Ranked Retrieval and Extraction of Open Source Intelligence

Paul Thompson
PRC Inc

1.0 Introduction

The present uncertainties confronting the planner of future intelligence analysis systems are largely the result of the rapid disintegration of the Soviet Union as a world power and the resulting collapse of the Warsaw Pact. The impact of these political changes on the intelligence community has been enormous. With the Soviet threat significantly diminished, the rest-of-the-world has suddenly escalated in importance—a process from which a new set of national concerns and priorities has emerged. One force shaping this changed intelligence mission is the nearly explosive growth of open-source intelligence (OSINT). For example, with the breakdown of tight controls on the release of information from the former Soviet Union whole new libraries of technical literature are now flooding into the public domain. Within the United States, subscription database services are proliferating rapidly. This demand for information will continue to grow within our society as its economic base continues to shift from industry to services. Furthermore, as global economic competition intensifies, economic data and assessment of the relative position of various countries' economies will become a focal point for intelligence analysis. It will be necessary to find ways to effectively handle this new wealth of information.

Even before the advent of the current uncertainties concerning the rapidly evolving threat environment, the unified and specified (U&S) commands and the national foreign intelligence community faced the problem of the increasing volume and volatility of available information. As collection system capabilities continue to improve and the degree of interoperability between command elements broadens, the capacity of these organizations to identify, analyze, and apply all of the important data will be strained to the point of failure. Ultimately, critical pieces of information will be overlooked and denied to decision makers. This problem is made more acute by those programs attempting to integrate multiple complex functions requiring the rapid fusion of many different sources and formats, i.e., multimedia, of information, as well as by the reduction in personnel resources available to work the problem. Even those intelligence activities that rely largely upon formatted raw data (e.g., ELINT data processing) have serious problems in effectively handling the increasing volume of input. For those activities relying heavily on textual input, the problem is exacerbated even further.

The data glut currently inundating the intelligence analyst is due in large measure to inability to deal effectively with textual information. The user is compelled to sift through mountains of text searching for those scattered pieces of useful information that will allow him to do his job. Since the system that provides the analyst with information cannot *understand* the content of the text with which it deals, it cannot properly match his interests against its holdings. As a result, the analyst must do his own filtering and selecting—a time consuming task that leaves less time for his real job of extracting and summarizing the scientific and technical significance of the available information.

Even if the analyst finds just the right source documents to deal with the problem at hand, a further problem is presented. Should information contained in these documents need analysis, the analyst must manually extract the data of interest, reformat it as needed to perform the needed analyses (statistical tests, cybernetic modeling, etc.), and then finally

organize it into summary tables, graphs, maps, and charts for display purposes. In a system designed to support the whole cycle from text to graphical display, the above process would be broken down into two major phases: (1) text-to-database and (2) database-to-display. During the first phase, the system would read the source text, extract the relevant concepts, and store the resulting templates of information into a formatted database. Ideally, this process of updating the database would take place without user intervention. During the second phase of the cycle, the user would be presented, not with raw text, but with formatted data summarized in the form of charts, graphs, tables, and maps. The user would not need to review the textual source materials unless interested in the context of the data presented or unless it was felt that all information of interest had not been extracted. This might happen because information which should have been extracted was missed, or because the analyst required information that the extraction process was not designed to extract. For these reasons it is necessary to complement extraction with an information retrieval system. Conventional retrieval systems often produce output overload, i.e., too many documents are returned in response to a query. Ranking retrieved documents according to probability of relevance provides a solution to this problem. At PRC we have developed both types of system through independent research and development projects, and are moving towards an integrated system combining the capabilities of both. Our information extraction system, is PAKTUS (PRC Adaptive Knowledge-based Text Understanding System). Our ranked retrieval system, ADIIR, is being developed under our Automated Document / Image Indexing and Retrieval (ADIIR) project

2.0 Information Retrieval of Textual Open Source Intelligence

The first long-term objective of the ADIIR project is to develop effective, computationally implemented retrieval algorithms that can efficiently scale up for very large text and image retrieval systems. The second is to develop an automatic indexing engine that will capture key elements of a scanned document based on document class and common zone patterns, eliminating costly manual indexing, as discussed in section 4.0. The third is to develop indexing and retrieval algorithms that can take advantage of structural information contained in documents, e.g., those formatted in the Standard Generalized Markup Language (SGML).

