Manipulating Full-text Scientific Databases: A Logic-based Semantico-pragmatic Approach*

R. MARSHALL†
Computer Science Department, Loyola College, Baltimore, MD 21210, USA

The syntactic, semantic, and pragmatic interconnections inherent in a portion of NASA's full-text scientific database, ENVIRONET, have been captured in the form of facts and rules using relevance and modal logic. Such a logic formulation of the contents of the database permits content-based retrieval of textual information using arbitrary search granularities. It also allows users to manipulate data contained in tables and graphs which occur as part of the natural text. The system takes into account user factors in deciding on the quantity and quality of information retrieved and how the information is displayed. The organization of the facts and rules is quite independent of the order in which the original text appears.

Received April 1988, revised February 1991.

1. INTRODUCTION

1.1. Purpose of the Study

The research study reported herein focuses on the theoretical and practical implications of employing non-standard-logic-based expert systems components to aid in the retrieval of information from full-text scientific database systems. For purposes of experimentation, the ENVIRONET scientific database has been used. ENVIRONET is a bibliographic, numeric and full-text information database at the NASA Goddard Space Flight Center in Greenbelt, Maryland.

The expected utility of incorporating an expert system component in ENVIRONET the retrieval process is predicated on the component's ability to provide, in some intelligent fashion and on a *user-customised* basis, scientific and technical users of ENVIRONET only *relevant* yet contextually *complete*, *coherent* and *consistent* information on space vehicles and the dynamic and static environment which obtains both inside and outside these vehicles. This information is essential to users for the proper design, manufacture, test, and operation of scientific payloads meant to be carried on the space vehicles. Effective isolation of a scientists experiment module from other scientists experiment modules requires that scientists be made aware of potential sources of data corruption and that they be provided with information from areas of expertise other than their own which might have an impact on their experiments. Scientists using ENVIRONET will need to be guided by the system in some intelligent fashion to obtain the necessary information without their having to laboriously navigate through a major portion of the database.

1.2. A Brief Description of ENVIRONET

ENVIRONET is a multi-disciplinary full-text database which contains information on the following topics: vibration and acoustics, electromagnetic interference, thermal and humidity effects, loads and low frequency dynamics, microbial and toxic contaminants, molecular contamination, surface interactions, natural environment, particulate environment, and orbiter motion. Interspersed in the full-text are tables of numeric data, charts and figures. ENVIRONET is accessible to the scientific community and the public, in the US and abroad, via phone lines. Typical users of the database include NASA and non-NASA scientists designing experiments for the space station and shuttle projects, engineering contractors and management personnel. ENVIRONET is resident on a MicroVAX II which is a node on the NASA Space Physics Analysis Network (SPAN).

```
B -> BROWSE text retrieval system
N -> Bulletin board notices
D -> Download specific chapter
M -> Mail system
       mail us your comments on the system
F -> Function calculation system
       natural environment models
G -> Graphics
       download high resolution graphs
L -> Logoff
       End ENVIRONET session
  Enter appropriate letter, followed by <RETURN>:
```

Figure 1. ENVIRONET main menu.

Information for the database is collected and maintained by three NASA technical panels – the experimenters panel, the natural and induced environment panel, and the information management panel. These panels are responsible for identifying users requirements and specifying needed environment data, making preliminary assessments of the reliability and traceability of such data, and indicating areas of use to experimenters.

ENVIRONET is basically a menu-driven system containing a set of hierarchically arranged menus. The root (i.e. main) menu has options which include a text retrieval subsystem named BROWSE, a table of contents reflecting the diverse scientific topics on which information is available, a bulletin board, downloading of full-text and simple ASCII text graphs, a function calculation system which can be invoked to exercise the various atmospheric models which have been developed to NASA and elsewhere, access to an electronic mail utility, and downloading of high resolution graphics. The

* The work reported here was conducted under research grant NAG-5893 from the NASA-Goddard Space Flight Center, Goddard, MD.

† Now at: Department of Computer Science, Florida Institute of Technology, Melbourne, FL 32901, USA.

CHAPTER	Section–Page
Introduction	1–1
Thermal and humidity	2–1
Vibration and acoustics	3–1
Electromagnetic interference	4–1
Loads and low frequency dynamics	5–1
Microbial and toxic contaminants	6–1
Molecular contamination	7–1
Natural environment	8–1
Orbiter motion	9–1
Particulate environment	10–1
Surface interactions	11–1
Definitions and acronyms	12–1

(P)age/Section (T)able of contents (Y)memory (E)xit (R)eturn (I)ndex (H)elp
Enter Page or section number, and ⟨RETURN⟩

Figure 2(a). BROWSE main menu.

(A)head page (T)OC (P)age/section (R)eturn (M)ain topics (B)ack page (I)ndex (S)ection (F)ind word (E)xit

Figure 2(b). BROWSE inner menu.

