emphasis on the first. The system, known as HINTS, is based upon an XML store containing three types of information, standard bibliographic information, domain specific information - in this case the health domain, and information specific to the intelligence analysis requirement. A key part of the system is to assign specific values to metadata elements automatically. To this end an information extraction process looks for specific entities (e.g. disease names, locations), and a document **summariser** assigns **keywords** and extracts a summary.

A strong goal of the design was to make the system generic and easily applicable to other intelligence domains. To this end we selected technologies that operate on text at a surface level, and a component-based architecture for integration.

2 XML Data Storage

We chose an RDF implementation of XML to provide a generic storage mechanism, and to allow concepts, and relations among concepts, to be defined. RDF also supports a class system analogous to Obect-Oriented programming and modeling systems. The RDF specification has now progressed to a "Proposed Recommendation" to the W3C (see http://www.w3c.org/TR/PR-rdf-syntax).

3 Metadata Extraction

It is often useful for users to read some sort of condensed or structured surrogate for a document. The purpose may be varied: it may be to determine if the document is worth reading in full, or to extract specialised information from the document, either by reading specific portions, or even without reading it at all. Some standard structures to support these requirements, (e.g. MARC) and abstracts, keywords, etc. have been in place for a long time. This

information is designed to tell anyone, regardless of their role or background, what a document is about.

Bibliographic metadata does not assist if the user requirement is to find a particular fact, such as an instance of an environmental event, like a volcanic eruption, or references to particular entities such as a company or person. This information may be imbedded in a document whose main topic is something completely different, and it is impractical to extract such objects without some detailed knowledge of the interest of the reader. Such information can certainly be seen as metadata, being domain specific or even user specific.

The existence of metadata also provides a set of indexing mechanisms to enable particular documents to be retrieved, with improved precision over full text retrieval. e.g. Find Jones as an author.

A major problem for determining content-based metadata has been the effort and skill required to create or extract it. Over the past 10 years, and especially in the past five years, techniques have been developed to try to automate some of these processes with greater or lesser success. The US government has given a major impetus to such techniques through the TREC and MUC conferences.

Two software packages were used to provide the automatic assist in metadata generation in HINTS:

- FXBench, and
- InTEXT Analyser

3.1 FXBench

FXBench was developed in DSTO in the Electronic and Surveillance Research Laboratory in Salisbury, South Australia. It comprises a suite of tools that can assist a Language Engineer to identify patterns in text, and to write Fact Extractors that produce formatted data from documents.

Fact Extractors are of two forms:

- Closed vocabulary lists of nomenclature (e.g. disease lists) where words or phrases can be mapped onto a preferred form; and
- Pattern definitions defining regular expressions using the PERL syntax.

HINTS uses FXBench to fill in domain specific metadata fields.

3.2 InTEXT Analyser

InTEXT Analyser was developed by InTEXT Systems Inc. It is based upon AIDA, the summarising and keywording software developed by Computer Power Group in conjunction with the Australian Federal Parliamentary Information Systems Office in the early 1990s (Jones [2] and Thistlewaite [3]). Another variant of this software was used in TREC 4 (Burnett [4]) in an experiment to determine the effectiveness of establishing a reduced full text index using only the automatically extracted keyphrases.

HINTS uses InTEXT Analyser to extract keyphrases, and sentences to fill in the Keyword and Summary metadata fields.

4 Technology Integration

The architecture used to integrate the various technological components employed an n-tier distributed set of components. It is a complex architecture incorporating:

- a web server (MS Internet Information Server);
- an XML data server (eXcelon)
- DCOM and CORBA components; and
- a Client/Server (JavaBeans) Component.

The HINTS operational prototype runs on an NT server, with browser and Java-application clients. The user interface is described in the next section.

5 Searching And Viewing The Metadata

Dynamic views and searches of the data in the metadata repository are served out to web clients and to a Geospatial Java application.

5.1 Browser access

A browser provides access to the following functional features:

i. Metadata Basic Search;
ii. Metadata view by Entity;
iii. Metadata Advanced Search;
iv. Modify Metadata;
v. Basic Search ;
vi. Advanced Search;
vii. Content Classification;

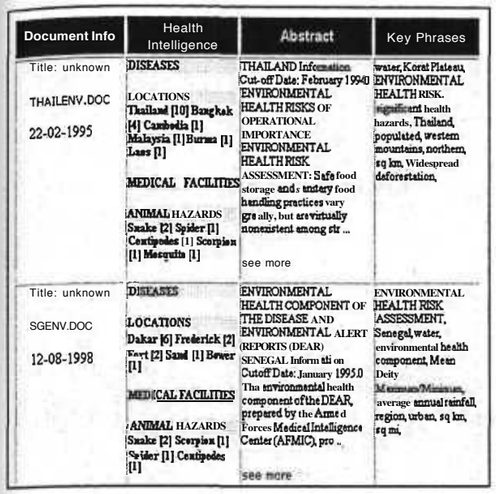
Features v-vii are provided by UltraSeek, a commercial text retrieval product available from InfoSeek. It was configured to index both the internal HINTS collection of documents, and a set of external health-related web sites.

Searching, viewing and modifying the metadata repository used active server pages, linked with server-side extensions of the eXcelon server, to access the XML store.

The screen shot shown in Figure 1 shows part of a document hitlist following a metadata query.

