SECURITY SCREENING AND KNOWLEDGE MANAGEMENT

IN THE DEPARTMENT OF DEFENSE

David Lloyd and Dr. Nancy Spruill

Office of Undersecretary for Acquisition, Technology and Logistics

Abstract

One of the biggest fears of operators of unclassified e-mail and other document systems in the Department of Defense is that classified material will get posted to these systems. The Office of the Under Secretary of Defense for Acquisition, Technology and Logistics has developed a full-text scanning package to search for classified data and other unauthorized network traffic. This paper will discuss that package, which is a natural language, real-time process for scanning text material. It uses advanced stemming, parsing and relevance ranking algorithms to detect unauthorized text material. For example, it is capable of discerning the difference between prose and outline formats. The types of network traffic filtered include mail messages/attachments, collaborative system documents, and correspondence. It is capable of filtering 25,000 7-page message/attachment packages per hour on-the-fly. It reads PDF, HTML, native Lotus Notes, and 200 other common formats including zipped attachment files. This package is written for NT or Linux environments and is integrated with MS Outlook and Lotus Notes mail systems. The paper will also discuss the more complicated full-scale, fully automatic taxonomy generation project, which uses the same engine and is being developed to provide the Under Secretary's staff with the capability to dynamically create a table of contents for any set of collections of documents. This enables the user to organize large, disparate collections of text information in one convenient nested outline format for issue research purposes. Any given topic or subtopic in the outline is linked to a list of relevance ranked source documents which can be displayed as required by the user.

Keywords: Search Engines, Network Security Screening, Knowledge Management, Taxonomy Generation, Statistically Based Text Searching, Local Area Context Model

I. Knowledge Management in OUSD(AT&L)

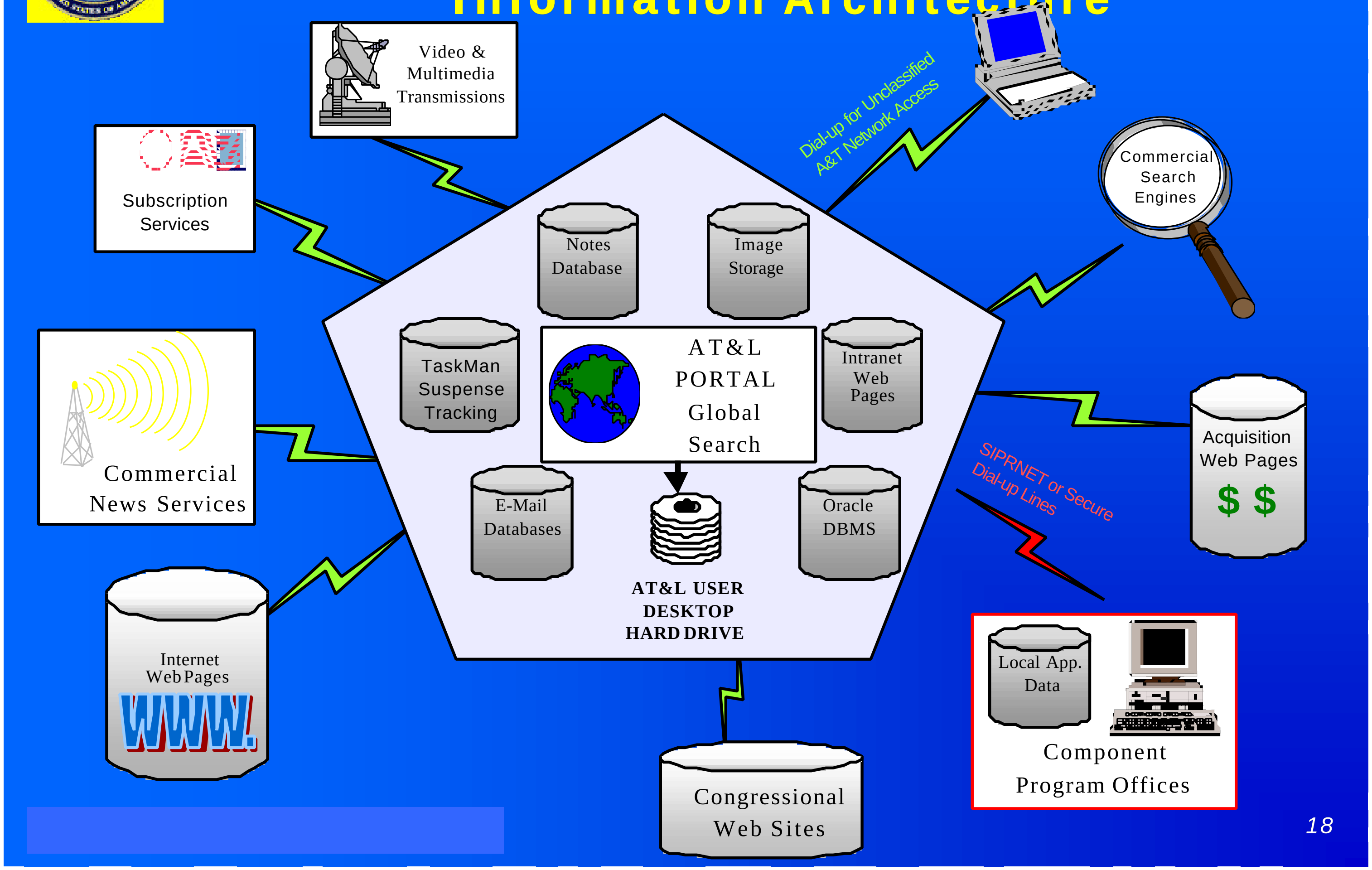
OUSD(AT&L) is responsible for overseeing and developing DoD policy in a variety of areas in acquisition, technology, development and logistics and provides computer support to approximately 1,500 people. Our organization is no different from most in having to deal with the tremendous increase in information available for policy decisions as well as the phenomenon of developing policies in a compressed window of time.

