AFRL-IF-RS-TR-2006-78
Final Technical Report
March 2006

IFED ENGINEERING SUPPORT

BAE Sysems, Inc.

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

STINFO FINAL REPORT

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-2006-78 has been reviewed and is approved for publication.

APPROVED: /s/

JOHN SPINA
Project Engineer

FOR THE DIRECTOR: /s/

JOSEPH CAMERA
Chief, Information & Intelligence Exploitation Division
Information Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	MARCH 2006	Final DEC 03 – DEC 05

4. TITLE AND SUBTITLE
IFED ENGINEERING SUPPORT

5. FUNDING NUMBERS
C - F30602-01-D-0083/0041
PE - 62702F
PR - 558T
TA - Q6
WU - 41

6. AUTHOR(S)
Michael S. Bilinski

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
BAE Systems
111 East Chestnut Street
Rome NY 13440

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory/IFED
525 Brooks Road
Rome New York 13441-4505

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TR-2006-78

11. SUPPLEMENTARY NOTES
AFRL Project Engineer: John Spina/IFED/(315)330-4032/ John.Spina@rl.af.mil

12a. DISTRIBUTION / AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***
This paper explores the use and development of knowledge discovery and information extraction tools. It can be concluded that these tools can provide a valuable analysis capability and insight for finding patterns of information that do not necessarily fit into the norm of patterns within that data. This allows an analyst more time to perform intelligence analysis on highly relevant information while spending less time on information retrieval. Information extraction tools make it easier for intelligence analysts to find, extract, visualize, and otherwise exploit data of interest from unformatted sources of text of multiple types. Though there are limitations associated with this technology, its benefit to the Air Force and analysts at all levels will be vital to our intelligence gathering capabilities far into the future.

14. SUBJECT TERMS
Information understanding, knowledge discovery, information extraction, extraction technology, information analysis

15. NUMBER OF PAGES
11

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

ABSTRACT

This paper discusses if we can provide superior tools for systems analysis, identification, evaluation, integration, demonstration, and transition of Information Understanding technology products to ensure focused development on operational requirements.

TABLE OF CONTENTS

1.0 INTRODUCTION

The mission of the Information Understanding (IU) lab is to significantly improve the design, development, evaluation, and deployment of Government and commercial software applications for the information analyst. The IU Branch, now the Situation Awareness (SA) Branch, looks to evaluate and integrate applications and technologies that have emerged from numerous Government programs, as well as concepts developed in-house. The IU/SA lab looks to develop a clearly defined set of evaluation metrics that would assist in truly achieving software and information superiority and dominance. The IU/SA lab acquires tools from the Top Sail and Eagle programs as well as from Small Business Innovative Research (SBIR), other Government, and in-house efforts. It performs evaluations on these tools to help formulate a baseline for required equipment and manpower. Tools are classified, evaluated for performance, integration, and transition. Another function of the IU/SA lab is the development of real-world scenarios to test and evaluate the software. Finally, the lab looks to create large data sets to test current and future tools and evaluate their scalability.

1.1 INFORMATION UNDERSTANDING BACKGROUND

The Information Understanding Branch (IFTB) IU lab consisted of three Government scientists, a number of summer students, and one contractor. In late 2003, this group began to evaluate technological innovations and to identify potential enhancements and improvements to knowledge discovery, information extraction, question-answering, and cognition reasoning tools. This group worked to identify new sources of information to test tools, find capabilities and functions within tools that could be used with other tools, and define information flow among tools. This group worked this mission through 2004 when the IFTB became the Situation Awareness (SA) branch (IFED). The mission of the SA branch is to perform leading-edge research and development of technologies to enable the realization of computationally intelligent systems for predictive situational awareness, complex reasoning, and situation understanding. IFED provides an integrated suite of tools and techniques empowering the war fighter, from the Commander down to each combatant, with a comprehensive knowledge of the battle space in support of their decision-making process.

1.2 KNOWDLEGE DISCOVERY

Intelligence analysts are often asked to analyze large amounts of data to find that small vital piece of the puzzle or a hotspot of information in a huge dataset. Hotspots in information are loosely defined as a pattern of information that does not necessarily fit into the norm of patterns within that data. When implementing knowledge discovery techniques on large amounts of data, an analyst is more interested in finding a flow of information through patterns of information

models rather than trends within the data. Information patterns or models represent a warning signal to the intelligence analyst. By identifying relevant, understandable patterns or models within the information, it is possible to transform that information into knowledge.

One method of finding these patterns of information within data is Knowledge Graph Matching. The process of graph pattern matching involves comparing graphs for similarities. Analyst use knowledge discovery tools to arrive at accurate intelligence decisions by allowing the analyst to review and utilize large amounts of information in a timely and orderly manner. These types of tools help the analyst disregard information that may be unrelated or outdated. Knowledge discovery tools allow the intelligence analyst a greater amount of time to perform intelligence analysis on highly relevant information and less time spent on information retrieval.