If the user selects option F on the BROWSE menu, the following prompt appears:
Enter string (Q to quit) and ⟨RETURN⟩
If the string is not found, the prompt displayed is:
String not found. Search next chapter? (Y or N)
If the string is found, the first page where it occurs is displayed along with the menu shown below:
(N)ext occurrence, (P)revious occurrence,
(Q)uit search

Figure 2(c). BROWSE search string menu.

main menu of ENVIRONET is shown in Fig. 1. The various menus associated with the BROWSE text-retrieval sybsystem are shown in Figs 2(*a*), 2(*b*), and 2(*c*). Users selecting the BROWSE subsystem on the main menu are presented with yet another menu. Among the various options available to the user in this menu are the following: a listing of the major sections of the database, a table of contents for a particular section, an index of keywords, scrolling facilities, and searching for the occurrence of an arbitrarily specified string in a given section or chapter. The menus which are brought up when the function calculation system (natural environment models) is invoked are shown in Figs 3(*a*) and 3(*b*).

1.3. Motivation and Previous Work

In our investigation of incorporating expert systems components into ENVIRONET, we have focussed on those expert systems aspects which facilitate the identification of basic syntactic, semantic and pragmatic interconnections between different segments of the database. In our research, we have adopted the view that ENVIRONET is several different full-text scientific databases under one umbrella. Since we are dealing with data representing many domains of expertise (electromagnetics, acoustics, thermal effects, etc.) and since some of these domains of knowledge do overlap to varying degrees, it is essential to capture the degrees and types of interconnectedness between these domains. This perspective of ENVIRONET has the following implications: which database to search depends on who the user is, how much information the user needs and how much the user already has, in what areas of the database the user is interested, and whether the user requires any analysis or synthesis of the retrieved data.

One of the stated goals of the designers of ENVIRONET is the following: ENVIRONET should be capable of providing the most appropriate and up-to-date information for the proper design of experiments to be conducted on the space shuttle by the technical and scientific community, the targeted users of ENVIRONET. Very often experiments are designed based on incomplete information possessed by the designers of the experiments; consequently this results in the duplication of effort, data already available is ignored, data that need to be collected in the course of an experiment do not get collected, and, more importantly, data designed to be collected by one experiment on the space vehicle may be unintentionally corrupted by the presence of other experiments in the vehicle. Scientists need to be made aware of potential sources of data corruption so that they can effectively isolate their experiments. They will have to be provided with information from areas of expertise other than their own which might have an impact on their experiments. To make such information readily available, scientists using ENVIRONET will need to be guided by the system in some intelligent fashion. This will obviate any need on a scientist's part to laboriously browse through a significant portion of a very large database to extract the relevant pieces of information.

Since our research is concerned with issues involving hypertext and multimedia systems, a brief review of related work in these areas which has been reported in the information processing and computer science literature is presented in the next paragraph.

Hypertext refers to non-sequential literature arranged for associative reading while a multimedia system refers to a system which facilitates computer processing of text, graphics, image and voice data. A survey of work done in the area of hypertext is to be found in Ref. 1. The feasibility of applying a wide variety of artificial intelligence techniques to aid in the information retrieval process is an active area of research at various institutions. For example, an expert systems approach to document retrieval is reported in Ref. 14. Experimental systems such as MINOS,[2] MUSE,[7] and CODER[4] employ artificial intelligence techniques to aid in the preparation, presentation, and retrieval of multimedia documents and in the formulation of user queries. An approach to handling both multimedia data and hypertext structures is discussed in Ref. 20. A heuristic approach to conceptual information extraction and retrieval from natural language input is presented in Ref. 15.

2. METHODOLOGY

2.1. The Nature of Scientific Text

Full-text scientific databases differ from conventional databases and non-scientific databases in two important ways.

Firstly, the syntactic and semantic elements found in the text are subject-area specific and consequently there is only a limited amount of overlap with those pieces of

Models

1. MSIS 86 neutral thermosphere model
2. Magnetic field model
3. International reference ionosphere
4. Quit

Enter appropriate response followed by ⟨RETURN⟩

Figure 3(a). Function calculation system menu.

MSIS model

Input ranges

Day number	1 to 365	Altitude (km)	85 to 1000
Local solar time (hrs)	0 to 24	Magnetic index AP	0 to 400
Average F10.7 flux	65 to 300	Current F10.7 flux	65 to 300
Geodetic longitude (deg)		−180 to 180	
Geodetic latitude (deg)		−90 to 90	

Input parameters		Output values	
1. Day of year	44	H (number/cm3)	1.41E+07
2. Altitude	100	N (number/cm3)	6.54E+05
3. Latitude	40	HE (number/cm3)	9.64E+07
4. Longitude	−75	O (number/cm3)	4.03E+11
5. Local time	12	N2 (number/cm3)	7.85E+12
6. Average F 10.7	100	02 (number/cm3)	2.02E+12
7. Current F10.7	200	AR (number/cm3)	1.19E+11
8. Magnetic index AP	300	Total (gm/cm3)	4.91E−10
		TN (INF) (deg K)	1300.8
		TN (deg K)	194.0

Do you want to (R)un the model with the current values, change some (1–8) or (A)ll the values, or (Q)uit?

Figure 3(b). MODEL menu.

text which are also subject-area specific but whose subjects of discussion are different.

Secondly, scientific text is, in general, rich in pragmatic content. The word 'pragmatic' is used primarily in a linguist's sense of the work. In speech, the pragmatic content of an utterance is readily available through 'supra-segmental' features such as stress, intonation, and accent. In full-text manipulation, extraction of pragmatic content is based on two sources of information – (*a*) the text itself, and (*b*) the particular query or query sequence input by the user. The pragmatic aspects relevant to scientific text include items such as references to works by other authors, figures, tabular data, charts, and mathematical expressions. Mathematical expressions are of particular importance since a scientifically-inclined user may wish to manipulate such expressions in their symbolic form or perform parameter substitutions to obtain numeric values and graphs.

