REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
18-03-2009	Final	From 15-12-2006 to 15-12-2007

4. TITLE AND SUBTITLE
Adaptive Multi-Modal Data Mining and Fusion For Autonomous Intelligence Discovery

5a. CONTRACT NUMBER
W911NF-07-1-0059

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Edward J. Wegman

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
George Mason University
4400 University Drive
Fairfax, VA 22030-4444

8. PERFORMING ORGANIZATION REPORT NUMBER
TR ARL 02

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
US ARMY RDCOM ACQ CTR - W911NF
4300 S. Miami Blvd
Durham, NC 27703

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO/ARL

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
This proposal addressed the autonomous discovery of relevant information in massive, complex, dynamic text and imagery streams. We began development of a prototype system to mine, filter and fuse multi-modal data streams and dynamically interact with the analysts to improve their efficiency through feedbacks and autonomous adaptation of the algorithms. The plan was to implement four core capabilities: 1) Text and image mining for feature extraction, 2) Multi-modal data fusion, 3) Agent-based adaptive information filtering, 4) Cognitively friendly information visualization. The focus in the first phase of the work was multilingual text search systems as well as geospatial mapping of documents and images.

15. SUBJECT TERMS
automated data mining, streaming data, geospatial Internet localization, Arabic language text mining

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Edward J. Wegman
U	U	U	UU	15	19b. TELEPHONE NUMBER *(Include area code)* 703-993-1691

Table of Contents

List of Illustrations

1 Statement of Problem Studied

Consider the plight of the military situational analyst, who is faced with multimedia sources that stream in data constantly. Data can be structured text, unstructured text, voice, images, and video. The data likely are not English language, the data are likely to be massive in scale, and the data are streaming. The premise of the project was that the analyst needs a system tool to integrate, filter, and present to the analyst for his or her consideration the data that are most likely to be useful. The tool should be a query system that must operate transparently and without significant human fine tuning. The plan was to implement four core capabilities: 1) text and image mining for feature extraction, 2) multi-modal data fusion, 3) agent-based adaptive information filtering, and 4) cognitively friendly information visualization. Together, these will enhance the capabilities of the analysts to discover, assess, and act on embedded intelligence in near real-time. Figure 1 shows an overview of the planned architecture.