In support of these objectives we are developing a retrieval system/testbed incorporating:

- Multiple retrieval models and combination methodology
- Intelligent interface including user modeling and relevance feedback
- Techniques for rapid query processing
- Acquisition of large document/image test collections
- Automatic image indexing.

2.1 The ADIIR Approach and Progress to Date

The ideal text retrieval system retrieves documents based on an understanding of the meaning of the query and the document. Due to ambiguity in the use of words and concepts, keyword retrieval falls far short of such an ideal (Furnas et al. 1987, Blair and Maron 1985). Furthermore, scanned documents, or those in SGML format, provide structural and layout information of which a retrieval system should take advantage. All

such information, however, provides at best clues to document relevance. Thus, a probabilistic approach is warranted.

Figure 1 shows the prototype ADIIR retrieval system. The two arcs from the incoming document stream indicate that some documents are in ASCII format, others are paper, requiring scanning and optical character recognition (OCR). The analysis system provides automatic indexing. Natural language processing (NLP) is provided by an NLP module created using PRC's NLP shell, PAKTUS (Loatman 1987, Thompson 1992). The intent is to provide concept-based indexing with broad, shallow, domain-independent NLP without extensive handcrafting. Analyzed documents are stored in an object-oriented repository customized for document retrieval, LEND (Chen 1992). LEND supports various types of data including a comprehensive computerized lexicon for English derived from machine readable dictionaries, factual domain knowledge, large digital, hyper-linked, and the multimedia archives from which retrieval takes place. Finally, the retrieval system matches document representations derived by the analysis system to query representations derived from the user.

[Insert Fig. 1]

The first stage of this research has involved: a) analytical work developing and implementing the individual retrieval algorithms and the methodology for combining them in an overall ranking algorithm; b) acquisition of large test collections of paper, image, SGML-tagged, and ASCII documents; c) evaluation and acquisition of commercial products for scanning/OCR and SGML parsing, and of university research prototypes for document retrieval (Turtle and Croft 1991, Chen 1992). The second stage, presently underway, is to build the prototype/testbed system. These stages include the following tasks.

Task 1 – Multiple retrieval models and combination methodology. Research has shown that different retrieval models retrieve different sets, only slightly overlapping, of more or less equally relevant documents (Katzer et al. 1982, Fox et al. 1988). Accordingly, a major thrust of this project is to implement the Combination of Expert Opinion methodology (Thompson 1990) for combining the results of multiple probabilistic retrieval models into an optimal ranking of documents.. Each retrieval model is viewed as a retrieval expert.

Task 2 – User modeling and relevance feedback. Current retrieval systems provide little support in query formulation. User models allow a system to adapt to each user, enabling mixed-initiative interaction. In this project relevance feedback algorithms will be investigated.

Task 3 – Techniques for rapid query processing. Fox, in collaboration with this research, is investigating techniques for rapid processing of queries against very large text databases using minimal perfect hash functions (Fox et al. 1992). These are especially important for accessing data on optical storage, since they guarantee collision-free hashing, thus reducing the number of seeks required on inherently slower optical storage devices.

Task 4 – Acquisition of large document/image test collections. Standard test collections are small compared to today's terabyte databases. It is uncertain how retrieval techniques will scale up. Furthermore, these collections are not well suited to testing iterative, feedback-based retrieval, nor to image retrieval. We are obtaining access to large document text and image collections for our testbed. Through participation in the Text Retrieval Conference (TREC), we have acquired an approximately half million

document, 2 gigabyte corpus of ASCII text (Proceedings of the TREC Conference, 1993). PRC participated in TREC as a team with Professor Edward Fox and his colleagues at Virginia Polytechnic Institute and State University. A very simplified version of our Combination of Expert Opinion algorithm produced results comparable to most of the other systems. The algorithm we now have implemented is capable of much better performance.

