METHODS AND MEANS USED IN PROGRAMMING INTELLIGENT SEARCHES OF TECHNICAL DOCUMENTS

David L. Gross
Computer Engineer
Analex Space Systems, Inc.
NASA Kennedy Space Center
Post Office Box 21206
Kennedy Space Center, FL 32815-0206
Phone: 407/861-5716
Fax: 407/861-5774

ABSTRACT

In order to meet the data research requirements of the Safety, Reliability & Quality Assurance activities at Kennedy Space Center (KSC), a new computer search method for technical data documents was developed. By their very nature, technical documents are partially encrypted because of the author's use of acronyms, abbreviations, and shortcut notations. This problem of computerized searching is compounded at KSC by the volume of documentation that is produced during normal Space Shuttle operations. The Centralized Document Database (CDD) is designed to solve this problem. It provides a common interface to an unlimited number of files of various sizes, with the capability to perform many diversified types and levels of data searches. The heart of the CDD is the nature and capability of its search algorithms. The most complex form of search that the program uses is with the use of a domain-specific database of acronyms, abbreviations, synonyms, and word frequency tables. This database, along with basic sentence parsing, is used to convert a request for information into a relational network. This network is used as a filter on the original document file to determine the most likely locations for the data requested. This type of search will locate information that traditional techniques, (i.e., Boolean structured key-word searching), would not find.

INTRODUCTION

The need to search technical documentation for desired information is a labor intensive activity. In the past, data searches have been restricted to human effort with limited computer searching, (generally Boolean key-word searching). This is primarily due to the type of information that is being searched and referenced. Technical documents are partially encrypted by the author's use of acronyms, abbreviations, and shortcut notations. At Kennedy Space Center (KSC), this problem is magnified further. A researcher who is searching for information based on an engineer's or a technician's notes is faced with notes that are usually more encrypted and/or abbreviated than those which are contained in the actual document. The problem is further compounded by the volume and dispersal of documentation that is produced during normal shuttle operations. The CDD addresses these problems. The commercial potential of this system is evident from the savings in man-hours alone. Any profession that devotes time to specific subject review and research would benefit greatly from this time-saving system, (e.g. legal, medical, information specialist, etc.).

BACKGROUND

In 1990 NASA funded a project to improve the data retrieval and dissimilation methods used by the Safety, Reliability & Quality Assurance (SR&QA) directorate. Systems and quality assurance reviews were identified as likely candidates for improvement. This procedure requires accessing a large number of technical documents and uses a large percentage of available man-hours. A project was initiated to develop a more time-efficient method of doing these searches. Several commercial packages were evaluated, but none met SR&QA's needs. Finally, a decision was made to develop custom software.

Software algorithms from the Artificial Intelligence (AI) field were used in an attempt to duplicate human search methods. The three methods that showed promise were:

1. Sentence parsing used in natural language processing
2. Confidence factors or weights from heuristic searching
3. Network connection and propagation from connectionism

Parsing analyzes the syntactic structure of sentences. To adapt this technique to technical data queries, parsing is used to identify word and phrase relationships such as subject-verb, verb-object, and noun-modifier (Figure 1). The parser uses knowledge of language syntax, morphology, and semantics. In technical document searches, sentence parsing is used to identify word types, (i.e. noun, verb, adjective) based on the context in which the abbreviation or acronym is used.

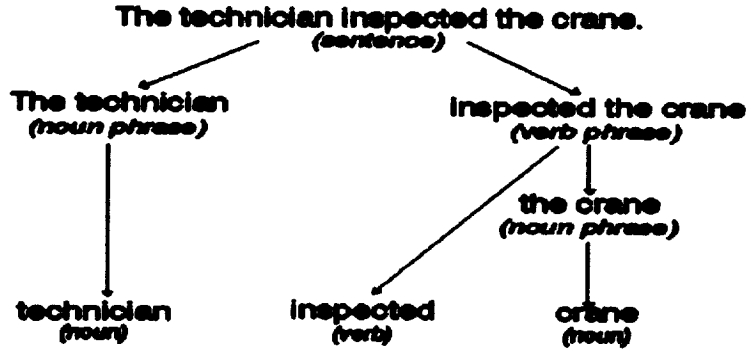

Figure 1. Sentence Parsing

Confidence factors or weights are normally used to measure the confidence level in a rule-based system. These factors combine to usefully measure uncertainty. In the CDD they are used to measure the probability that a given search parameter is correct. For instance, if a search query uses an acronym that has two possible meanings, a search for both would be performed using a lower weighing value than if the acronym had only one possible meaning.