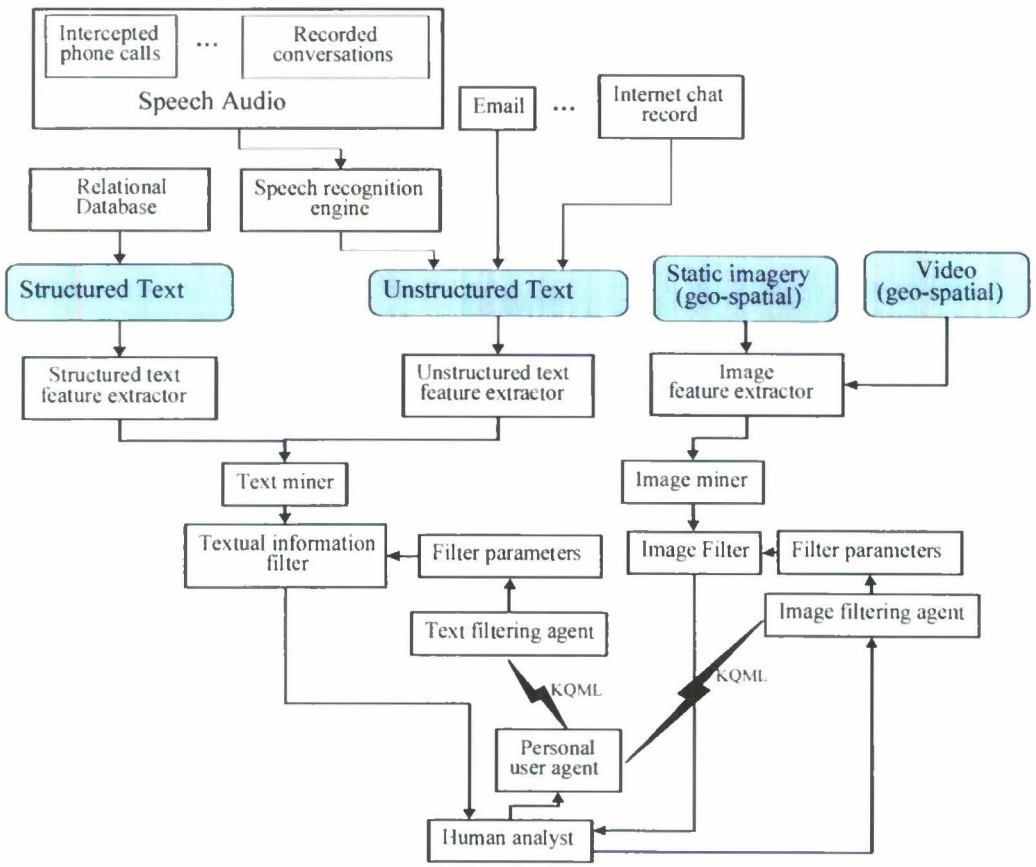

Figure 1: Overall architecture of the planned system tool.

The strategy for development of this rather ambitious tool was to divide the components of the architecture into modules that could be addressed by teams that were to work under the guidance of the investigators capitalizing on the particular skills of the graduate students involved. We had prior experience with speech recognition tools as well as with

relational data bases. The planned decomposition was to focus on the unstructured text feature extraction, on the geospatial imagery, video, and internet IP locators, and on streaming text data mining.

1.1 Mixed Language Text Database Search

A particularly useful component that was under development was on a mixed language text database search of open literature and intelligence documents. Because of the impetus from the fighting in Iraq, we initially developed a prototype for an Arabic and English mixed database. This project was directed at building a bilingual Arabic/English Web-Based Text Processing system that provides analysts with search results in Arabic based on English language queries. A massive amount of multilingual open source text data is available. Arabic is the official language of Algeria, Bahrain, Comoros, Chad, Djibouti, Egypt, Eritrea, Iraq, Jordan, Kuwait, Lebanon, Libya, Mauritania, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Somalia, Sudan, Syria, Tunisia, United Arab Emirates, Yemen, Western Sahara, African Union Arab League, and the United Nations. The structure of Arabic is quite different from Western languages not only because of the alphabet, but also because of conceptual difference in construction of words and sentences. Typically, vowels are omitted from written Arabic, which is written from right to left. There are approximately 10,000 independent roots and each root can have multiple meanings.

Particles of construction are the letters used to form a word. The particles of signification are used to form sentences. Particles of signification are used to modify verbs, nouns, or both. In Arabic, a noun is a word that indicates a meaning by itself without being connected with the notion of time. Nouns can be, masculine or feminine, qualified or qualificative, definite or indefinite, and diminutive or relative. Nouns also can be singular, plural, or dual. Singulars have more than one plural form and some plurals are not derived from the singulars. Nouns can be proper nouns. A proper noun is the name of a specific person, place, organization, thing, idea, event, date, time, or other entity. Proper nouns can be masculine or feminine, simple and composite. Proper nouns in Arabic do not start with capital letter as in English, which makes it complicated to try to extract them by machines. A simple proper noun is the proper noun composed of only one word, e.g., Muna (feminine) منى, Muhammad (masculine) محمد. Arabic names like Abdul Rahman عبدالرحمن , Abu Saleem أبوسليم, and Abdu Allah عبدالله are considered composite proper nouns. Composite proper nouns are composed of more than one word.

Because automatic language translation is an extremely difficult task, we were not attempting automatic language translation, rather we are suggesting to the analyst a reduced set of Arabic language documents he or she might consider. Arabic is one of the most common languages in the world, but technology has been slow in development for Arabic. This is mostly due to the complexity of the written structure of the language. Unfortunately due the language differences, this technology is usually limited to the language in which it was developed (usually English) and cannot be easily transferred to in different linguistic environments.