1.3 INFORMATION EXTRACTION TECHNOLOGY

Information extraction is data extraction and data organization, combined with visualization capabilities that make it easier for intelligence analysts to find, extract, visualize, and otherwise exploit data of interest from unformatted sources of text of multiple types (e.g., message traffic, news wire feeds, document databases, and open sources) and across multiple domains. This structured information can then be stored in a database and used to feed analysis and visualization tools, such as data browsers and link analysis displays, in order to enable analysts to support the war fighters and decision makers with mission planning, campaign assessments, and other analysis products and services.

1.4 TOOL EVALUATIONS

Below is a review of some of the tools evaluated by the IU/SA lab.

TMODS (Terrorist Modus Operandi Discovery System) is developed by 21st Technology. It is an exact, inexact, and partial graph matching tool. TMODS is an attributed, directed graph, where the nodes' entities such as groups, individuals or resources, and the edges represent relationships among the nodes. Both nodes and edges can have attributes that describe their properties. The primary purpose of TMODS is to find sub-graphs with certain topological or attribute characteristics, within the input graph. These sub-graphs are referred to as patterns. The output of TMODS is a list of exact or partial matches of a target pattern, along with a measure of the confidence of the match.

LAW (Link Analysis Workbench), developed by SRI International, is a web-accessible tool where analysts and machines collaboratively perform link analysis. They do this by defining hierarchical and temporal patterns that include uncertain and qualitative elements. They also do this by defining domain dependent and independent search strategies for pattern application,

through a graphical user interface (GUI) that supports direct graphical browsing and editing of patterns, search strategies, and summaries and details of resulting matches.

Proximity, developed by the Knowledge Discovery Laboratory (c/o Professor David Jensen, Director, Department of Computer Science, University of Massachusetts, Amherst, Massachusetts), allows users to easily understand and modify large relational data sets. In addition, users can build statistical relational models to enhance their understanding of the data and to make predictions about new data.

POLESTAR, developed by BAE Systems, is a fact collecting and argument construction tool that enables an analyst to collect text from websites displayed in Microsoft Internet Explorer and Microsoft Word documents in a collaborative environment. An analysis of the information collected, as well as additional metadata to describe source, reliability, and classification is provided to the Fact Collector component for the generation of intelligence products.

ISX Tool Kit, developed by ISX, is a combination of three tools – (1) the Thematic Argument Group (TAG) Manager, (2) the Semantic Navigator, and the (3) Thematic Argument Group (TAG) Log. The TAG Manager is a Groove-based application designed to provide role-based information access and collaboration process control. It provides a natural interface for information sharing and collaboration process within the Groove collaboration environment. Individuals may choose the image to use to represent them within the workgroup. The Semantic Navigator is a Groove-based application designed to make information and not "files" the currency of analysis, collaboration, and exchange. The TAG Log is a Groove-based application designed to capture workgroup activity and provide a temporal view of that activity.

HITIQA is a question-answering system, driven by natural language human-computer dialogue. HITIQA is sponsored by ARDA (Advanced Research Development Activity) and the State University of New York at Albany.

WebTAS, developed by Northrop Grumman and the Air Force Research Laboratory (AFRL), is a modular software toolset that supports fusion of large amounts of disparate data sets, visualization, project organization and management, pattern analysis, activity prediction, and various presentation aides.

CADRE (Continuous Analysis and Discovery from Relation Evidence), developed by BAE Systems, automatically links and matches data with user-specified threat patterns, evaluates alternative hypotheses, and automatically generates queries for more data.

TBM Reasoner and CAAT (Theater Ballistic Missile Reasoner and Course of Action Analysis) tool, developed by BAE Systems, is a knowledge-based decision aid that helps analyst rapidly identify time critical targets and named areas of interest from ground moving target indicator (GMTI) data.

Catalyst, developed by General Dynamics Advanced Information Systems, is a modeling application for representing and analyzing problems, particularly those pertaining to potential threats and adversaries' courses of action. Catalyst provides an explicit external framework that supports the analytic decision making process and allows analysts to develop a problem- or task-specific organizational structure that can be used for organizing incoming or discovered data. It also allows analysts to develop organizational structures in advance or to develop and refine their structures as data and information become available.

ITEA (Intermediate Text Extraction), developed by General Dynamics is a tool that exploits text documents and /messages (particularly the unstructured prose text portions of the documents) to support intelligence analysts in their job of supporting decision makers, mission planners, and war fighters. This structured information can then be stored in a database and used to feed analysis and visualization tools such as data browsers and link analysis displays. This can enable analysts to support war fighters and decision makers with mission planning, campaign assessments, and other analysis products and services.

IDP (Information Discovery Portal), developed by JANYA, allows an analyst to view important data from thousands of documents and integrates several search and browsing capabilities with visualization and charting tools. IDP also has the capability to create Cross-Document Entity Profiles. It is built with InfoXTract technology for text extraction.