Network connection and propagation refers to the construction of multi-layer networks and how the weights of the nodes are patterned. In this search technique, the nodes of the network are words (or phrases) with the pattern of the weighing determined by a set of heuristic rules. This network then can be used to develop a set of conditions that evaluate each area of a document.

METHODOLOGY

In any highly developed field, especially a highly technical one, there are a number of words, phrases, and acronyms that have specific meanings. These can be considered a specialized knowledge base for that particular field. Developing an intelligent search system for a specialized field must utilize that knowledge base, along with more general information of the English language.

In developing this knowledge base for NASA operations, a general database of acronyms, abbreviations, and synonyms was used as a starting point. Specialized acronyms and abbreviations used in normal shuttle operations were added to this database. In addition, word frequency tables were developed to identify the most commonly used words.

The first step in processing a query is to break down the sentence structure. Initially, the sentence or sentences are separated into individual word objects. These prime words form the first level nodes of the filter, with the order of the words maintained through the links between nodes (see Figure 2). The node object includes weighing variables for the word and for the links between nodes. The weighing variables for the prime nodes are set to a benchmark reference value of 100.

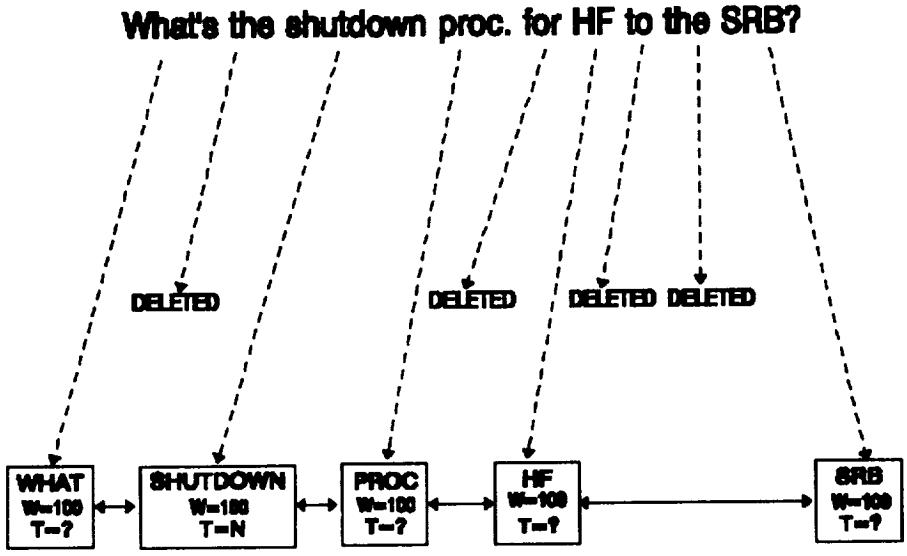

Figure 2. Generation of Prime Nodes

The knowledge base is used to expand the single level of prime nodes into a multilevel node network. Each word in the first level is referenced in the knowledge base. If a match is found, the reference values from the knowledge base become new nodes at a lower level. For example, in Figure 3, the node with the value "SRB" is matched in the acronym table with the value "Solid Rocket Booster". This value then becomes a new sub-node with links to the same nodes that "SRB" has. If more than one value is found, then more than one sub-node is created for each prime node.

In Figure 3, the parsing function identifies the prime node "PROC." as being used as a noun. A sub-node of this produced from abbreviation tables is "PROCEED," a verb. Since the word types do not match, the sub-node, "PROCEED" can be eliminated along with any synonyms produced from it. Eliminating the node this way would require assuming that the original query was structured syntactically correct. An alternate method is to reduce the weight of that node to indicate a much lower probability.

