Raising the IQ in Full-Text Searching via Intelligent Querying

Robert Kero, Lucian Russell, Craig Swietlik
Argonne National Laboratory

Charles Morgan
DOE/Environment, Safety & Health (EH-5)

Charlynn Clayton, Kelly Dunlap
DOE/Office of Scientific and Technical Information

Abstract

Current Information Retrieval (IR) technologies allow for efficient access to relevant information, provided that user selected query terms coincide with the specific linguistical choices made by the authors whose works constitute the text-base. Therefore, the challenge is to enhance the limited searching capability of *state-of-the-practice* IR. This can be done either with augmented clients that overcome current server searching deficiencies, or with added capabilities that can augment searching algorithms on the servers. The technology being investigated is that of deductive databases, with a set of new techniques called cooperative answering. This technology utilizes semantic networks to allow for navigation between possible query search term alternatives. The augmented search terms are passed to an IR engine and the results can be compared. The project utilizes the OSTI Environment, Safety and Health Thesaurus to populate the domain specific semantic network and the text base of ES&H related documents from the Facility Profile Information Management System as the domain specific search space.

Introduction

The Department of Energy, Office of Special Projects and Argonne National Laboratory (ANL), along with the Department of Energy, Office of Scientific and Technical Information (OSTI) have designed and implemented the Environment, Safety and Health Facility Profile Information Management System (FPIMS), which is a personal computer based information retrieval system. This system is the central repository for Department of Energy (DOE) oversight documents consisting of Baseline Compliance Assessments, Action Plans, follow-on Progress Assessments, site specific information (e.g., maps, fact sheets, profiles), etc.

The system itself is highly modular, consisting of two major components. The first component is the graphical user interface which displays a map of the United States with icons representing field offices and associated DOE sites that have been assessed. The documents available for each site are displayed in an intuitive bookcase, bookshelf and book level progression of screens. The second component is a robust, commercial information retrieval engine used to provide powerful searching capabilities over the many user-selectable sets of documents housed within the system.

The information retrieval engine, (Personal Library Software's PC LAN-based Personal Librarian), found within the original FPIMS facilitates searches on collections of documents by providing standard Boolean logic (i.e., AND, OR, NOT) as well as natural language style searching capabilities. These capabilities are further supplemented with the ability to dynamically formulate concept-based searches. These capabilities were provided initially in a stand-alone system, and then were migrated to a Local Area Network (LAN) [MORG93].Advancing technologies on the Internet such as the World Wide Web and the emergence of an advanced commercial information retrieval product, Web-based Personal Librarian, have created new cost effective opportunities for disseminating important DOE information, as well as providing the tools necessary to support the migration of the original FPIMS to a Web-based FPIMS. This opportunity is currently being addressed.

With the opportunity, however, comes a challenge. As the number of documents in FPIMS grows and the number of potential users increases it becomes ever more a challenge to ensure effective use of the corpus of information. With this in mind Argonne National Labs has embarked on a project to raise the IQ capability of the FPIMS interface.

Motivation/Incentive

A major challenge is to provide all user classes a natural and easy means of using the FPIMS. The primary motivations are to improve specifically EH's, and broadly both DOE's and the public's accessibility to the data and to consolidate the information management needs of EH. This is consonant with EH's proactive management approach. It is also an initiative as called for in DOE's Mission Statement and the administration's National Information Infrastructure (NII). Furthermore, the proposed use of the SGML

standards, as a neutral format in which to save all EH documents is in full accord with the DOE mandate and requirement. Also, this approach allows for the exploitation of emerging advanced public domain technology, such as (Wide Area Information Servers (WAIS), World Wide Web (Web) servers and Mosaic clients), to fully leverage prior investments in FPIMS.

One technical challenge in meeting FPIMS dissemination and usage goals is communications. The migration of the FPIMS to the Internet through the use of the Web allows DOE environment, safety and health data to be also shared with outside organizations such as local, state and federal agencies, as well as the public. Hence, remote DOE operations offices and DOE sites will be provided a more consistent means of access to the information via the Internet instead of using slow phone connections to the original system's local area network. With this technical infrastructure in place, the attention is directed to the system and its potential for improving productivity and effectiveness.