Consider the following typical portion of scientific text:

> Summaries of the forces and torques produced by Skylab operations are given in Fig. 7, Table 12 and Table 13. Using a shuttle mass of 85,000 kg (see Ref. 10), this gives accelerations between 0 and 5 milli-g.

The typical user might be, for example, interested in (*a*) looking at figure 7; (*b*) examining tables 12 and 13 (such tables and figures need not occur in the immediate vicinity of the displayed text) and possibly manipulating them, and (*c*) obtaining additional details on the article cited in reference 10. Syntactic elements such as 'table', 'figure', and 'ref.' in scientific text provide important pragmatic clues which exist as part of the textual material (i.e., endocentric aspects). As for pragmatic clues which are exocentric in nature (i.e., clues which are found outside the text), the intensionality of the user needs to be taken into consideration.

2.2. A 'Possible Worlds' Model

To capture some of the intensional aspects which are present in any user-machine dialog, we have resorted to a 'possible worlds' model utilising modal and relevance logics[11] in our reformulation of the scientific text and in the query handling process. A brief explanation of the 'posible worlds' model is presented in the ensuing paragraph.

If we assume two scientific disciplines, say C1 and C2, as representing two worlds, where two propositions, say p1 and p2, are true in C1 and C2 respectively, the validity of a third proposition p3 which is the conjunct of p1 and p2 is dependent on the 'connectedness' of C1 and C2. The 'connectedness' of C1 and C2 may be predicated on syntactic, semantic or pragmatic grounds.

Given p1 (C1, X) = true; p2 (C2, X) = true;
R = C1; C2 = 0
Then (p1 and p2) (C1, C2, X) = true in R

In the above, X is a subset of the connectivity set which is defined to be {syntactic link, semantic link, pragmatic link}. The semantic, syntactic, and pragmatic connectivity

of more than two worlds with respect to a set of propositions is also defined using the same approach.

The referential connectivity of a piece of text or a collection of texts may be established on syntactic, semantic, and pragmatic grounds. Most text usually contains anaphoric references, polysemous constructs and ellipsis which provide clues to syntactic and semantic connectivity while syntactic entities such as 'figure', 'table', and 'reference(?)' are good indicators of types of pragmatic connectivity. Thus text is a set of syntactic elements whose 'meaning' is conditioned by a set of semantic or concept elements and a set of pragmatic elements.

In our work, we have adopted the view that a piece of text, say T1, is representable as a triple (Syn, Sem, Prag) where Syn is the set of syntactic elements of the text {s1, s2, .. sm}, Sem is the set of semantic elements {c1, c2, .. cn}, and Prag is the set of pragmatic elements {p1,p2, .. pk}. Likewise, a query is representable as the triple {Syn′, Sem′, Prag′}. However, in the case of a query, the elements comprising the three sets are defined quite differently. The set Syn′ contains the base syntactic element which is user-supplied, and elements which are 'syntactic collocates' of the base term. These latter elements are supplied by the system. Syntactic collocates are restricted to synonyms, paraphrases, and variant spellings including acronyms. The set Sem′ contains the base concept word or phrase which is implied by the user-supplied syntactic term, and the concept word or phrases 'semantic collocates' which are restricted to paraphrases of the base concept word or phrase, and variants. The Prag′ set contains user-supplied data such as user category, search granularity, and topic of interest as well as 'pragmatic collocates' for these items.

2.3. Logic Net

Our 'possible worlds' model of information-retrieval in full-text scientific databases is an integrated model in that it takes into account not only a wide range of user perspectives (including user categories, capabilities and intensionalities) but also the basic functional aspects of scientific text. The reader will no doubt observe that our formulation of the text, the user and the text-user interface is a comprehensive network model where the nodes in the network are linked, either physically or logically, by three different types of links – syntactic, semantic, and pragmatic links. These typed links are, in general, many-valued. In order to traverse the network, extensive use of inheritance and transitivity rules have been used.

Inheritance rules are of the form:

If q(X, Y) and p(X, s), then p(Y, s)

In the above, 'p' is a property and 'q' is a relation. The rule may be paraphrased to read: if relation 'q' holds between entity 'X' and entity 'Y', and entity 'X' has the value 's', then entity 'Y' has the same value 's'.

Transitivity rules are of the form:

If q(X, Y) and q(Y, Z), then q(X, Z)

The above rule may be paraphrased to read: if relation 'q' holds between entity 'X' and entity 'Y', and the same relation holds between entity 'Y' and entity 'Z', then relation 'q' also holds between entity 'X' and entity 'Z'.

In establishing the various physical as well as virtual (i.e., inferred) semantic and pragmatic interconnections between textual portions at the inter-sentential, inter-paragraph, and inter-chapter levels, the notion of relevance and concepts pertaining to relevance and modal logic,[12,18] and text grammars, [3,5,8,10,13,19] have been used in recasting the contents of the different chapters into the facts and rules associated with each chapter.