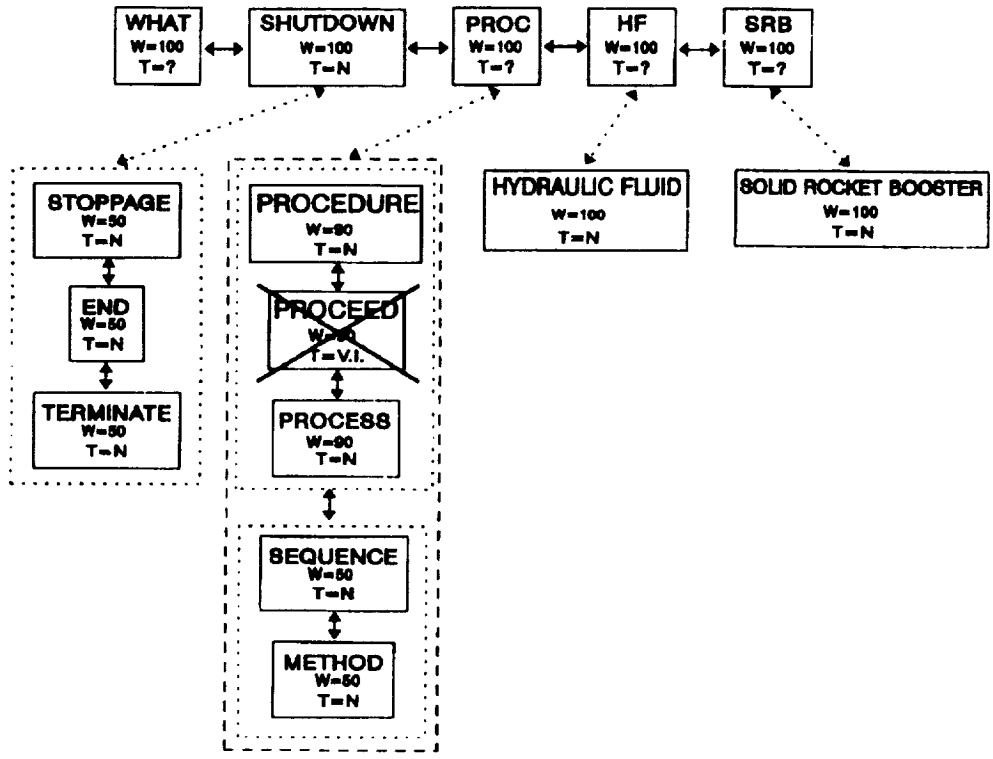

Figure 3. Generation of Sub-Nodes

After this multilevel network is created, a set of heuristic rules is used for setting the weights for each node. If the conditional part of the rule tests true, then the rule sets the weights of that sub-node, and can adjust the weight for the prime node(s) of the sub-node. Table 1 lists some of the heuristic rules used to set these values.

A Boolean search condition is then generated for every node in the net. This search condition is of the type:

if <node> exists then VALUE = VALUE + WEIGHT

The list of these conditions forms the filter function. The filter function generates a value for different areas of the document.

IMPLEMENTATION

The search technique described in the proceeding section was implemented, along with standard document handling techniques, into a system called the Centralized Document Database (CDD). This system has an extensive database of technical documents supported on a Local Area Network (LAN). The system provides a single point for accessing and searching technical documentation. The system is designed to access ASCII formatted files of various sizes and types and different physical storage locations.