From the perspective of EH, one of the purposes of FPIMS was to assist the DOE in planning future activities which might require access to ES&H documents. The idea was that any future activity of the DOE involving ES&H matters could be *a less expensive and higher quality effort* if plans could be made on the basis of already completed surveys and assessments. One example: a simplification of interactions with the field. Headquarters personnel could review all on-file records of interactions with the field offices regarding any specific area they are interested in. With this information in hand, activities involving headquarters and field offices might be conducted *by telephone instead of needing in-person visits*. In the same vein, new activities could *eliminate duplicative effort* by building upon already known and completed activities. Another clear example is the advantage of having a corpus of material that can be read by new employees, whether federal or contractor. FPIMS' documents, and the efficient means of locating specific areas of interest, can make these personnel much more effective in a shorter time frame.

The key to realizing this promise of increased efficiency is integrating the use of FPIMS into all relevant activities. In addition to the necessary conditions for this to take place (e.g. communications and software), there is one overriding condition: the government employee must be assured that the *time* spent interfacing with the system is *cost effective*. This means that the system, must provide a high degree of assurance that the necessary information will be found *without* the need to review large numbers of irrelevant documents. If this is so, the employee will continue to gain trust in the FPIMS knowing that its use is an efficient investment of their time. Any factors which mitigate this trust must be identified and eliminated.

One historical factor that mitigates the usefulness of any Information Retrieval system, is the difficulty in formulating search strategies which will obtain the *right resultant set* of documents. The Federal employee who must investigate ES&H matters is a professional in a related discipline, not necessarily a specialist in document searching. Therefore the exposure to FPIMS is likely to be intermittent and casual. Such exposure

puts a burden on the system that it contain aids for quickly navigating within the morass of detail that any such system must necessarily contain. To lift this burden, it appeared that the introduction of Cooperative Answering technology might be of assistance, and so experiments were conducted to test this hypothesis.

Advancing Efficient Full-Text Searching through Cooperative Query Construction

Locating specific full-text information throughout DOE increases in difficulty as the information continues to grow in size and complexity. Hypothetically speaking, a particular technology (which on average retrieves 7 irrelevant documents for every 3 relevant ones), is rendered unusable when document collections grow and precision to recall ratios remains static (30 %) . This information explosion problem therefore makes it unreasonable to expect that users possess enough knowledge about existing information and its organization to accurately compose useful queries consistently - *the burden must be shared by the information system they are interacting with!*

Current Information Retrieval (IR) technologies do very little to facilitate this problem since they only allow efficient access to "relevant information", provided user selected query terms coincide with the specific linguistical choices made by the authors whose works constitute the corpus being searched. For casual or intermittent users of the system this cannot be guaranteed: they cannot be expected to know the context in which documents were indexed, categorized, or the terminology of the specialists who generated the documents. Cooperative Answering strategies assist these users by providing access to the conceptual framework of the specialists who created the documents.

Therefore by accessing the context used by the authors of the documents in FPIMS, the IQ software can guide the user during query construction. The benefit of this approach is that it permits earlier detection of poorly composed queries - e.g., queries that (1) are to general, or (2) contain misconceptions, or (3) are not linguistically equivalent to that of the corpus. The system can provide feedback and suggestions of alternative words that will make the queries more effective.

The proof-of-concept IQ system currently available utilizes the Department of Energy's Environment, Safety and Health thesaurus/dictionary (which includes taxonomies - broader/narrower, synonymous, and meronymous terms and phrases), to populate a domain specific semantic network. This semantic network is used to navigate between possible query search term alternatives. The domain specific HTML-based corpus is a subset of the FPIMS' documents.