We have not used any statistical or probability measures in formulating the rules. For a statistical semantic approach to the performance of key-word information systems, see for example Ref. 6. Our network approach, while akin to hypertext, is quite different from other published work on hypertext in that the formulation and traversal of the network is grounded rigorously on formal logic; it is neither done on an *ad hoc* basis nor does it incorporate any heuristics. Very often, relations among objects in a hypertext structure are insufficiently defined; semantic information and structural information are not clearly distinguished. Our network is not deficient in this regard since different types of node linkages are maintained in the net. The query formulation and presentation aspects of the system include both menus and natural language input; the systems which have been reported in the literature employ one or the other, rarely both.

To give a concrete example of how the notion of 'relevance' and non-standard logics such as relevance logic and modal logic are used in designing the expert system, a simple example is presented below.

If a user were to specify the search string 'ram', conventional information-retrieval systems would attempt to retrieve text containing one or more occurrences of 'ram' regardless of whether (1) the user is an aerodynamics expert or a computer engineer, and (2) the specified string occurred in a portion of text dealing with computer memories or aeronautics. On the other hand, if the user describes himself as an aerodynamics specialist prior to exercising the database, under most situations, the expert system would only retrieve text containing 'ram' which has aerodynamics as the underlying theme. As for the applicability of modal logic, a computer scientist performing a text search using the string 'rom' would necessarily be also interested in text containing references to related (in an immediate sense) terms such as 'rom, prom, prom', etc. and be possibly interested in terms such as 'ram' which are related to memory though not in an immediate sense. Modal logic has also been found to be useful in the proper codification of semantic collocates. For example, 'if phrase X occurs, phrase Y must also occur in the same context', and 'if phrase U occurs, phrase V is likely to occur as well'. Such a codification allows for some latitude in handling both syntactically and semantically ill-formed input. This ill-formedness pertains not only to the original document or text but to user queries as well.

The connectivity of text T_i with respect to text T_j is given by

$$T_i\ T_j = (S_i\ S_j, C_i\ C_j, P_i\ P_j)$$

The connectivity of text T_i to two or more pieces of text is defined likewise. It should be noted that connection between text T_i and T_j will not, in general, be the same as the connection between text T_j and text T_i since the

necessary P	= not possible not P
contingent P	= possible P and possible not P
noncontingent P	= not possible P or not possible not P
consistent (P, Q)	= possible (P and Q)
inconsistent (P, Q)	= not possible (P and Q)
(strict) implication (P $\succ$ Q)	= not possible (P and not Q)
(strict) equivalence (P $\leftrightarrow$ Q)	= not possible not (P = Q)

[= represents material equivalence]

Figure 4(a). Interdefinability of modal operators.

P	$\Diamond$ P	$\Box$ P	∇ P	$\triangle$ P
T	T	I	I	I
F	I	F	I	I

Figure 4(b). Monadic modal operators truth table.

P	Q	P $\bigcirc$ Q	P $\oplus$ Q	P $\rightarrow$ Q	P $\leftrightarrow$ Q
T	T	T	F	I	I
T	F	I	I	F	F
F	T	I	I	I	F
F	F	I	I	I	I

Note: I is not a third truth value. It becomes T or F depending on the instantiation.

Figure 4(c). Dyadic modal operators truth table.

connectivity is established on the basis of implications and not on the basis of definitions; the former posits a one-way inferencing while the latter, a two-way inferencing.

2.4. Modal Operators

In formulating our rules for the expert systems rule base, we have used the following monadic and dyadic modal operators:

monadic: necessary ($\Box$), possibility ($\Diamond$), contingency (∇), non-contingency ($\triangle$),
dyadic: consistency ($\bigcirc$), inconsistency ($\oplus$), strict implication ($\rightarrow$), strict equivalence ($\leftrightarrow$).

The above eight modal operators are inter-definable; we have chosen the possibility operator as the basic one and used it to define the other seven operators. The operators' definitions are shown in Fig. 4(*a*).

The truth status of a proposition involving these modal operators is explained in the ensuing paragraphs using a possible worlds paradigm.

Let P be any proposition, A be the actual world (i.e., the scientific area or domain under consideration), and B represent the remaining worlds or domains. If P is true in A and in B, then P is necessarily true. Likewise, if P is false in A and in B, then P is necessarily false. If P is true in A but it is false in at least one element of set B, then P is possibly true. In addition, if P is necessarily true, it is also possibly true. On the other hand, if P is false in A but is true in at least one other world contained in B, then P is possibly false. Also, if P is necessarily false, it is also possibly false.

Irrespective of the truth status of P, if P has the ascription of 'necessary', then P is an non-contingent proposition. Otherwise it is a contingent proposition.

The remaining four modal operators, which are dyadic in nature, may be explained thus. Let P and Q be propositions. P and Q are inconsistent only if there is no world, actual or otherwise, in which both are true. P and Q are consistent with each other if and only if there is some world in which both are ture. As for implication, P implies Q if and only if in *each* of the worlds where P is true, Q is true as well. P and Q are equivalent if and only if there is no world in which they differ in truth value, i.e., if P is true/false in some world, Q is true/false in that same world.

The truth tables for the modal operators are shown in Figs 4(*b*) and 4(*c*).

3. SYSTEMS DESIGN

In the following two sections, we focus on the design aspects of the expert system. In Section 3.1 we address the issue of how text retrieval is reformulated and retrieved in our system and present several examples. In Section 3.2 we give a brief description of the system architecture, existing as well as proposed.