This new development in information management resolves itself into essentially two components: (1) The ability to easily and conveniently organize information from multiple sources such as email, correspondence, legacy data bases, legislation, various office automation applications, handheld devices and; (2) The ability to conveniently, in an organized way, to retrieve, view and transmit this information.

OUSD(AT&L) has viewed this as a multi-front initiative and has implemented a knowledge management program using three basic technologies:

> **A. Portals:** This allows users to access and display a variety dynamically updated sources of information and automated applications both inside and outside of the organization on one screen without having to go to the special application areas and open up the applications. OUSD(AT&L) has selected Plumtree Inc. for use in the knowledge management portal application project.
>
> **B. Automated Taxonomy Generation:** This allows multiple collections of data to be quickly organized as one database for use in policy issue resolution. Multiple sources of internal legacy databases, the Internet, email, calendars, formal correspondence, news services and various special automated systems can be summarized under one category, topic or sub-issue. This taxonomy is generated on-the-fly without need for "key words" or any other effort at cataloguing.
>
> **C. Network Scanning and Retrieval.** Retrieval requires a robust search engine that can be used against multiple types of databases and formats of data. In addition, the OUSD(AT&L) knowledge management program included a requirement for real-time scanning of network correspondence and email on issues being staffed in order to provide for quick coordination of staff issues on very short deadlines. The scanning technology was first applied to security requirements. The requirement for being able to detect a classified document on the unclassified mail system became an early priority and this application was made operational first. The intent is to use this same technology for the broader use of reporting network traffic on specific issues when deadlines require such rapid coordination.

AT&L Knowledge Management Program: Information Architecture



Video & Multimedia Transmissions

Subscription Services

Commercial News Services

Internet Web Pages

WWW.

Dial-up for Unclassified A&T Network Access

Commercial Search Engines

Acquisition Web Pages

$ $

SIPRNET or Secure Dial-up Lines

Local App. Data

Component Program Offices

Congressional Web Sites

Notes Database

Image Storage

TaskMan Suspense Tracking

AT&L PORTAL Global Search

Intranet Web Pages

E-Mail Databases

Oracle DBMS

AT&L USER DESKTOP HARD DRIVE

II. Network Security Screening in OUSD(AT&L).

A. Requirements

As noted above, this application grew out the more general effort to monitor network message traffic on a real-time basis to satisfy the need for close coordination under deadline conditions. Harvest Mail addresses the fundamental security problem of detecting classified mail and classified attachments traversing the network on a real-time basis. This product is a full-text natural language scanning package that uses a document analysis process to resolve previous problems usually associated with most of the commercial products. The most serious problem in this area is the staff time required to resolve whether or not a message and/or attachment flagged by the system is really classified. In those cases when the system flags a document but the document is not really classified, a false positive has occurred with the system. Clearly, most government staffs cannot spend large amounts of time resolving false positives delivered by the software.