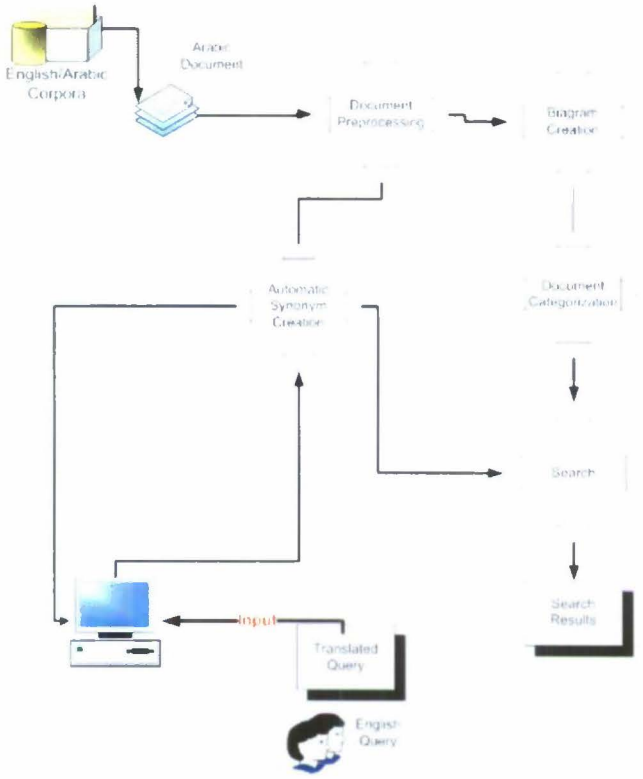

Figure 2: Block diagram of Arabic-English Text Mining System

In a preprocessing phase, Arabic language documents are background processed, stemmed, and de-noised, i.e. stop words are removed. We developed a capability for synonym creation based on LSI techniques. The version of the system we developed currently supports 81,500 Arabic/English words. For every document in the corpus, the *bigrams* are constructed. A bigram is a word pair where the order of the words is preserved. For example, if the sentence is "Hell hath no fury like a woman scorned." The stemmed and denoised sentence is "Hell has no fury like woman scorn." The bigrams are: *hell has*, *has no*, *no fury*, *fury like*, *like woman*, *woman scorn*, and *scorn .*; The sentence ending full stops (like . ! ? : ;) are treated as a word for purposes of bigram development. Bigrams are used as terms in the system so that searches can be made on terms like *nuclear weapons*, *biological weapons*, and *mass destruction*. The demonstration database can be bilingual Arabic and English and it currently supports html, Microsoft Word, and text documents. Figure 3 is the enhanced block diagram. Figure 4 is a screen shot from the system. This system, while not implemented is extensible to a Korean-English system. The work reported here was implemented by Eiman Alshammari, a Ph.D. student who is a native Arabic speaker, under the supervision of Dr. Edward J. Wegman and Dr. Yasmin H. Said, who is also a native Arabic speaker and a U.S. citizen.

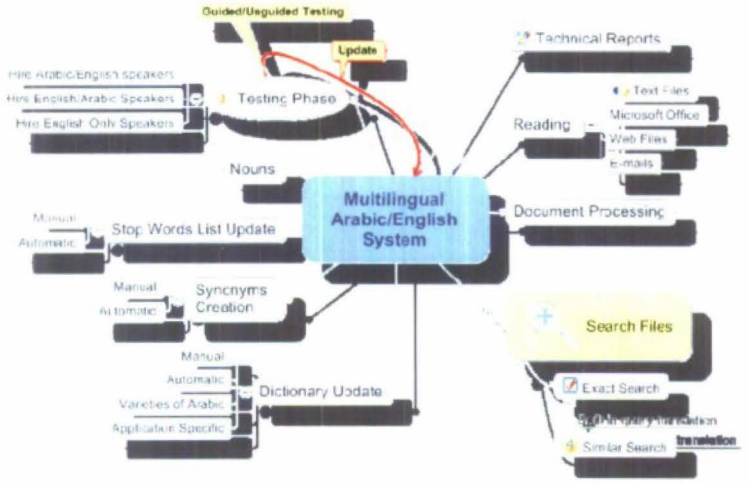

Figure 3: Enhanced block diagram for the Arabic-English System.