The System Overview

The problem, as shown in Figure 1, is to navigate the maze of documents so as to locate the most likely required document subset for which to peruse - this subset could be found distributed across multiple IR engines. An integral part of the puzzle is the IQ-based

FPIMS user interface. By using a visual interface metaphor, of a map and bookcases located at various sites on the map, it is possible to provide user-selectable control over the total FPIMS document set. Therefore, the overall query space can be initially constrained by site, by document type or by combinations of both. The intelligence of the system, is found in the Semantic Net, a logical database of terms having one or more relationships to one another. This set of relationships contains at its core the terms developed by the OSTI team that created the DOE ES&H Thesaurus. The IQ software uses this semantic net to supply the user with possible new terms which might make their searches more effective. The following Figures are taken from the proof-of-concept IQ system; they are provided to allow for a better understanding of the process.

The example takes a compound search, "arm or joint" as the starting point as shown in Figure 2. In the original test there were over 100 records (portions of documents) which satisfied the original query - too much to peruse quickly. (This situation is typical in large document collections.) Note that the screen identifies the server as a Wide Area Information Server (WAIS) node. This is not necessary for IQ searching to be of value - any type of search engine could be substituted.

The next screen (Figure 3) shows how to make use of the Cooperative Answering capability - made narrower. The OSTI Thesaurus was searched for which words in the query could be changed, made either more general or more restricted or related in meaning, to refine the query's semantics. In the example shown both words could be "relaxed" - replaced by other related terms.

The system then displays (Figure 4) the choices which can be made. The word "joint" was relaxed through a narrower interpretation, and the 7 types of specific joints were provided for possible selection and subsequent query term replacement.

The selection of "shoulder" was made (Figure 5) and the resulting query (Figure 6) revealed that 21 possible documents were found, a savings of 500% in document perusal time!

This simple example shows that the incorporation of the IQ system can greatly improve the querying process.

Benefits and Value for DOE

The concept of using a discipline-specific thesaurus to assist a DOE user in formulating and enhancing queries, effectively leverages the experience of the subject matter experts in various disciplines for the refinement of the users query. The OSTI expertise invested in the thesaurus can now be used as an integral part of IQ-based information systems.

The IQ technology makes it possible for DOE and public users to access highly specific technical information within large document collections in an efficient and highly cost effective manner. Coupled with advances in Internet technologies, the dissemination of

the FPIMS' environment, safety and health oversight information is helping to achieve the DOE's openness initiatives. This technology enables DOE to work better and cost less.

Current and Future Research Directions

Although Cooperative Answering strategies have been successfully incorporated with deductive and relational databases in the past, our research will now investigate its incorporation with information retrieval. The overall goal of the project is to provide a more efficient user class independent (e.g., domain experts through the general public) query construction technology.

Acknowledgments

The authors wish to acknowledge the contributions of the organizations and staff who have been involved in the development of the systems. Department of Energy: Mark Gilbertson. Argonne National Laboratory: Michelle Bernard, Terry Gaasterland, Robert Haddad, Julia Lee, Wes Maciorowski, Michele McCusker-Whiting. Personal Librarian Software: Deborah Senay, Paul Campbell. The development of these FPIMS-based systems and experiments represent a close collaboration of Government, Department of Energy, National Laboratories and the commercial sector.

Work supported by the U.S. Department of Energy, Assistant Secretary for Environment, Safety and Health, Office of Special Projects under Contract W-31-109-Eng-38.

References

[AND93]

Marc Andreesan: NCSA Mosaic Technical Summary, NCSA Technical Document, NCSA Springfield, Champaign Il, May 8, 1993.

[CLA93]

Charlynn Clayton, et. al: INFOTECH'93, "WAIS EVALUATION PROJECT REPORT", Office of Scientific and Technical Information U.S. Department of Energy, DOE Technical Information Meeting, Oak Ridge, Tennessee, Oct. 20-22, 1993.

[GAAS92]

Theresa Gaasterland, "Generating Cooperative Answers In Deductive Databases", University of Maryland, August 1992.

[KERO94]

Kero, R., C. Swietlik, Migrating the Facility Profile Information Management

System into the World Wide Web, DOE 11th Office Information Technology Conference, Chicago, Aug. 23-25, 1994.