#	Condition	Change
1	Word is a prime. Word is unique (not in any table of knowledge base).	Value = 400
2	Word is a prime. Word is not in frequency table.	Value = 200
3	Word frequency level > 10	Value = 0
4	Word frequency level > 5	Value = Value/2
5	Word (or phrase) is a sub-node. Phrase is only possible acronym meaning.	Value = Value of prime node
6	Word (or phrase) is a sub-node. Phrase is one of several possible acronym meanings.	Value = (Value of prime node) / (number of acronym meanings - 1)
7	Word is a sub-node. Word is a synonym.	Value = (value of prime node) / (number of synonyms)
8	Word (or phrase) is a sub-node. Phrase is only possible abbreviation meaning.	Value = Value of prime node
9	Word (or phrase) is a sub-node. Phrase is one of several possible abbreviation meanings.	Value = (Value of prime node) / (number of abbreviation meanings - 1)

Table 1
Weighing Conditions

A basic menu system is used to call up and display all of the available document files (see Figure 4). It uses a number of filters for common word processors and mainframe printer formats. These are simple filters designed to mask the command codes used by the different application programs that produced the document. The end result is that a document in almost any format, (e.g. Word Perfect, Displaywrite, or mainframe redirected printer output), can be displayed (somewhat distorted) and used by the system.

The program provides immediate access to any part of a document through the use of special pointer files. The CDD program uses these pointer files to speed-up direct location access. These pointer files can be used for pages, sections, record numbers, or any string value. When a document is selected, the software will check for all available pointer files and add the options to the option menu. These pointer files are created outside of the CDD to fulfill specific needs within the SR&QA community. The CDD requires the pointer files to be in a particular format and location, but any programming language can be used to create them.

The CDD has the capability to perform several different types and levels of data searches. The simplest type is a basic Boolean key-word search. This type of search is a useful and fairly common type of search that can locate a specific string using standard AND/OR logic. The program provides an improvement to this type of search by expanding the Boolean logic to include any acronyms and abbreviations of the search strings from its built-in database.

The program has a fully operational version of the intelligence searching technique explained previously (Figure 5). The initial query is broken down into its related components (words and phrases). Network nodes are established and expanded through the methods described previously. A set of heuristic rules are used to assign weights for the nodes in the new levels.

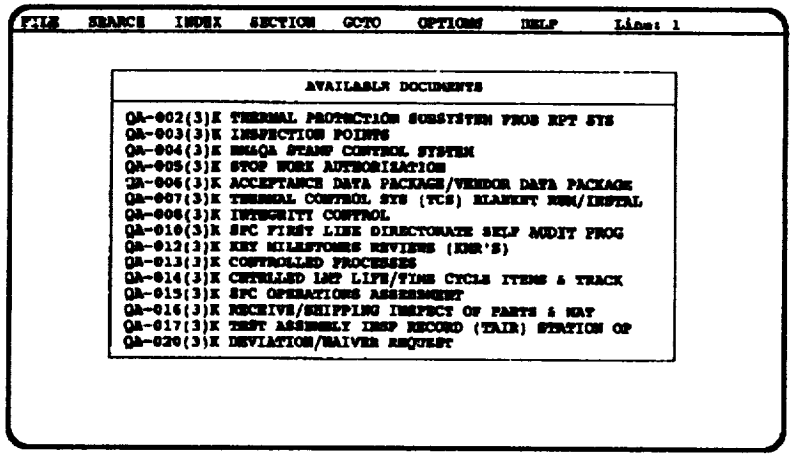

Figure 4. Document Selection in The CDD

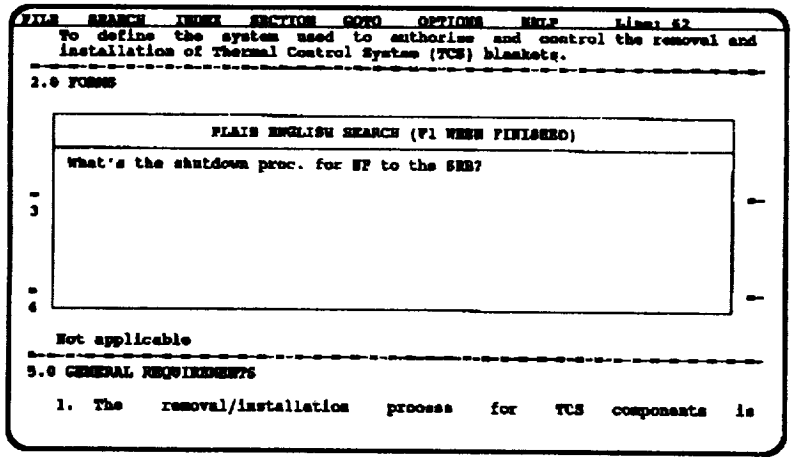

Figure 5. Intelligent Search Query Screen

This network can identify and rank key areas of the document that are likely to contain the information requested. The program does this by filtering the document through the network. This relational network returns a weight value for the section of the document that is currently passing through the network. The user is appraised of the search status as the document is being processed (Figure 6). A list of pointers, to the sections of the document that had the highest values, is the final result of the filtering. The software will immediately display the area of the document that had the highest weighing (Figure 7). If the user does not find the needed information, the software will move to the next highest weighted area of the document.