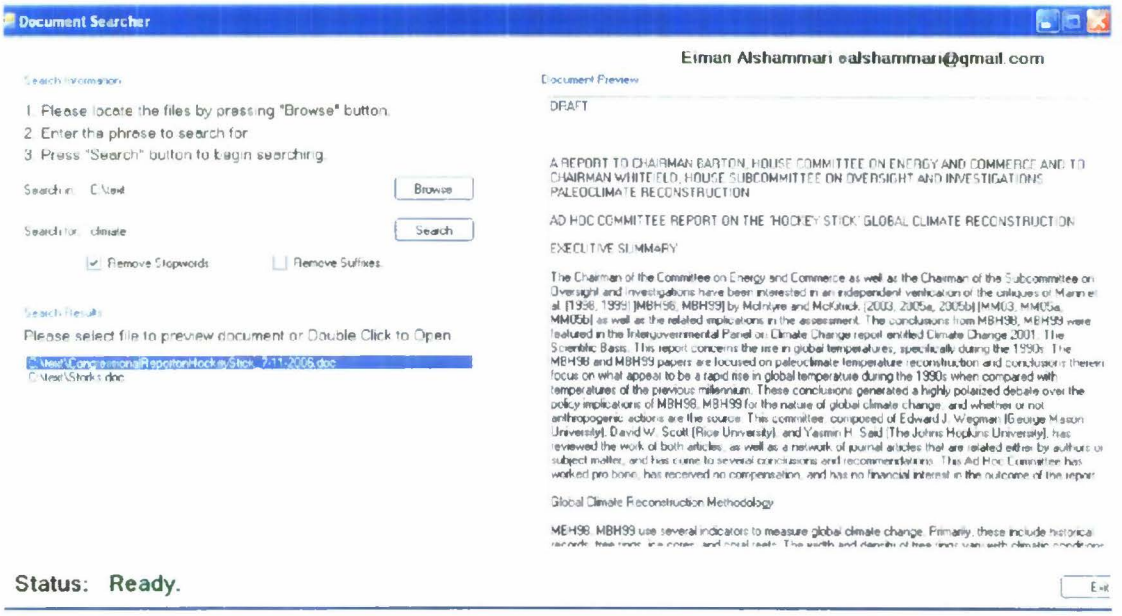

Figure 4: Document search screen shot.

1.2 Streaming Text Data Classification

Text processing is usually performed on a fixed corpus of text. The natural question to ask is what to do in the case of streaming news articles, weblogs, military reports, or even research articles when you want to examine them as they evolve in time? Traditionally, in text mining, each document is represented as a vector, $x_i \in \mathbb{R}^L$, $i = 1, \ldots, D$ where D is the number of documents in the corpus and L is the number of words in the lexicon. Each dimension of the vector corresponds to a different term in the lexicon. Then, x_{ij} is the TFIDF weight for word j in document i. TFIDF is the so-called term frequency inverse document frequency weight. TFIDF is proportional to the number of times the j^{th} term in the lexicon appears in document i and is inversely proportional to j^{th} term's rate of occurrence in the corpus. Specifically, $x_{ij} = TF_{ij} log\left(\frac{D}{b_j}\right)$ where TF_{ij} is the number of times the j^{th} term in the lexicon appears in document i, D is the number of documents in the corpus and b_j is the number of documents that contain the term j. Terms that occur in every document will have $\frac{D}{b_j} = 1$ so that $x_{ij} = 0$. In contrast, relatively rare words will have a large weight.

Steaming text data presents an alternate problem because the size of the lexicon and the size of the corpus are changing with time. Suppose at time t a new document is observed. Let $X_j(t) \in \{0,1\}$ indicate whether term j occurred in the document read at time t. Then the approximate document frequency for word j at time t is

$$DF_j(t) = \alpha X_j(t) + (1 - \alpha) DF_j(t - 1),$$

an exponential window on the document frequency so that at time t the

$$\text{TFIDF} = TF_{ij} log\left(\frac{1}{DF_j(t)}\right).$$

We are interested in tracking changes in article topics. To this end we suppose that there are allowing for N topics to be considered. We formulate a graph model. A graph G consists of $V(G)$, the set of vertices or nodes that represent the N topics, and the set of edges $E(G)$ where $v_i v_j \in E(G)$ is an edge from vertex i to vertex j. This is a directed graph. The number of vertices is the order of the graph and the number of edges is the size of the graph. For a graph of order N, the adjacency matrix A is an $N \times N$ matrix with $a_{ij} = r_{ij}$ if $v_i v_j \in E(G)$ and $a_{ij} = 0$ otherwise. Node represent topics, i.e. accumulations of similar articles. Each node has an associated set of words and counts of those words from articles that have been assigned to each node. The edge weights, $r_{ij} \in [0,1]$, encode similarities between the nodes. A new article get assigned to the node to which it is most similar based on the cosine similarity. If the new article similarity is less similar to any node than some threshold, τ, it initializes a new node or resets an existing node.