Thus, the security screening requirements must meet competing requirements:

> Sensitivity: The ability to flag for analysis any and all documents which should be sent to the software analysis module for evaluation as to whether the text is "really" classified. Theoretically, we would want this to be 100%.
>
> Precision: Once the document has been flagged for analysis, the analysis module would make the correct decision as to whether the document really is classified or not and then appropriately notify the security officer with priority email and stop transmission of the mail package. Ideally, this error rate would be small - preferably 0 on 100% of the documents forwarded for analysis.

As a matter of practicality, if ALL documents are flagged, (sensitivity = 100% in all cases) performance of the system would suffer since the subsequent analysis of all text would require either unacceptable delays or require a huge investment in processing equipment. The fundamental issue is the appropriate trade-off under acceptable management risk.

B. Screening Process for Classified Documents. In order to arrive at the appropriate trade-off, heuristics must be applied to the set of events in a set volume of mail traffic and a set of tests run.

The event categories are as follows:

1. Email correctly flagged as potential problem and

correctly identifies markings which appear to make the document classified.

2. Email processing which results in false negatives.

(Email which should have been flagged as potential problem but which was not.)

3. Email processing which results false positives.

(The software correctly flags document as potential problem but incorrectly decides markings are classified markings.)

4. Email that has been correctly assessed as no potential
problem and not passed for further analysis.

Thus, the general formulas for sensitivity and precision based on a daily load of 30,000 messages (message and all attachments are counted as one document) would be:

Sensitivity = D/D+B or 30,000/30,000 + 1 or 99.997%

Precision = D/D+C or 30,000/30,000 + 9 or 99.97%

In the above example, one document would miss being flagged as a potential problem (sensitivity) and nine documents would be falsely determined to be classified by the software after the document analysis (precision).

In terms of staff time spent on false positives, the goal was to have precision levels set at about 99.97%. For our stream of 30,000 messages per day, this meant that no more than about 8-10 documents per day would be reported as false positives by the software. Thus, 8-10 documents per day would have to be examined by a security staff specialist to make sure the false positive was really false.

In those cases where classified information is inadvertently transmitted with no markings indicating classification, the system currently has no way to detect a problem. However, since Harvest Mail is based on natural language interpretation of all text, it would be possible to intercept certain classes of information based on special descriptors entered into the "filter" for the package. For example, documentation such as security guides for specific areas could be the basis for such filtering.

C. Implementation Experience To Date.

With respect to the metric on sensitivity, we could afford greater sensitivity since the processing load for detecting what might be a problem was less. Thus, for our mail stream, we could afford to set up the software so that only one mail message in 30,000 should actually be considered a potential problem by the software but the software would miss that potential problem. There is, of course, no way one can establish whether the software is missing potential problems for certain. Our level was initially based on a test data set sent through. However, we have the advantage of the mail being read by humans and they usually do report security violations. Thus, we do have a reasonable empirical measure of the sensitivity metric as operational experience continues.

The software did exhibit two cases of false negatives that we know about in a stream totaling 3.6 million messages over six months: One document was passed through incorrectly because a certain NATO security marking was not sensed. The software was modified to detect it. A staff member reported that violation. In another case, the ink was smeared on the document so badly, the PDF document was not properly interpreted by the software. This violation too, was reported by staff.

D. Other Possible Implementations.

The broader knowledge management applications of the scanning tool have not yet been implemented in the OUSD(AT&L) due to the effort spent for the security screening efforts. The scanning software could be used in a similar way to "flag" messages on critical issues such as pending legislation reaching deadline votes on the Hill.