3.0 Information Extraction of Textual Open Source Intelligence

PAKTUS is a NLU shell providing full syntactic and semantic text processing. PAKTUS has been used to develop information extraction, or automatic database generation, applications. In these applications messages are processed to extract information to fill relational database, or knowledge base, records. Such records can then be used as part of a conventional relational database or as formatted fields of a full-text database record, or as input to an expert system knowledge base. In addition PAKTUS has been used for conceptual information retrieval. This section will describe the PAKTUS shell and its application to text extraction, particularly in the context of the Message Understanding Conferences (MUC).

PAKTUS is written in Common Lisp and uses features of PIKS, a frame-based AI development language developed at PRC. PAKTUS provides a coherent knowledge representation scheme that can be used to process texts in a given domain. Its knowledge representation structures include a core lexicon and conceptual representations, or case frames, which are domain-independent, along with an additional lexicon and conceptual patterns which are domain-dependent.

The architecture of a PAKTUS-based NLU extraction system is summarized in Figure 2. Knowledge representation structures are represented by ovals. Processing begins with the arrival of an electronic stream of text and ends with the output of extracted information as database or knowledge base updates. Alternatively messages are disseminated and/or archived for later retrospective querying. This overall architecture can be grouped into three broad areas: a) linguistic processing, which proceeds through the conceptual analysis component; b) discourse analysis; and c) information extraction. All of these components are described below.

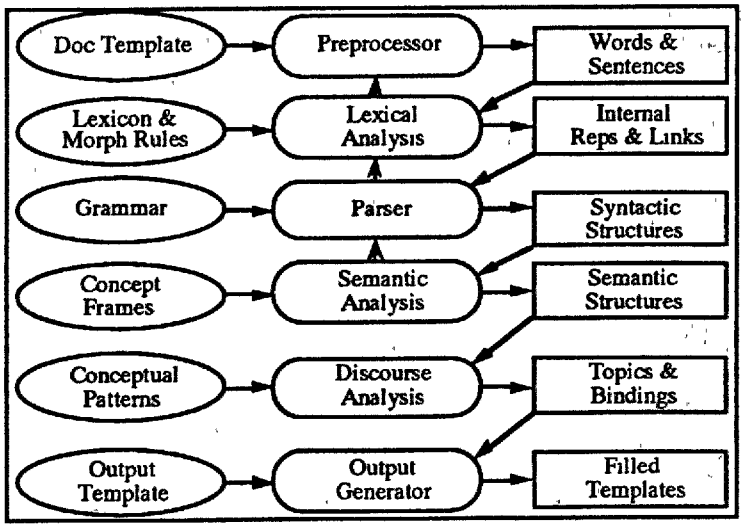

Figure 2. The PAKTUS Architecture

3.1. Problems and Issues

The development of a prototype NLU system intended to process messages in real time for automatic database generation, selective dissemination of information, and retrospective querying raises a number of issues not necessarily addressed by purely theoretically motivated systems. These include: linguistic, pattern matching, and filtering issues.

3.1.1 Linguistic Issues

Natural language understanding of unrestricted text is not feasible given the current state of the art. However if the linguistic domain can be suitably narrowed, NLU is feasible. Domain knowledge associated with a sublanguage can be encoded into knowledge bases or patterns for extraction such as with PAKTUS. In a domain such as scientific articles on superconductivity the notion of a sublanguage has been expanded to include large areas of physics and chemistry. Nevertheless, PAKTUS's information extraction capabilities have been successfully applied to this domain among others.

3.1.2 Pattern Matching Issues

The highly structured nature of the case frame representations of the parser's output suggested that it might be possible to exploit the structure for extraction directly through pattern matching. With this approach a developer takes the case frame representation of a representative clause from a message and edits the representation. The edited case frame is matched directly with the representations of clauses from messages being processed. If a match takes place, the case frame fillers are mapped into the appropriate template slots.