In order to maintain the changing lexicon, currently the word counts for each of the N nodes are kept in a $N \times L$ matrix, X, which will be a very sparse matrix. Once the

lexicon reaches its maximum, L, the oldest m terms are eliminated. The corresponding columns of X are eliminated so that the new X will be an $N \times (L-m)$ matrix. Then add new terms (columns) onto X. In order to update graph nodes, Suppose a new document is read. Let $z_i = (z_{i1}, \ldots, z_{iL})$ be the count vector for the terms in the document. Suppose the document is assigned to node n. Let z_n be the count vector for the node n prior to the assignment so that $z_n = (z_{n1}, \ldots, z_{nL})$. Then the entries of z_n will be updated by $z_{nk} = z_{ik}$ if $z_{nk} = 0$ and $z_{nk} = \beta z_{ik} + (1-\beta) z_{nk}$ otherwise. This implements an exponentially smoothed average for the counts in the node. β is adaptive. Let s be the similarity between the current article and the closest graph node. Let τ be a threshold so that any article with the maximum similarity less than τ resets the graph node. Let β depend on the maximum similarity by $\beta = 1 - s$ if $s \geq \tau$ and $\beta = 1$ if $s < \tau$. This makes large changes in the node when similarity between the document and the node is small, and small changes when the similarity is large.