Analyst Notebook, developed by i2 Technologies, provides powerful solutions for accumulating, investigating, analyzing, and displaying complex information and relationships. The Analyst Notebook is a graphical software product that is designed to display and analyze intelligence relating to an investigation. It provides a wide range of methods to support analysis, help navigation through large networks of data, unravel complex relationships, and discover underlying interconnections quickly.

SEAS (Structured Evidential Argumentation System) or **SRI Early Alert System**, developed by SRI International, aids analysts in predicting potential opportunities and crises. It is implemented as a web server that supports the construction and exploitation of a corporate memory filled with analytic products, methods, and their interrelationships, indexed by the situations to which they apply. Objects from this corporate memory are viewed and edited using a standard browser client, with the SEAS server producing temporary HTML based upon the contents of the SEAS knowledge base that constitutes corporate memory.

1.5 DEVELOPING METRICS FOR KNOWDLEGE DISCOVERY AND INFORMATION EXTRACTION

The task of defining metrics for knowledge discovery and information extraction tools can be a difficult and challenging process. It is a process that engineers within IFTB/IFED have struggled

with for some time. When developing metrics for these types of tools, the following functional performances are the basis for our metrics development:

- θ Accuracy
- θ Efficiency
- θ Stability
- θ Improvement
- θ Error handling

Other factors to consider include usability, adaptability, and scalability. Lab engineers have found many problems when trying to develop these metrics. It is hard to define usability and adaptability in the lab environment. Experiments are modeled to match real world intelligence missions and scenarios but it is very difficult to model these environments to provide an accurate assessment of the tools' capabilities. In addition, usability can be in the eye of the beholder – what works for one analyst and his or her mission may not work for another. Accuracy, which is vital to intelligence missions, can also cause problems when developing metrics to test these types of tools.

In the real world, ground truth is often impossible to determine. Ground truth in the lab environment can be replicated, but at the cost of using much smaller and less cluttered data sets than will be found in the operational world. Metrics need to effectively show what the tool can do. Often the inability to replicate the real world situation hampers the development of metrics as it pertains to accuracy.

For the other basic factors of metric development, the problems faced in the lab are not often data or scenario driven, but rather they are driven by the immaturity of the tools. Most of the tools in this lab are prototypes. The prototypes contain system bugs and other problems that hinder the ability to test efficiency, stability, improvement, and error handling. The lab continues to improve metrics to better understand and evaluate tools that will support the Air Force in the future.

2.0 SUMMARY AND ASSESSMENT

Several lessons were learned from the development of the IU lab that can be used to guide and improve the ongoing efforts in the SA lab. The main lesson that can be taken away from the development of the IU lab is how difficult it is to develop accurate metrics for knowledge

discovery and information extraction tools. Because of the nature of this technology and our inability to obtain real-world ground truth data, we have struggled to identify metrics we believe show the true capabilities of many of the tools that we have evaluated.

Another lessoned learned is that when dealing with prototype tools, it is difficult to evaluate tools that often break easily or are so unstable that is nearly impossible to complete experiments on them. It is also very difficult to modify data to be used with these tools. Academic institutions and companies that have little experience working with intelligence analysts or within the field of intelligence analysis developed many of the tools. As a result, many of the tools suffer these shortcomings:

- θ Too difficult (from a technical stand-point) to be operated by your typical intelligence analyst

- θ Require an enormous amount of time to access the large amounts of data that the analyst must examine

These issues negate their intended purpose of helping the analyst maximize their efforts. It also makes it extremely difficult to recommend the transition of these tools to the analyst and the organizations that rely on their accurate and timely products. AFRL/IFED has made contact with several agencies regarding the potential transition of tools to their analysts. At this point, we have had a difficult time transitioning tools because the tools are either unstable or are difficult for the analyst to use (or learn) from a training stand-point. It may take the analyst too long to learn the tool and time is of the essence for an intelligence analyst.

Another problem is the conversion of intelligence data to be used by the tools. Often these tools require information in a specific format. This conversion of information can be a lengthy process that often only the tool developers can perform as opposed to the local intelligence analyst. However, the IU/SA lab has made great strides to make contact with the Air Force, the DIA, and other agencies to introduce new tools to intelligence analysts when the technology is ready for transition. The SA lab will continue to look for tools that will help our intelligence analyst complete their jobs in the most timely and accurate manner possible.

Even with the many problems faced by the IU lab, the effort was a success. Tools that had languished on shelves were examined and evaluated to the greatest extent possible. Observations, comments, and recommendations were used to improve many of the tools, which led to their ultimate deployment or identification of future enhancements. Many of the tools were integrated with other tools, assisting in the development of both tools. Deficiencies found in many of the tools have led to the development of other tools, expanding the scope and need for additional research and development in the field of knowledge discovery and information extraction.