3.1. Text Reformulation and Text Retrieval Examples

NASA scientists with expertise in the subject areas represented by the data in ENVIRONET have been consulted in building the various expert system components for the following obvious reasons: (1) The logic rules to be formulated in setting up the rule-base of the expert system are dependent on the particular subject areas under consideration. (2) Rules for establishing physical and virtual links between subject areas can be set up only in consultation with experts in the subject areas which are being linked. (3) Ascertaining the nature and extent of usage of domain-specific information by the scientists is necessary if the user-machine interface is to be properly designed.

A fairly detailed analysis of ENVIRONET users has been carried out to establish the different categories of users based on frequency of usage of the ENVIRONET database, scientific and technical background of the user, the computer skills of the users, the access mechanisms employed by the users, the portions of the database the user accesses, and the purported reasons for accessing the database. This information has also been utilized in reformulating the text through the specification of suitable facts and rules.

In reformulating the text as a set of facts and rules using first-order predicate logic, a content scanning approach has been followed. A content scanner has the basic capability to scan text and store only the significant aspects of the text. The utility of this approach lies in the fact that user queries are mostly content-based or content oriented and that most pieces of text have a smaller amount of significant semantic content than the actual text would seem to imply. A considerable amount of text space is taken up by style and assorted grammatical

1. input_skill, main_menu, help, explain, end_search
2. accelerometer, disturbances, thrusters
3. reaction_control_system, reaction_control, control_systems
4. orbiter_maneuvering_system, oms, prcs, vrcs.

Figure 5(a). Simple facts.

1. thruster(symbol), section_9_1_1(symbol)
2. disturbance_is(symbol), disturbance_periodic(integer)
3. table(integer), figure(integer), ref(integer), tablell(integer, symbol, real, real).

Figure 5(b). Collocation predicates.

1. disturbance_is(gravity_gradient):
   section_9_2_2(gravity_gradient),
   section_9_2_2(table9a),
   section_9_2_2(table9b),
   section_9_2_2(figure_3).

2. disturbance_is(venting): section_9_2_6(intro),
   venting(scheduled), venting(contingency), venting(failure).

3. start: input_skill, explain, main_menu.

4. activate_skill(3): skill(5), not(skill(3)),
   assertz(skill(3)).

5. formulall(A1, A2, T1, T2):
   skill(5), A2 = 1.5*A1, T2 = 0.1*T1.

Figure 5(c). Sample rule base.

requirements and publishers' manuscript submission requirements. This approach for scientific full-text databases may be defended on several grounds. One, scientific data and scientific articles generally contain a lot of redundant information as, for example, in articles where authors discuss previous work and results. It would be pointless for ENVIRONET to display entire articles on a specific topic; instead, the system should have the capability of only presenting the significant content of articles relating to a specific topic and pointing the user to the sources if the user is interested in delving into the articles. It should be noted here that the 'significant content' of an article is quite different from the abstract of an article. For example, the significant content of a scientific research article includes a statement of the problem, experimental procedures followed, equipment used in the experiment, results obtained, figures and tables. The extraction of conceptual information from documents is an active area of research in natural language processing and text understanding systems.[16]

Unlike a keyword-based information retrieval system where the basic search units are keywords, in our system the search units are predicates and propositions which form part of the fact base of the expert system. The predicates and propositions are based on significant syntactic, semantic, and pragmatic units which occur in the text and in the retrieval environment. The latter is posited on text structure, user categorisation, and other extra-textual and extra-linguistic features.

The collocation units, which in *part* are akin to the thesauri entries in a lexicon-based information retrieval system, occur as clauses in the rule base of the expert system; some collocation units occur in the fact base as well.

The syntactic, semantic, and pragmatic connectivities are captured throughout the specification of implicational rules. The clauses which occur in a rule employ predicates involving free variables as well as propositions, the former typifying physical connectivity in text while the latter, virtual or inferential connectivity. Some of these rules also involve the modal operators of necessity, possibility, contingency and consistency. Predicates involving modal operators have been found to be extremely useful in capturing the pragmatic aspects of text and the intensional aspects of the typical user of ENVIRONET. Examples of propositions, predicates, and rules employed in the system are shown in Figs 5(*a*), 5(*b*) and 5(*c*).

In Fig. 5(*a*), four sets of propositions are shown. The first set deals with the user-interface and is representative of the exocentric pragmatic aspect of the system. Individual facts in this set may be viewed as 'procedure calls' which result in the invocation of rules containing these facts. The second set is representative of syntactic and semantic entities which occur in the text. The third set typifies syntactic collocates. Any user query which refers to, for example, the phrase 'reactor control system' has

9.2.5. Crew motion

Extensive measurements of the forces provided by crew motion were made on Skylab (See Ref. 11). For the most part, people have divided these forces by the mass of the shuttle orbiter to calculate the acceleration. These calculations are corroborated by a few measurements performed aboard Spacelab (Refs. 1, 4, 5). Summaries of the forces and torques produced by Skylab operations are given in Fig.

9.2.5. Crew motion

Extensive measurements of the forces provided by crew motion were made on Skylab (See Ref. 11). For the most part, people have divided these forces by the mass of the shuttle orbiter to calculate the acceleration. These calculations are corroborated by a few measurements performed aboard Spacelab (Refs. 1, 4, 5). Summaries of the forces and torques produced by Skylab operations are given in Fig.