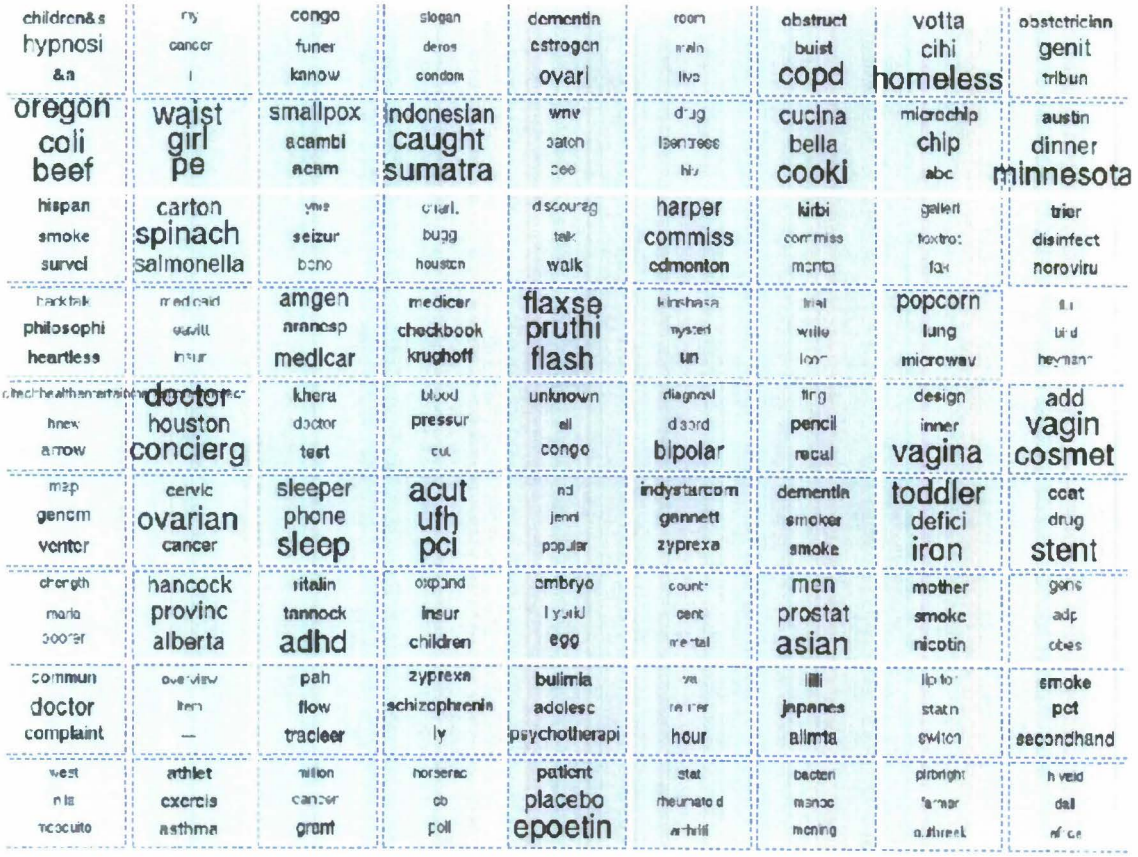

Figure 5: Example of node tracking by the graph-based algorithm. Here there are 81 nodes. The size of the font represent the frequency of the word in a particular node. This is a screen shot from an animation that can be found online at http://www.galaxy.gmu.edu/stats/colloquia/AbstractsFall2007/HEALTHNEWS_FAST.GIF

This work was tested on a Google News dataset. It consisted of approximately 350 articles per day, 70 in each of five categories, World, US, Business, Science and

Technology, and Health. This work was carried out by Dr. Elizabeth Leeds Hohman under the supervision of Dr. Edward J. Wegman. It was part of her Ph.D. dissertation.

1.3 Transient Geographic Mapping System

The geolocation tool is designed to enable real-time identification of incoming threats and attacks. The geospatial visualization tool is used for both display and query. The intent is to provide a tool to: 1) geolocate the source IP address of data packets on a world map projection with the idea of locate potential hostile attacks, 2) locate image and video sources based on geospatial metadata, and 3) query a database using geospatial coordinates to obtain multimedia documents. The application was written using Windows API application interface, C++, OpenGL for animation, and the WinPcap library for packet handling. The IP address geolocation tool was based on world map data courtesy of Professor David Wong of GMU, IP location data courtesy of GeoBytes, and GMU traffic data developed under a project with Dr. Wegman as PI funded by the AFOSR. Three projections were considered. The Mercator projection is a conformal mapping that preserves angles and is good for zooming and panning, the Plate Carrée is the simplest to implement and the Winkel Tripel. The latter minimizes distortion so is best for whole Earth maps. On the downside, it is hard to invert, which is only possible numerically.