This approach operates primarily at the phrase or clause level. Thus it provides very little discourse analysis, i.e., intersentential analysis. Consequently, this approach has been modified so that pattern matching takes place against previously constructed topic structures, rather than against the original case frames from each individual sentence (Loatman 1992). Topic structures are sets of case frames having common topic objects and times. Topic objects are defined as fillers of certain case roles, specifically, 16 of the total 40 case roles used in PAKTUS. The most notable case role that is *excluded* as a topic object is the Agent. This is because topic structures are meant to represent information about entities that are being affected or focused upon in some way, whereas a single Agent can operate on several different entities. After all case frames have been assigned to topic structures, domain-specific conceptual patterns are compared to the case frames, topic-by-topic rather than sentence by sentence as had been done previously, binding pattern variables to information that is extracted and put into event reports whose format is specified by a domain-specific template.

Figure 3 shows the current approach to discourse analysis and extraction. This approach was first used for MUC-4, but is generic for expository text, such as news reports. In Figure 3, only the conceptual patterns and filter are MUC-4-specific, and these are part of the extraction component, not discourse analysis.

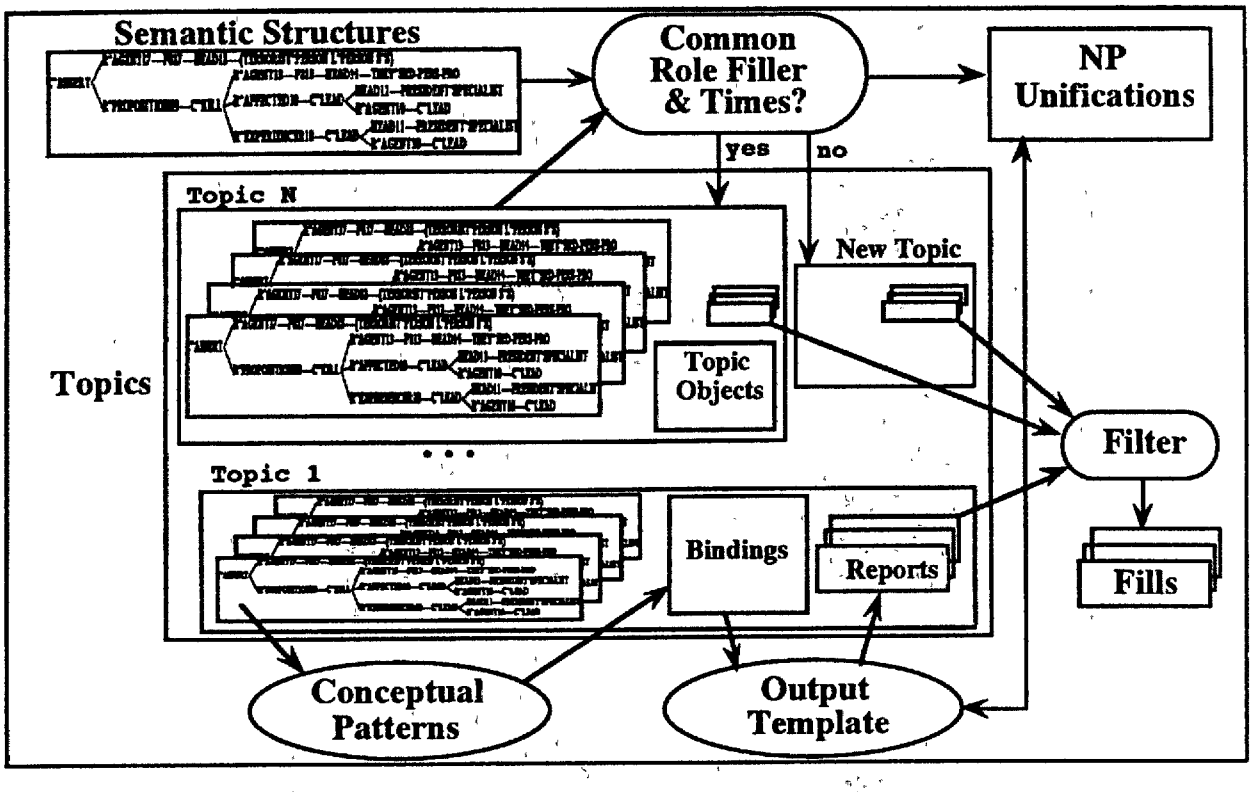

Figure 3. Discourse Analysis and Extraction Details

3.1.3. Filtering Issues

Integration of a conceptual querying module with the MUC version of PAKTUS would provide a system which could perform case frame based retrieval using a full parse of both query and message. Such an approach is too computationally intensive. What is required is a filtering mechanism that will allow full parsing to be done only where necessary and a method of concept-based indexing so that a case frame representation of a query could be compared to a concept-based surrogate of a record, rather than to the conceptual representation of every sentence of the message. The filtering process is currently being implemented. Jacobs et al. (1991) describe a similar filtering approach. Using this filtering, they were able to eliminate the need for parsing 75% of the MUC-3 text during the filtering, or preprocessing stage.

3.2 PAKTUS and the Message Understanding Conferences (MUC)

The intent of the Message Understanding Conferences has been to compare message understanding system performance on a common test set of messages (Sundheim 1989). This has been a black-box style of evaluation where only the output produced for the task is examined, rather than the system internal processing. The fourth message understanding conference (MUC-4) was held in June 1992, in McLean, Virginia (*Proceedings of the Fourth Message Understanding Conference (MUC-4*, 1992). The MUC-3 corpus was used and, with slight modification, the extraction task remained the same as for MUC-3. For MUC-4 the major enhancement to PAKTUS was the new discourse analysis module using topic structures which greatly improved intersentential discourse analysis, which in turn significantly improved the integration of multiple templates generated from a single text. PAKTUS development time for MUC-4 was