[MORG93]

Morgan, C., C. Swietlik, B. Kissinger, Environmental, Safety and Health Facility Profile Information Management System (FPIMS), INFOTECH'93, Office of Scientific and Technical Information U.S. Department of Energy, DOE Technical Information Meeting, Oak Ridge, Tennessee, Oct. 20-22, 1993.

[PIT94]

James E. Pitkow & Mimi Recker: First International Conference on the World Wide Web, "RESULTS FROM THE FIRST WORLD WIDE WEB USER SURVEY", College of Computing, Georgia Institute of Technology, Atlanta, GA, May 25-27, 1994.

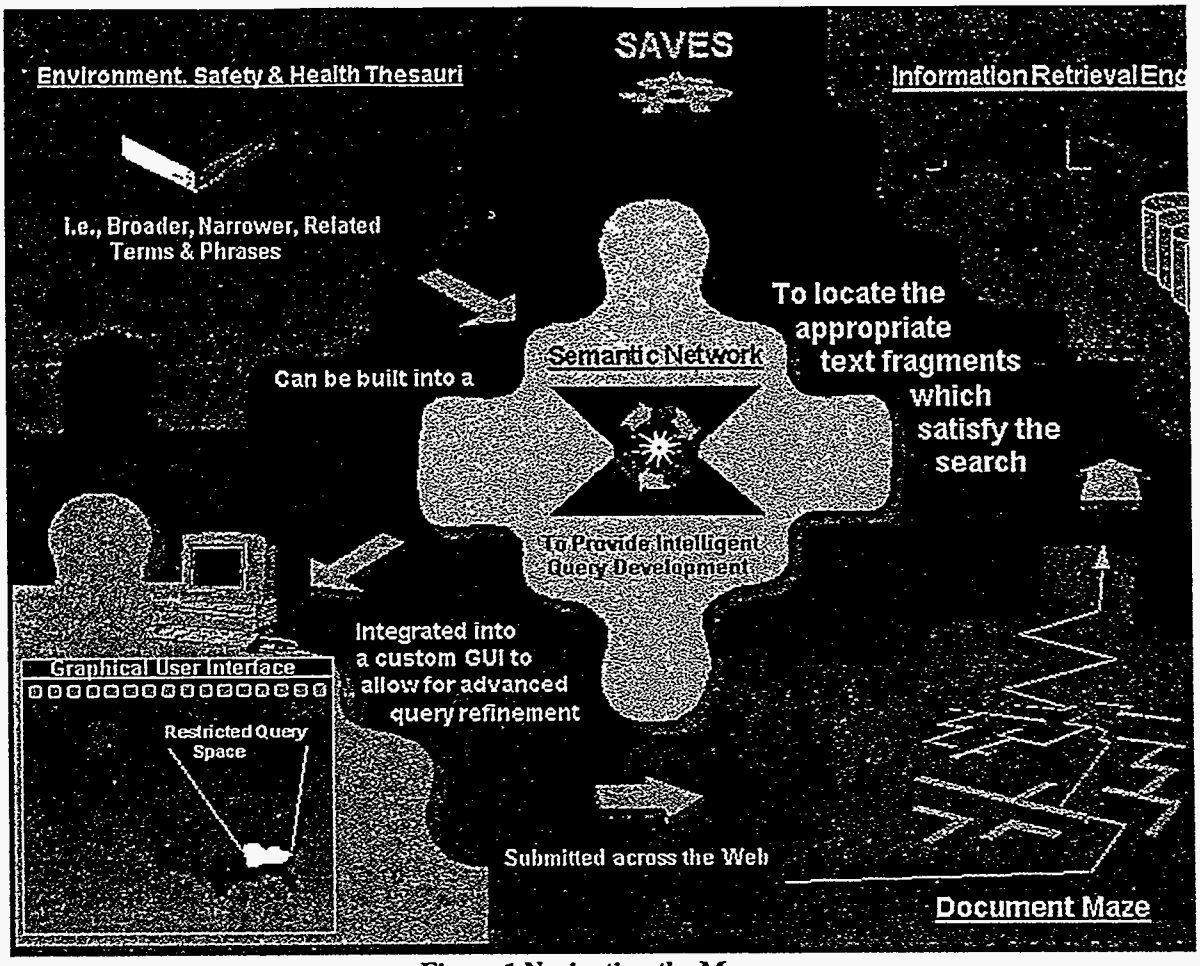
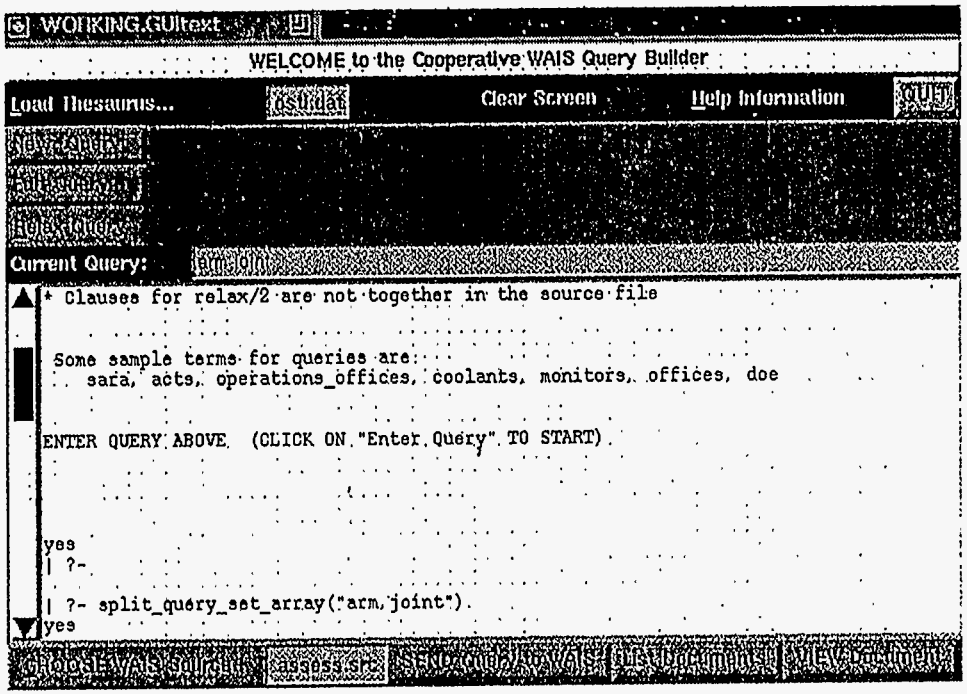
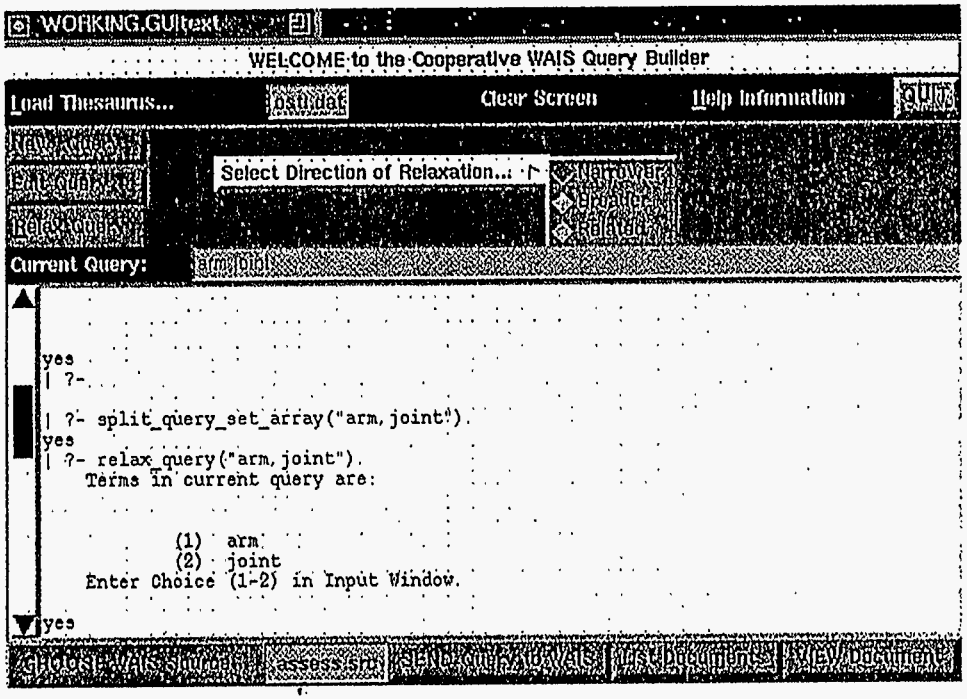
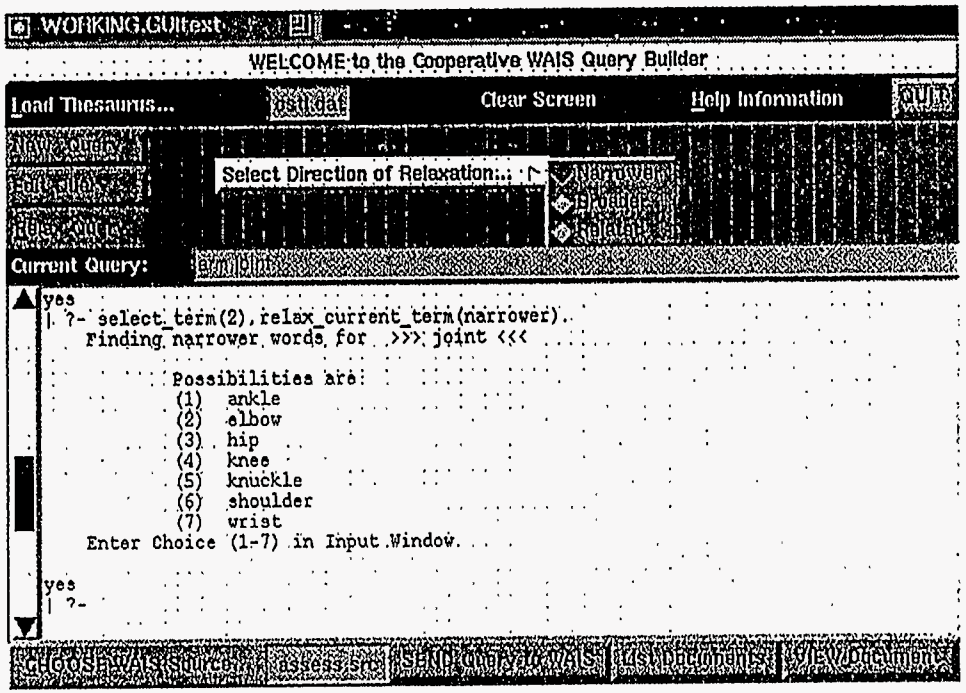
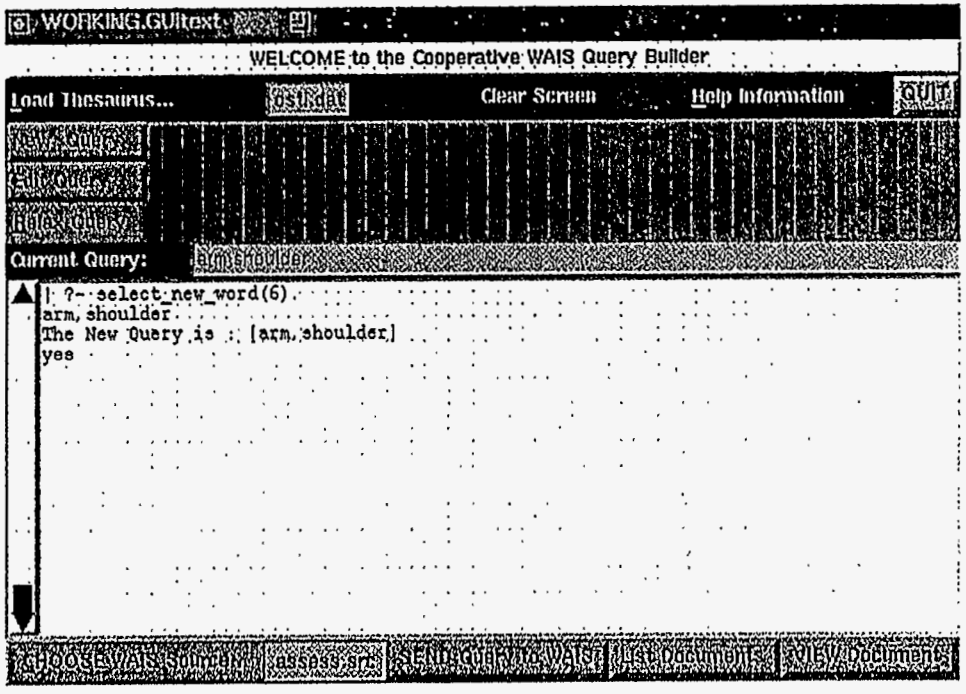