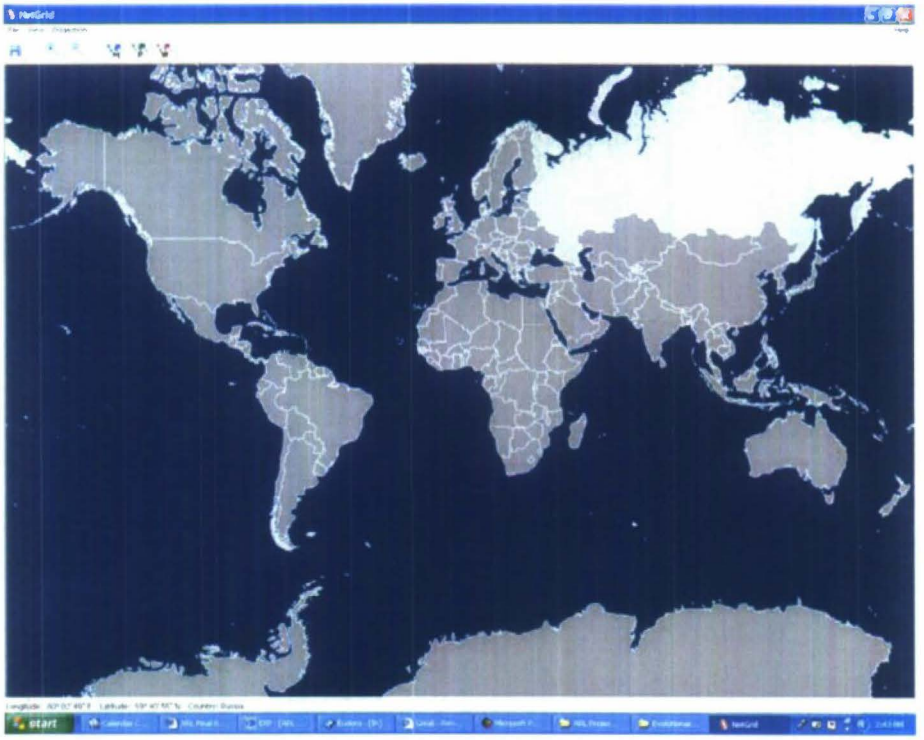

Figure 6: Mercator projection of map application with Russia highlighted.

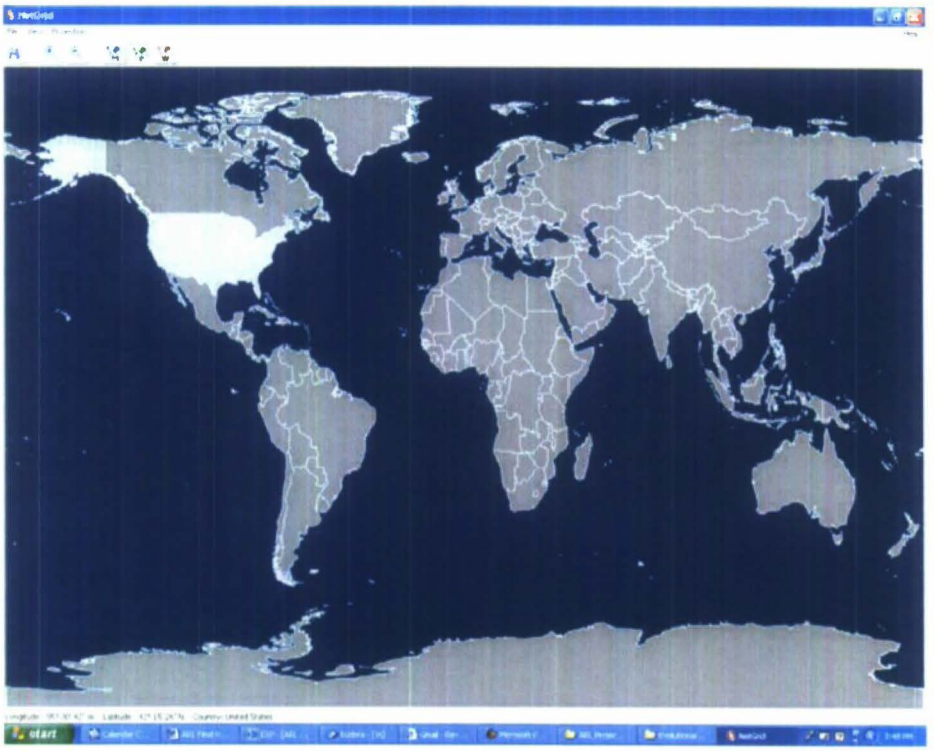

Figure 7: Plate Carrée projection of map application with USA highlighted.

Figure 8: Winkel Tripel projection of map application with Antarctica highlighted.

The maps are based on a Delauney triangulation of the surface of the Earth. The drawings are based on the OpenGL software tool. The idea of the transient geographic mapping is to geolocate and event or an object in real time. The point location algorithm locates the Delauney triangle where the given destination point belongs. Once the triangle is known, the country to which the triangle belongs is identified by a table lookup. The algorithm has pre-calculated index points in 1 degree increments. The algorithm starts from the closest pre-located index point and searches triangle by triangle until it locates the triangle to which the given destination point belongs. The complexity for the point location algorithm is $O(n^{\frac{1}{3}})$ where n is the number of points. Complexity is improved by using the pre-calculated index points rather than a unique starting point. The GeoBytes database is a database for IP location that provides for each subnet latitude, longitude, city, region, and country. Thus the GeoBytes database provides a the lat-long information and the point location algorithm locates that lat-long information on the map. The WinPcap library provides low-level network access for reading packet headers either real-time or from a stored database. The incoming IP address is read from the packet header and is identified geospatially from the GeoBytes database. The basic idea of transient geographic mapping was to display the source IP geographically in order to dynamically follow where packet headers are coming from, in particular, to rapidly identify whether they were coming from hostile adversaries. The idea is that for each packet, we display a point at the latitude and longitude found in the GeoBytes database. The point has a fading time so that many packets coming from the same source will be represented by a bright point. The default fade time is 1 second, but this can be adjusted. Each point fades out gradually using OpenGL alpha transparency and alpha blending. This work was carried out by In-ja Youn and Felix Mihai under the direction of Dr. Edward Wegman. It is based on an idea suggested by Wegman and Marchette (2003).