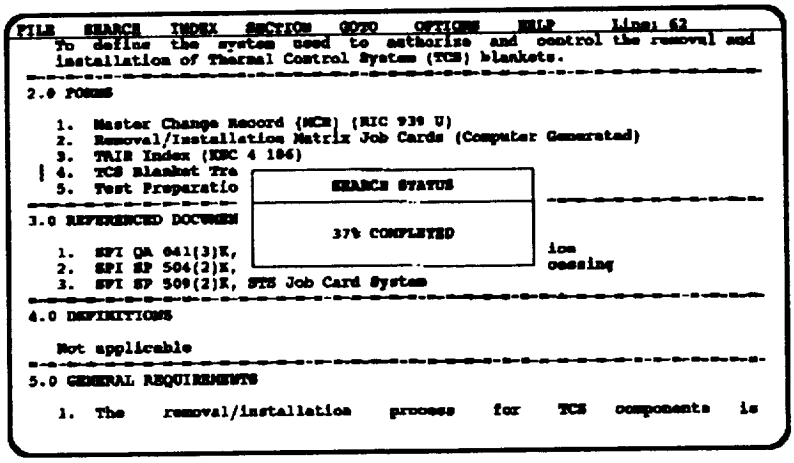

Figure 6. Search Status Screen

The results of any search are displayed on the terminal and can be exported to other applications. The program has the option to copy part or all of a document to a file or printer. The data is copied in standard ASCII format that can be imported into most word processor and database applications. The data also can be printed to any network printer.

Modifying the program to work in other fields (legal, medical, and business) would require the creation of a specialized database of words, acronyms, and abbreviations used in that field. In most cases this database already exists in the reference documentation used in that field.

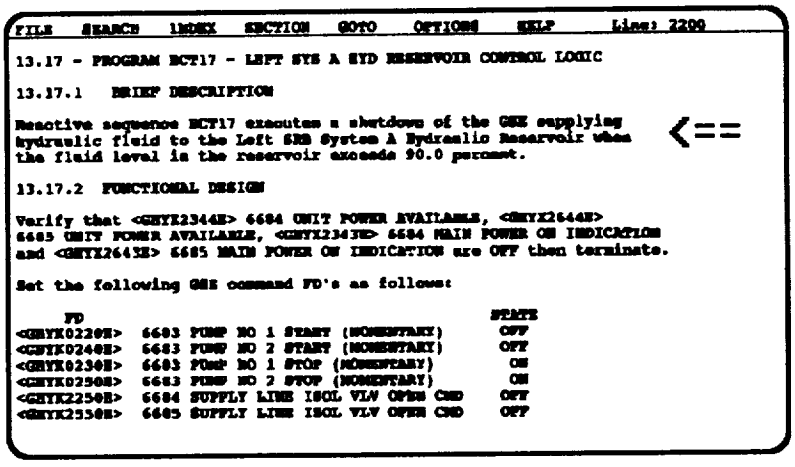

Figure 7. Results of Search

REFERENCES

[1] Luger, G. and Stubblefield, W. : Artificial Intelligence and the Design of Expert Systems, Benjamin/Cummings Publishing Company, Inc. 1989.

[2] Minsky, M. : Semantic Information Processing, The Massachusetts Institute of Technology Press, 1968.

[3] Sombe, L. : Reasoning Under Incomplete Information in Artificial Intelligence, John Wiley & Sons, Inc. 1990.

[4] Whittington, R.P. : Database Systems Engineering, Oxford University Press, 1988.