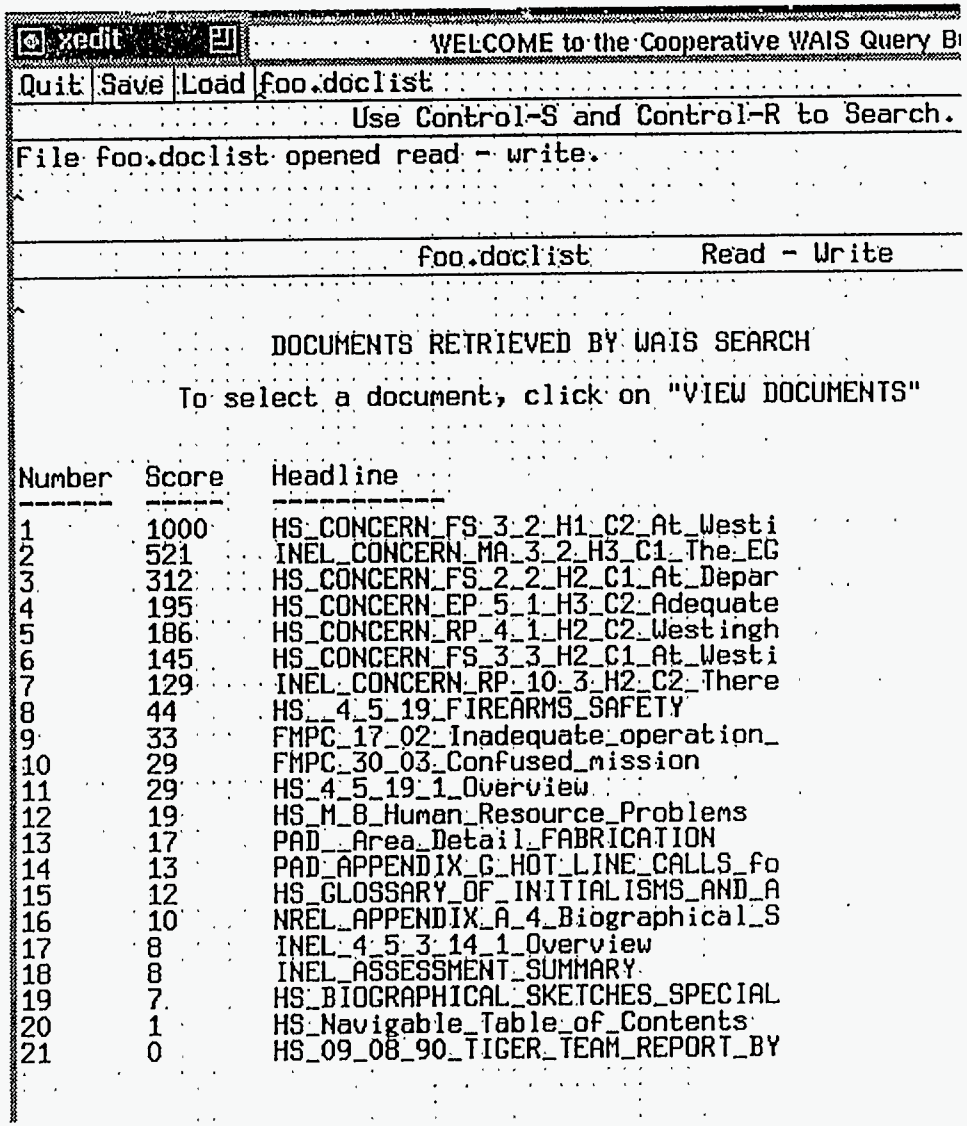
Figure 1 Navigating the Maze


WORKING.GUItext
WELCOME to the Cooperative WAIS Query Builder
Load Thesaurus...
Clear Screen
Help Information
QUIT
Current Query:
* Clauses for relax/2 are not together in the source file
Some sample terms for queries are:
sara, acts, operations_offices, coolants, monitors, offices, doe
ENTER QUERY ABOVE (CLICK ON "Enter Query" TO START).
yes
| ?-
| ?- split_query_set_array("arm,joint").
yes


WORKING.GUItext
WELCOME to the Cooperative WAIS Query Builder
Load Thesaurus...
Clear Screen
Help Information
Select Direction of Relaxation...
Current Query:
arm joint
yes
| ?-
| ?- split_query_set_array("arm,joint").
yes
| ?- relax_query("arm,joint").
Terms in current query are:
(1) arm
(2) joint
Enter Choice (1-2) in Input Window.
yes


WORKING.GUItest
WELCOME to the Cooperative WAIS Query Builder
Load Thesaurus...
Clear Screen
Help Information
QUIT
Select Direction of Relaxation:
Current Query:
yes
| ?- select_term(2),relax_current_term(narrower).
Finding narrower words for >>> joint <<<
Possibilities are:
(1) ankle
(2) elbow
(3) hip
(4) knee
(5) knuckle
(6) shoulder
(7) wrist
Enter Choice (1-7) in Input Window.
yes
| ?-


WORKING.GUItext
WELCOME to the Cooperative WAIS Query Builder
Load Thesaurus...