Viewing by location, medical facility, hazardous animal, or disease is available for the current set of documents from a drop-down menu.

These dynamic views allow greater access to the information stored in the repository, and show how different entities and concepts are related across a set of documents.



Document Info	Health Intelligence	Abstract	Key Phrases
Title: unknown THAILENV.DOC 22-02-1995	DISEASES LOCATIONS Thailand [10] Bangkak [4] Cambodia [1] Malaysia [1] Burma [1] Laos [1] MEDICAL FACILITIES ANIMAL HAZARDS Snake [2] Spider [1] Centipedes [1] Scorpion [1] Mosquito [1] see more	THAILAND Information Cut-off Date: February 1994 ENVIRONMENTAL HEALTH RISKS OF OPERATIONAL IMPORTANCE ENVIRONMENTAL HEALTH RISK ASSESSMENT: Safe food storage and sanitary food handling practices vary greatly, but are virtually nonexistent among str ... see more	water, Korat Plateau, ENVIRONMENTAL HEALTH RISK. significant health hazards, Thailand, populated, western mountains, northern, sq km, Widespread deforestation,
Title: unknown SGENV.DOC 12-08-1998	DISEASES LOCATIONS Dakar [6] Frederick [2] Egypt [2] Saud [1] Bower [1] MEDICAL FACILITIES ANIMAL HAZARDS Snake [2] Scorpion [1] Spider [1] Centipedes [1] see more	ENVIRONMENTAL HEALTH COMPONENT OF THE DISEASE AND ENVIRONMENTAL ALERT (REPORTS (DEAR) SENEGAL Information Cutoff Date: January 1995.0 The environmental health component of the DEAR, prepared by the Armed Forces Medical Intelligence Center (AFMIC), pro ... see more	ENVIRONMENTAL HEALTH RISK ASSESSMENT, Senegal, water, environmental health component, Mean Deity Maximum/Minimum, average annual rainfall, region, urban, sq km, sq mi,

Figure 1: Document metadata.

The screen shot in Figure 2 is a location view of the document set following an "ebola" query.

The locations contained in the document set are presented as an ordered list, with the most commonly occurring location at the top. For each location, the set of documents that contains that location is shown as a sub-list. These documents are ordered by their relevance to that term, with a colour-code indicating the value of the relevance.

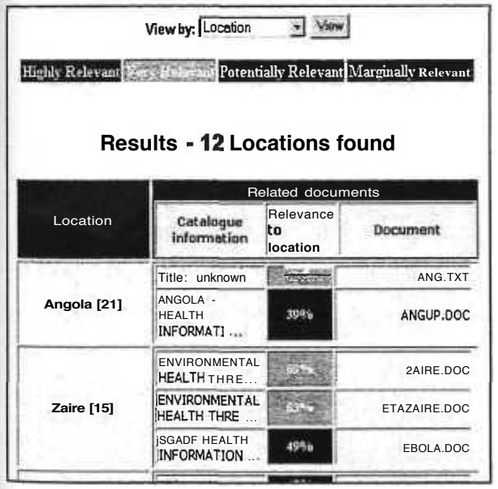


Figure 2: A location view.

5.2 Geospatial access

A geospatial interface is a natural interface to documents that often deal with one or more locations. Using a highly detailed gazetteer, location names are mapped to co-ordinates that allow data to be displayed on a map of the world. Using a health layer in the OpenMap(tm) package (http://javamap.bbn.com:4711) a user can display document information on a map at various magnifications, and drill down to very detailed views of a region.

Figure 3 is a screenshot showing the results of an "ebola" query, with folders indicating documents containing occurrences of ebola.

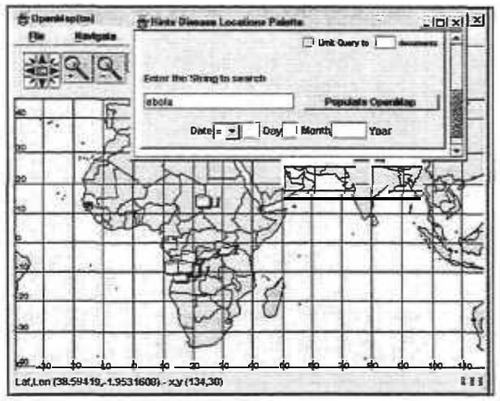


Figure 3: Querying the metadata via Openmap.

6 Conclusion

This paper has sketched the design and implementation of a generic system for intelligence analysis. Particular attention has been paid to the extraction and use of document metadata.

References

[1] *Proceedings of the Seventh Message Understanding Conference (MUC-7),* Morgan Kaufmann, San Francisco, 1998.

[2] R.L. Jones. AIDA - the Artificially Intelligent Document Analyser, *In Libraries and Expert Systems, ed. McDonald and Weckert,* Taylor Graham, pp 49-56, London, 1991.

[3] P.B .Thistlewaite and S.Blume. Offloading Information Overload: the AIDA Project. In *Proceedings of the Seventh Conference for Librarians in the Criminal Justice System,* Canberra, January 1990.

[4] S.M Burnett, C. Fisher and R.L. Jones - InTEXT Precision indexing in *TREC4. ed. : D. K. Harmon NIST Special Publication 500-236,* pp 287-294, 1995. http://trec.nist.gov/pubs/trec4/t4_proceedings.html