Other broad applications of the scanning system would include the ability of senior managers to conduct an audit of subject interest areas for all topics in the organization based on both real-time filtering and the filtering of legacy mail archives. This could enable senior managers to have visibility of all of the personnel who are working and staffing certain topic areas based on their email exchanges during the normal course of work. It has often been the case that expertise in a given area was available but various management levels were simply unaware the experience and talent was there. Such a personnel talent tool used in conjunction with the taxonomy tool described below might be useful in planning reallocation of staff. Privacy issues and personnel awareness issues would first have to be resolved prior to this type of implementation.

III. Taxonomy Generation in the OUSD(AT&L) Knowledge Management Program.

A. Background of OUSD(AT&L) Taxonomy Issue

Senior managers in OUSD(AT&L) were informally surveyed in 1993, as part of the knowledge management planning process, to determine what the fundamental categories of information they dealt with in the daily course of resolving policy issues. Not surprisingly, the three categories mentioned the most often were simply people, organizations and the topics themselves.

OUSD(AT&L) investigated ways of using commercial software to sift through mountains of issue background material, such as legislation, to develop a "table of contents" or taxonomy of all of the collections of data bases available in the organization. At that time, there was no software we were aware of that could even approach the problem. The Library of Congress was contacted and meetings were held to address what seemed like a monumental problem. In 1994 the Library, OUSD(AT&L), Dept of Commerce, the Social Security Administration and other government and Defense organizations joined a Consortium of vendors and academic institutions noted at the beginning of the paper: The Center for Intelligent Information Retrieval (CIIR) at the University of Massachusetts at Amherst. The intent was to resolve the fundamental problems associated with searching and organizing very large collections of data in multiple formats and locations. Each member paid $25,000 dues for free use of licenses of developing software solutions and the ability to set requirements and influence operating designs. Since that time, two of the products developed into commercial products and have been integrated as part of the OUSD(AT&L) knowledge management program: Harvest Mail, a network scanning system, and Athena, an automated system for generating automatic ad hoc taxonomies on large collections of data. Both products are available commercially from Chiliad Corp.

B. Taxonomy Generation Using Athena

The OUSD(AT&L) requirements submitted to the CIIR originally for the generation of taxonomies which would assist in the organization staff work included the four fundamental characteristics:

1. Categorization of information based our perceived need for resolution of policy options and decisions:

   a. People

   b. Organizations/Companies

   c. Topics along with display of indefinite numbers of levels of topics making available for display short summaries of all relevant documents relating to the level selected.

2. Taxonomy is dynamic and data driven. As data is updated in the organization, the new data is incorporated into the new taxonomy real-time. (No human categorization or key-word usage is required.)

3. Ability to include any type of data or from any automated source including but not limited to email, legacy data files, internal organizational web sites, the Internet, voice-mail, teleconference streams, and various hard-copy files capable of being scanned.

4. Ability to effectively and conveniently "drill down" from the topic or subtopic presented to the user in the taxonomy directly to the ability to view the source document. Athena is able to carry out these functions. (Voicemail would be treated as text after interpretation by voice recognition software.) The architecture used by Athena is web-based and uses distributed servers where all servers in the system can communicate and transfer appropriate information enabling the creation of one virtual data base

C. Brief Description of Athena Engine

There are basically two general approaches to the overall architecture typically used:

1. The query is parsed and interpreted by the engine and put into a Boolean form of the request then submitted for the basic search against the indexed file.

2. The query is parsed and interpreted using a relevance ranking scheme based on probabilistic frequencies of occurrences of terms, words, phrases and their relative proximity to one another or related terms in the document or collection of documents.

Athena is a mix of both approaches with the intent of optimizing performance based on performance records and heuristics gleaned to date from researchers in the field.

The set of collections is pre-processed to form one unified set of text. Although the identity of each document is retained for later use, the entire set of text is treated as a whole. The set is broken up into pseudo-documents and each word is stemmed and parsed in each pseudo-document. The first process is to set up an analysis of the "query". Since, in the case of the generation of the taxonomy, each "topic" becomes a query applied to the set as a whole, the query analyzer creates a set of topics based on proximity and word frequency from the pseudo-document. The resulting query is then used to develop the Local Context Analysis process, which then results in the development of the taxonomy as a whole.