Figure 6(a). Text retrieval under BROWSE.

(Note: BROWSE highlights every line containing the search string *** [shown here in italics])

Sample query

```
Goal: disturbance (motion)
```

Text retrieved

From Section 9.2.5 Crew motion
Extensive measurements of the forces provided by crew motion were made on Skylab (see Ref. 11). For the most part, people have divided these forces by the mass of the shuttle orbiter to calculate the acceleration. These calculations are corroborated by a few measurements performed aboard Spacelab (Refs. 1, 4, 5). Summaries of the forces and torques produced by Skylab operations are given in Fig. 7, Table 12, and Table 13. Using a shuttle mass of 85,000 kg (See Ref. 10), this gives accelerations between 0 and 5 milli-g. This is in rough agreement with VFI and MSDR accelerometer results from Spacelab 1. These results show acceleration spikes of up to 10 milli-g.

Figure 6(b). Expert system text retrieval.

```
Goal: altitude (drag_torque).
```

From: Section 9.2.3. Aerodynamic Drag
Since the force has components parallel and perpendicular to the flight path, the orbiter experiences drag torques as well. These torques are given in Table 11. Note the strong altitude dependence: increasing the altitude by 50 per cent decreases the drag by an order of magnitude.

Table 11. Orbiter aerodynamic torques

Altitude (nmi)	Body axis	Max. Torque (ft-lb)	Min. Torque (ft-lb)
100	Roll	7.2	−7.2
	Pitch	16.4	−16.4
	Yaw	7.6	−7.6
150	Roll	0.7	−0.7
	Pitch	1.6	−1.6
	Yaw	0.8	−0.8

Figure 7(a). Compound query.

```
Goal: tablell(Altitude, roll,
       Max_Torque, Min_Torque).
```

Altitude	Max_Torque	Min_Torque
100	7.2	−7.2
150	0.7	−0.7

```
Goal: tablell(150, Axis, Max_Torque, Min_Torque).
```

Axis	Max_Torque	Min_Torque
roll	0.7	−0.7
pitch	1.6	−1.6
yaw	0.8	−0.8

Figure 7(b). 'Computational' queries.

```
Goal: altitude(solar_activity).
```

From: Section 9.2.3 Aerodynamic Drag
Figure 5. Calculated atmospheric drag on the orbiter at 300 km altitude over the period of one orbit (90 min).
The extremes indicate the variations in drag due to solar activity. The peaks are caused by the two tidal bulges in the atmosphere. (Modified from Reference 4).

the necessary connotation of 'control systems'. A reference to 'control systems' in a user query has no such implications of necessity as far as the phrase 'reactor control systems' is concerned; it, however, has connotations of mere possibility. It must be noted that any proposition which is necessarily true (false) is also possibly true (false); the reverse does not hold. The last set of propositions is representative of semantic and pragmatic collocates. Oms, prcs, and vrcs are acronyms which stand for orbiter manoeuvring system, primary reaction control system and vernier reaction control system, respectively, and these are entities which occur in the discussion on thrusters in the ENVIRONET text dealing with the Shuttle's altitude control system.

Examples of three sets of predicates are shown in Fig. 5(*b*). In the first set, the first predicate allows a user to retrieve text dealing with specific kinds of thrusters while the second permits the user to retrieve text dealing with, for example, thrusters in a constrained portion of text, namely Section 9.1.1. The second set of predicates permits user queries on different types of disturbances such as periodic and transient disturbances as well as causal aspects of disturbances such as crew motion, rack vibration, and the like. Since there are a greater number of potential sources of disturbances than there are different types of disturbances, the system presents the user with a menu of potential disturbance sources once the user has shown interest in a specific type of disturbance. This is exemplified by the query 'disturbance_periodic (integer)' where the argument is the specific menu option selected by the user. The third set of predicates in Fig. 5(*b*) is a good example of the pragmatic aspects of text and of user intentions. When the user is viewing a portion of text containing references to tables, figures and charts in the text and citations to articles which definitely are not part of the text, these predicates permit the user to bring up instantly the figures, tables, etc. In the case of tabular data, the user is also given some capability to selectively manipulate the data. Such manipulation is restricted to retrieving horizontal and vertical subsets of the tables and performing extrapolation and interpolation (refer to Fig. 7(*b*) as well). It should be pointed out that these operations are akin to selection, projection and function computation operations in typical relational database systems. This computational aspect of our system is quite distinctive in that most traditional full-text information retrieval systems

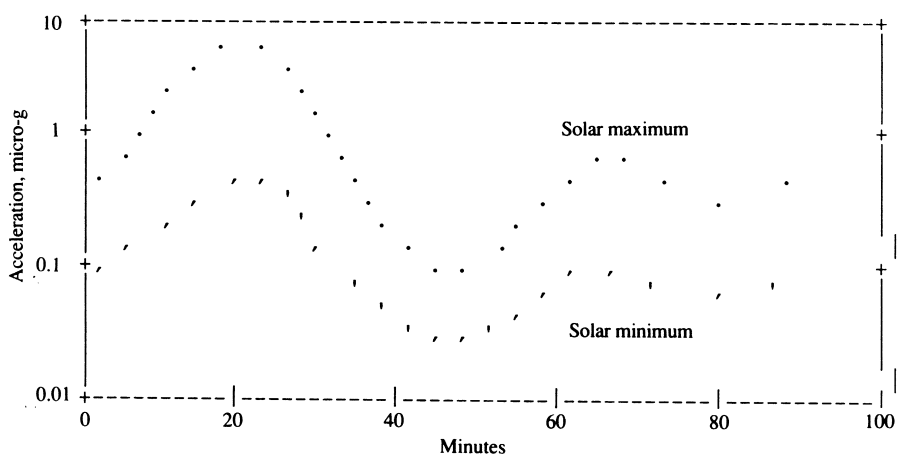

Figure 7(c). Text and graph retrieval query.

do not permit such on-the-fly numeric-type computations.