limited to only 4 person months of linguistic effort. Despite this, when MUC-3 results were rescored to MUC-4 guidelines, PAKTUS showed the most improvement of any system participating in both conferences, indicating how much improvement was made using the new discourse analysis technique. These scores placed PAKTUS's performance in the top half of participating systems.

4.0 Document Image Retrieval

As mentioned in section 2.0, one of the primary objectives of the ADIIR project is to apply advanced automatic indexing and retrieval techniques to document images. Thus, to complete the lists of tasks given in section 2.1, our final ADIIR task is:

Task 5 – Automatic image indexing. Converting a paper document to an indexed image currently requires manual intervention at three points: 1) indices creation; 2) indices to image link creation: 3) OCR correction. Our goal is to reduce manual involvement as much as possible. This task includes three subtasks.

- Evaluate COTS packages for zone OCR capability and SGML-tagging and parsing. We seek to delimit, or SGML-tag, zones in a raster image so that the full OCR process can be applied to the relevant zones. This provides a means of: a) filtering sections of documents so that OCR can be applied selectively; and b) utilizing structural information (e.g. SGML) to weight index terms extracted from certain critical zones (e.g., an abstract is weighted more highly).

- Establish expert thesaurus database and implement reference tables to improve OCR accuracy. The key terms of a particular domain form the basis for the thesaurus. Reference tables are also based on expert knowledge of the domain of the document database. They are used to verify and/or correct the OCR interpretation of special names and titles. Typical reference tables may be parts lists, glossaries, phone directories, or organizational charts.

- Automate indexing of scanned and OCR output. The same indexing engine will be used as for ASCII text with the inclusion of modules to handle the additional representational elements involved, e.g., a mechanism to identity zone boundaries.

5.0 Conclusion

In depth, domain-independent information extraction and display is beyond the present state-of-the-art. In the research reported here the focus has been on retrieval and extraction of information rather than display and analysis. An integrated open source system requires display and analysis capabilities, as well. In the area of information retrieval three issues stand out. First, the need to combine the results of different retrieval models in order to get maximum retrieval effectiveness. The Combination of Expert Opinion algorithm addresses this need. Second, ultimately natural language understanding of query and document is called for in order to obtain the highest retrieval performance. Integration of ADIIR with suitable modules of PAKTUS is a step in that direction. Third, well over 90% of open source literature is in paper format. Automatic retrieval or extraction of information contained in these documents requires prior scanning and OCR. Thus one of the main foci of ADIIR is automatic analysis of document images.