D. The Local Context Analysis Engine Used As Basis for Taxonomy Generation.

Based on the initial query analysis and the parsing and stemming in each pseudo-document, a list of top passages is created. The window for the selection of a passage in the text is arbitrarily set but might range from 120 to 300 characters. The list of passages (totaling N passages), is used as a basic start list. The first of the passages now becomes an independent query applied against the entire collection of passages. Other passages relevant to the first one are scored and the taxonomy is developed from the accumulated scores.

The scoring system is used for the fundamental ranking system that ranks individual words, groups of words and entire passages. The scores are ordered from low to high and form the basis of the value of a given document in the collection to a given topic and its children.

The ranking process results from a "belief" score that is computed in the following way:

$$bel(Q,c) = \underset{T,eQ}{Min}\ F(tf,\ idf_{c,}\ idf_i)$$

Where:

tf (term frequency) rewards topics co-occurring frequently with query terms.

idf_c (inverted document frequency topic) penalizes topics occurring frequently in same candidate passage set

idf_i (inverted document frequency for infrequent terms) emphasizes query terms that are infrequent in the passage.

Statistics accumulated on:

a. The number of occurrences of any query term t_i in any passage p_j

b. The number of occurrences of any concept c in any passage p_j.

c. The number of passages in the candidate sub-documents list (N).

d. The number of passages containing query term $t_l(N_i)$

e. The number of passages containing topic c (N_c)

Upon valuation of the minimum function, the top M topics are selected to be the results. The process is then iterated for each topic type for lower levels of topics.

The parsing and term recognizer modules select from the set of collections the set of companies/organizations. A second set of people's names is developed from the collection set. Each item the two lists is then evaluated against the M set of topics and scored and ranked.

The resulting list of people has associated with it a ranked list of topics for each person.

Likewise, the resulting list of organizations/companies has associated with it a ranked list of topics for each of the organizations/companies.

At the third level of the selection of the topic from a given person, for example, the results of that level include a combination of people, organizations or topics. In other words, at the lower levels in drilling down, "all" relevant categories associated with that item selected. The resulting lists of "items of interest" may be from any of the three categories.

The architecture of the engine is scaled for very large collections. Currently, the engine can easily handle over 760GB of data in one set of collections. The distributed architecture allows sizes of one terabyte to be processed for each server. The engine has built into it its own web server system and its own web crawler. The engine can collect and index files in any of 200 formats using internal and external data bases and web sites.

IV. Summary.

OUSD(AT&L) has implemented two applications, network scanning and taxonomy generation, as part of its current deployment of a portal-based knowledge management system. The two products, Harvest Mail and Athena, are now operational in the organization and have been successful in reducing the chances of network document security violations and in assisting staff and executives in locating and organizing relevant information for policy issue resolution.

Acknowledgements

We thank Forrest Landry, the chief architect of the Athena product, and Jeff Gelbard, the chief architect of Harvest Mail, for reviewing and making many useful suggestions regarding technical issues presented in this paper. Mr. Landry and Mr. Gelbard are currently Senior Scientists at Chiliad Inc., Amherst, Mass.

References

1. Jinxi Xu and W. Bruce Croft, "Query Expansion Using Local and Global Document Analysis", Proceedings of ACM SIGIR, 1996.

2. G. Salton, "Automatic Test Processing: The Transformation, Analysis and Retrieval of Information by Computer", Addision Wesley Longman, Reading, Mass. 1989

3. E. Keen, "Term Position Ranking: Some New Test Results," Proc 15th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, 1992, pp 66-76; available on line at http://www.acm.org/pubs/citations/proceedings/ir/133160/p66-keen/.

4. D. Hawking and P. Thistlewaite, "Proximity Operators - So Near and Yet So Far," Proc. Fourth Text Retrieval Conf., D. K. Harman, ed., 1995; available online at http;//web.soi.city.ac.uk/~andym/PADRE/trec4.ps.Z.

5. R. L. Larson, "Directions for Defense Digital Libraries", D-Library Magazine, July/August 1998; available online athttp://mirrored.ukoln.ac.uk/lis-journals/dlib/dlib/dlib/july98/07larsen.html