Clear Screen
Help Information
QUIT
New Query
Edit Query
Refine Query
Current Query:
arm shoulder
| ?- select_new_word(6).
arm, shoulder
The New Query is : [arm, shoulder]
yes
CHOOSE WAIS Sources
SEND QUERY to WAIS
List Documents
VIEW Document


xedit
WELCOME to the Cooperative WAIS Query B
Quit
Save
Load
foo.doclist
Use Control-S and Control-R to Search.
File foo.doclist opened read - write.
foo.doclist Read - Write
DOCUMENTS RETRIEVED BY WAIS SEARCH
To select a document, click on "VIEW DOCUMENTS"
Number Score Headline
1 1000 HS_CONCERN_FS_3_2_H1_C2_At_Westi
2 521 INEL_CONCERN_MA_3_2_H3_C1_The_EG
3 312 HS_CONCERN_FS_2_2_H2_C1_At_Depar
4 195 HS_CONCERN_EP_5_1_H3_C2_Adequate
5 186 HS_CONCERN_RP_4_1_H2_C2_Westingh
6 145 HS_CONCERN_FS_3_3_H2_C1_At_Westi
7 129 INEL_CONCERN_RP_10_3_H2_C2_There
8 44 HS__4_5_19_FIREARMS_SAFETY
9 33 FMPC_17_02_Inadequate_operation_
10 29 FMPC_30_03_Confused_mission
11 29 HS_4_5_19_1_Overview
12 19 HS_M_8_Human_Resource_Problems
13 17 PAD__Area_Detail_FABRICATION
14 13 PAD_APPENDIX_G_HOT_LINE_CALLS_fo
15 12 HS_GLOSSARY_OF_INITIALISMS_AND_A
16 10 NREL_APPENDIX_A_4_Biographical_S
17 8 INEL_4_5_3_14_1_Overview
18 8 INEL_ASSESSMENT_SUMMARY
19 7 HS_BIOGRAPHICAL_SKETCHES_SPECIAL
20 1 HS_Navigable_Table_of_Contents
21 0 HS_09_08_90_TIGER_TEAM_REPORT_BY