Fig. 5(*c*) shows five examples of implicational rules used in the system. Text dealing with gravity gradient type disturbances is retrieved from a specific chapter/section of ENVIRONET when the first rule is invoked; the text will also contain context related tables and figures, the context being disturbances caused as a result of a gravity gradient. The second rule results in the retrieval of any text dealing with the topic 'venting'. The text retrieved contains an introductory discussion of venting, followed by expanded discussions of major types of venting and venting failures in space vehicles. The third rule deals with system initialisation. The fact 'input_skill' obtains (from the user) information on the user, the fact 'explain' is a concise explanation of how expert system based information retrieval occurs, and the fact 'main_menu' presents the main menu of ENVIRONET. The fourth is an example of a user categorisation rule which permits the system to dynamically compute the 'skill' level of the user in order to decide how much text is retrieved, how the text is displayed and what operations on the text are permissible by that user. The last rule is an example of the computational aspects within full-text. The particular example presented permits computations of variables A2 and T1 for user-specified values of A1 and T1 only if the system-computed value of the user-skill level is 5.

A portion of text as it actually exists in ENVIRONET and the same piece of text reformulated as facts and rules are shown in Figs 6(*a*) and 6(*b*) respectively.

The information which is retrieved by the expert system for three different types of queries is shown in Figs 7(*a*), 7(*b*) and 7(*c*). In Fig. 7(*a*), text and tabular data are retrieved when the query 'altitude (drag torque)' is input. The query may be paraphrased to read 'is there any connection, other than syntactic, between drag torque and altitude?'. It should be remarked here that a simple string search mechanism would merely look for one or more occurrences of the syntactic elements 'altitude' and 'drag torque'. 'Computational' queries on information (i.e. Table 11) retrieved by the previous query is shown in Fig. 7(*b*). A vertical subset of Table 11 is presented when the user specifies an altitude value of 150; likewise, a horizontal subset of Table 11 is presented when the user specifies the body axis to be 'roll'. A query on the connection between altitude and solar activity results in the retrieval of graphical and text data as shown in Figure 7(*c*). If a user wishes to obtain details on reference 4 cited in the text which is displayed, the user would obtain this information by formulating the query 'ref (4)'.

3.2. System Architecture

The architecture of the system as presently configured is shown in Fig. 8. An integrated version of the system, shown in Fig. 9, is currently being developed. A proposal to employ a distributed architecture version of the system is also under study. A multimedia message system on a distributed architecture has been reported in Ref. 17.

In order to adequately support the 'possible worlds' model of information retrieval by taking advantage of any existing, yet natural, partitionings of the full-text

Figure 8. Original system.

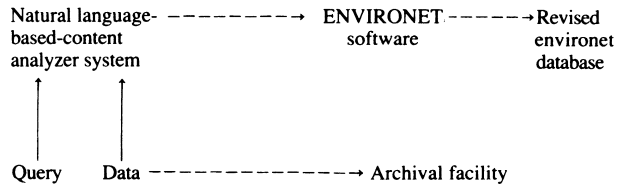

Figure 9. Revised architecture.

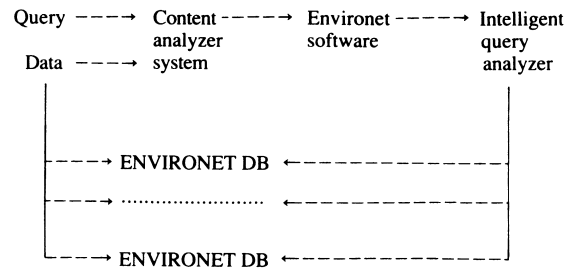

Figure 10. MBDS architecture.

database, the system architecture would have to be fairly sophisticated. To this end, we have proposed a modified version of the Multi-Backend Database System (MDBS).[9] The system, shown in Fig. 10, would permit parallel processing of user queries in an intelligent fashion. The natural language based content analyser system would be used to process not only user queries but the data entered into the ENVIRONET database as well. Since the database is now conveniently partitioned along scientific-topic lines, user queries involving data retrieval from more than one database would result in parallel searches of the relevant databases, such parallel searches being instantiated by the intelligent query analyser.

5. CONCLUSIONS

A prototype expert systems interface to a small portion of the ENVIRONET database has been built. In specific, the contents of Chapter 4 'Electromagnetic Interference' and Chapter 9 'Orbiter Motion' of the existing ENVIRONET database were reformulated in terms of facts and rules using two different versions of the logic programming language PROLOG, namely C-Prolog on Loyola College's VAX 11/785 and Turbo-Prolog on a MicroVAX II at the Goddard Space Flight Center.