An integrated open source solution for the general problem of information retrieval, extraction, display, and analysis is not one of complete automation. Rather it is a system that is highly interactive with the analyst. In order to make the analyst's task feasible, however, the retrieval, extraction, display, and analysis components of the system must be made to perform in as automated a mode as possible. The display and analysis components must be flexible enough that each analyst can work with the data in the manner he or she finds most useful. Extraction techniques can be very useful in domains where certain known types of information are to routinely extracted. Information retrieval techniques are still necessary, however, for several reasons. Extraction systems may miss information that they have been designed to extract. There may be no extraction systems developed for most domains of interest. Finally, there will always be a requirement to retrieve information the need for which was not anticipated when the extraction system was developed.

References

Blair, D. C. and Maron, M. E. 1985. "An evaluation of retrieval effectiveness for a full-text document-retrieval system" *Communications of the ACM* vol. 28 no. 3 p. 289-299.

Chen, Qi Fan. 1992. "An object-oriented database system for efficient information retrieval applications." Ph.D. Thesis, Virginia Tech Dept. of Computer Science, March.

Fox, E. A.; Chen, Q. F.; Heath, L. S. 1992. *Proceedings of the Fifteenth. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, June 21-24 Copenhagen, Denmark p. 266-273.

Fox, E. A.; Nunn, G. L.; Lee, W. C. 1988. "Coefficients for Combining Concept Classes in a Collection" *Proceedings of the 11th. International Conference on Research and Development in Information Retrieval* June 13-15, Grenoble, France p. 291-307.

Furnas, G. W.; Landauer, T. K.; Gomez, L. M.; and Dumais, S. T. 1987. "The vocabulary problem in human-system communication" *Communications of the ACM* vol. 30 no. 11 p. 964-971.

Jacobs, P.S.; Krupka, G.R.; and Rau, L.F. 1991. "Lexico-semantic pattern matching as a companion to parsing in text understanding" (to appear)

Katzer, J.; McGill, M. J.; Tessier, J. A.; Frakes, W.; DasGupta, P. 1982. "A study of the overlap among document representations" *Information Technology: Research and Development* vol. 2 P. 261-274.

Loatman, B. 1987. "A hybrid architecture for natural language understanding" *Proceedings of Applications of Artificial Intelligence V* SPIE - The International Society for Optical Engineering 18-20 May 1987 p. 416-422.

Proceedings of the Fourth Message Understanding Conference (MUC-4). 1992. San Mateo, CA: Morgan Kaufmann (to appear).

Proceedings of the TREC Conference. 1993 National Institute of Standards and Technology (to appear).

Sundheim, B. M. 1989. "Plans for a task-oriented evaluation of natural language understanding systems" *Proceedings of the DARPA Speech and Natural Language Workshop* . P. 197-202.

Thompson, P. 1992. "Concept-based Information Extraction" *Heuristics: The Journal of Knowledge Engineering*, Special Issue on Knowledge Extraction from Text (to appear).

Thompson, P. 1990. "A Combination of Expert Opinion Approach to Probabilistic Information Retrieval, Part 1: The Conceptual Model.; Part 2: Mathematical Treatment of CEO Model 3." *Information Processing and Management* Vol. 26 No. 3 p. 371-394.

Turtle, Howard and Croft, W. B. 1991. "Evaluation of an inference network-based retrieval model" *ACM Transactions on Information Systems* vol. 9 no. 3 p. 187-222.

FIRST INTERNATIONAL SYMPOSIUM: NATIONAL SECURITY & NATIONAL COMPETITIVENESS: OPEN SOURCE SOLUTIONS Proceedings, Volume II - Link Page

Previous	Government Information Wants to Be Free
Next	Painting the Future: Some Remarks Following the INTERVAL RESEARCH Brainstorming Session of 7 May 1992

Return to Electronic Index Page