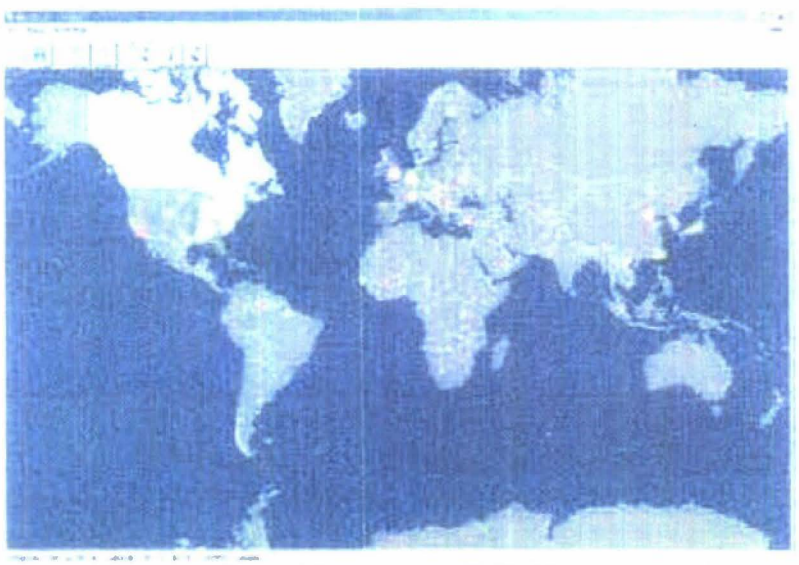

Figure 9: Mercator projection with IP point sources lit by intensity of traffic. Green sources are friendly, red are hostile, yellow is unknown.

2 Summary of Most Important Results

The overall idea of the project was to develop a tool to aid the analyst in identifying information rapidly that may be of interest for command decisions. The overall system diagram is included in Figure 1. The strategy we employed employed was to modularize the system outlined in Figure 1 and to employ Ph.D. students to develop subsystems whose implementation had enough intellectual challenge to be suitable for Ph.D. dissertations. With continuing funding, these modules would be integrated into the final system. The main achienvements of the current project funding were:

- The development of a multi-language search system using terms entered in English, but capable of finding relevant documents in multiple languages. The demonstration languages were English and Arabic, although a Korean-English version could also be rapidly developed based on the prototype. The proptotype system was implemented on a windows platform.

- The development of a streaming text classification system capable of automatically identifying evolveing topics. The prototype systems was also built on a Windows platform with test data deriving from Google News feeds.

- The development of a real-time dynamic geospatial location and query system system for locating incoming IP traffic as well as identifying geospatial location of multimedia documents. The latter functionality was not implmented.

Several papers were developed based on research carried out under this project. These papers expolited aspects of the developments here. They are Said et al. (2007, 2008), Said and Wegman (2009), and Wegman and Said (2007). Presentations were given in a number of forums that credited this contract.

3 Bibliography

Geobytes IP Locator Files, http://www.geobytes.com.

Ipbuker, Cengizhan and Bildirici, I. Oztug (2005) "Computer program for the inverse transformation of the Winkel projection," *Journal of Software Engineering.* 131(4), 125-129.

Mucke, E.P., Saias, I. and Zhu, B.-H. (1996) "Fast randomized point location without preprocessing in two- and three dimensional Delaunay traingulations," In *Proceedings of the 12th Annual ACM Symposium on Computational Geometry.* 274-283.

Shewchuk, J.R. (2002) "Delaunay refinement algorithms for triangular mesh generation," *Computational Geometry: Theory and Applications.* 22(1-3), 21-74.