In analysing the textual portion of the database, special attention has been paid to anaphora, ellipsis, and polysemous constructs for suitable reformulation in a facts and rules framework. Syntactic elements with significant pragmatic content such as 'figure', 'table',

and 'ref.' have also been extensively used in formulating the rules for text extraction.

The numeric portion of the database containing tabular data, charts and the like has also been represented using facts and rules in two different ways – one representation being the actual data, and the second (if possible), a mathematical expression or formulation. The latter allows users to perform computations in an information retrieval mode, e.g., simple interpolations and extrapolations.

Unlike conventional full-text retrieval systems, our system allows queries requiring numeric computations (e.g., functional queries), and the order in which the contents of the chapters in the database are stored is quite independent of the order in which the original text appears. The original text is in archival form and is not available unless explicitly requested by the user.

The analysis of the messages left by the users indicates that not enough information was being collected for generating a fairly comprehensive user profile. This information, though quite inadequate, has been incorporated into the experts systems program by redefining the rules and facts of the database to enable search and retrieval processes to be conducted on a customised user basis. In order to obtain a detailed user profile information on the following activities of the user of ENVIRONET is being monitored (such monitoring is being done with the explicit permission of the user owing to privacy statutes): (*a*) express reasons, other than curiosity, for accessing the database; (*b*) which chapters, sections, and paragraphs within sections are accessed; (*c*) whether any downloading of information occurs at the user end; (*d*) duration of the session with the database; (*e*) search strings specifications – in particular strings stored in the index section of the database, arbitrary strings (with and without wildcards), lengths of each user-specified search, and the number of strings specified during a given session with the database; (*f*) the menu choices made by the user at the inter-section and intra-section levels, and finally (*g*) whether the user abandons a particular search and the reasons for such abandonment (such reasons may include lack of interest in the document retrieved, wrong document being retrieved as a result of user error or system error, frustrations in using the menus and the search software, and inability, whether actual or user-perceived, of the system to satisfy the user query).

A comparative performance evaluation, from a user perspective as well as from a space/time trade-off perspective, of the BROWSE and Prolog-based systems is currently being conducted and we hope to publish the results soon.

REFERENCES

1. J. Conklin, Hypertext – An introduction and survey. *IEEE Computer* **20** (9) (1987).
2. S. Christodoulakis, *et al.*, Multimedia document presentation, information extraction and document formation in MINOS: A model and a system. *ACM Transactions on Office Information Systems* **4** (4) (1986).
3. S. C. Dik, *Functional Grammar*. Academic Press (1981).
4. E. D. Fox, Development of the CODER System: A testbed for artificial intelligence methods in information retrieval *Information Processing and Management* **23** (3) (1987).
5. W. Frey, *et al.*, Automatic construction of a knowledge base by analyzing text in natural language. *International Joint Conference on Artificial Intelligence* (1983).
6. G. W. Furnas, *et al.*, Statistical semantics: Analysis of the potential performance of key-word information systems. *The Bell Systems Technical Journal* **62** (6) (1983).
7. S. Gibbs and D. Tsichritzis, Document presentation and query formulation in MUSE. *Proceedings of the ACM Conference on Research and Development in Information Retrieval, Pisa, Italy* (1986).
8. R. Granville, Controlling lexical substitution in computer text generation. *COLING '84, Proceedings of the 10th International Conference on Computational Linguistics* (1984).
9. D. K. Hsiao, (ed.), *Advanced Database Machine Architecture*. Prentice Hall, Englewood Cliffs, NJ (1983).
10. E. Marsh, *et al.*, A production rule system for message summarisation. *Proceedings of the American Association of Artificial Intelligence* (1984).
11. R. Marshall, Inferencing in modal and relevance logic systems: A representation scheme with applications. *Proceedings of the Conference on Computers and Information Sciences*, pp. 735–740. The Johns Hopkins University (1987).
12. R. Marshall, A logic programming approach to full-text database manipulation. *Proceedings of the Conference on User-Oriented Content-Based Text and Image Handling*, MIT, Cambridge, Massachusetts (1988).
13. R. Marshall, Text understanding and discourse analysis: A functional grammar approach *The Computer Journal* (In the press).
14. S. Pollitt, CANSEARCH, An expert system approach to document retrieval. *Information Processing and Management* **23** (2) (1987).
15. L. Rau, Knowledge organization and access in a conceptual information system. *Information Processing and Management*. Special Issue on Artificial Intelligence for Information retrieval **23** (4) (1987).
16. L. Rau, Conceptual information extraction and retrieval using natural language input. *Proceedings of the Conference on User-Oriented Content-Based Text and Image handling*, MIT, Cambridge, Massachusetts (1988).
17. R. H. Thomas, A multimedia message system on a distributed architecture. *IEEE Computer* **18** (12) (1985).
18. A. Urquhart, Semantics for relevant logics. *The Journal of Symbolic Logic* **37** (1972).
19. T. A. Van Dijk, Connectives in text Grammar and text logic, edited T. A. van Dijk and J. Petofi. *Grammars and Description*, Walter de Gruyter, Amsterdam (1974).
20. N. Yankelovich and N. Meyrowitz, Reading and Writing the Electronic Book. *IEEE Computer* **18**, (10) (1985).