23

1. This document has been approved for public release and sale; its distribution is unlimited.

Department of Psychology
University of Washington • Seattle

11

AFOSR -68-0948

THE REACTIVE LIBRARY

EARL HUNT AND J. R. QUINLAN

DEC 1 1967
67-1-03

This research was supported by the
U.S. Airforce Office of Scientific
Research under Grant No. 1311-67

Abstract: Reactive Library

An information retrieval program has been written to aid in coordinating literature searches by separate individuals. The program utilizes rudimentary techniques originally developed in question answering and theorem proving programs in order to determine whether a fact input to it is relevant to previously stored facts. The list of relevant facts, together with a citation to the article reporting them, is then printed out. The program is currently being used by a group of physiological psychologists on a field trial basis.

The Reactive Library

Earl Hunt and J. R. Quinlan
The University of Washington

No one can keep up with literature. As with most complaints, this one hides several problems, only some of which can be helped by computers. We shall describe an information retrieval system called the "Reactive Library" (RL), which is designed to aid, but not replace, a person in the job of fitting new experimental findings into the previously available literature.

The RL is a tool for increasing communications within a small group of scientists, say somewhere between a dozen or a hundred workers dealing with closely related research topics. The basic assumption of the RL is that the different individuals in such a group will share a common language, but have somewhat disparate reading interests. Thus a biochemist and a physiological psychologist might both be studying the physical basis of memory, and be quite capable of communicating with each other, but still be scanning different journals. In such a situation person A may come upon a fact which has interesting connotations if considered in the light of a fact which person B has noted in his journal, but the connotation may not be at all apparent if either fact is considered in isolation. The problem is to establish an exchange of information between A and B. Traditionally this is done by social devices, such as discussions and seminars. While such meetings have their place, they also have well known disadvantages. In addition to personal contact, it would be useful if B could routinely be informed of the relevance facts in articles A was reading.

This informal description contains some implied assumptions which ought to be made explicit. The problem is one of relating the facts contained in articles to each other, and not of grouping relevant articles. While it should be possible to retrieve articles which assert given facts, the physical retrieval problem is not of primary interest. Cooper's (1963) distinction between fact retrieval and document retrieval is relevant. In our particular case a given article may contain several facts, each of which should be related to its own set of stored facts, which may appear in different articles. The second assumption we have made is that the scientists do indeed share a common language. While this seems trite, its importance cannot be stressed too highly. Previous studies on the use of language in social sciences (see particularly Stone, et al., 1967) have shown that the definitions of terms upon which scientists agree very largely determines their conceptual approach to their subject. The RL, to be useful to the scientific community in general, must be able to accept the definition of a language as part of its data base, rather than having special functions built into the program itself.

Our final implicit assumption is that we are dealing with a program for research groups which, by and large, do not have their own private computing equipment. The RL should be available through the sort of general purpose, public utility type computers which one might reasonably expect to find in universities and research laboratories in the next few years.

The RL program, which meets these requirements, is a fact retrieval system which is conceptually an amalgam of the techniques used by Raphael's (1964) the Semantic information Retrieval program and the methods developed by Stone et al.,(1967) for the analysis of scientific dictionaries.[2] The program was written in the Burroughs extended Algol language and operates as a user program in the University of Washington's Burroughs B5500 remote terminal computing system. The programming job was made easier by first writing portions of the RL in SNOBOL (Farber, et al., 1964), then using a SNOBOL to ALGOL conversion system (Quinlan, 1967) to produce the final program. The entire effort, from conception to initiation of field trials required less than two man-months.

The program is now being used on a field test basis by a group of physiological psychologists who are interested in the physical substrate of memory. Most of our illustrative examples will be taken from their work.

The User's View

In the RL the user is virtually isolated from the system. He need not even know how to turn on a teletype. Each user is given a number of standard forms, on which he is asked to take notes while reading articles. An example of a completed standard form is shown in Figure I. For convenience, users may keep carbons for their personal files. Since most scientists in this field take extensive notes on experimental reports, this is a minor addition to their normal practices, although it does make note taking slightly more formal than usual.

The standard form is divided into four sections. The first section, the reference, simply names the article. The second section, labeled "Conclusions" is the key section for the RL. Each conclusion must be a single sentence stating an assertion which the user, or reviewer, thinks can be made on the basis of the article. Figure 1 contains representative samples.

Section 3 is a "Comments" section. Information in this section is kept on file, but not input to the RL.

The fourth section permits the user to specify particular types of searches of previously stored data. Discussion of this section will be postponed until after the program's operation has been described.

Users deposit completed forms with a secretary. Periodically the secretary goes to a teletype near her office and inputs the forms to the RL program. While this requires a modicum of special training, it relieves the research specialist of the need of ever learning anything new, and in addition relieves him of the frustrations of waiting for a computing system to be available. The secretary also redistributes output to the scientific personnel.

Figure 2 shows a fragment of an output page. The output is again divided into sections, one for each fact which was input. The RL lists the input fact, its reference, and all previously stored facts which the system has selected as relevant to the input, along with their references. In addition, the output sheet also states under the term "Metafact", the reason why a particular stored fact has been retrieved.

If we examine a fragment of Figure 2 more closely, the general nature of the rules which the RL uses to establish relevance will become clear. Consider the input statement

THE POST-TRIAL INJECTION OF STRYCHNINE IMPROVES DISCRIMINATION LEARNING IN TRYON-S3S.[3]

The system makes the following retrievals.

1) AGREEMENT WITH STORED FACT

THE PRE-TRIAL INJECTION OF PICROTOXIN FACILITATES DISCRIMINATION LEARNING IN WISTARS

METAFACT ON WHICH AGREE

INJECTION CONVULSANT IMPROVE DISCRIMINATION RAT

and

2) DISAGREEMENT WITH STORED FACT

INTREPERITONEAL INJECTION OF 1757-IS HINDERS AVOIDANCE LEARNING IN TRYON-S1S.

METAFACT ON WHICH DISAGREE

INJECTION CONVULSANT IMPROVE LEARNING RAT

Examination of these sentences will reveal the retrieval pattern. All retrieved facts contain a verb which is either approximately a synonym or an antonym of the verb in the input fact. If the verb is a synonym, agreement is indicated; if the verb is an antonym, disagreement is indicated. Parenthetically, if the verb is "NOT v", it is interpreted as its own antonym. In addition, we see that non-verb forms also play a role in retrieval. Each "Metafact" is the basis of agreement or disagreement. All "noun-like" terms in the metafact (which we call operands) represent either operands in the input or generalizations of an operand in the input. (Note the replacement of "strychnine" by "convulsant" in the example.) Each retrieved fact contains within its operands specializations of the generalization proposed in the metafact.

Interior of the System

The way in which the RL establishes generalizations and relevancies between facts will now be described in more detail.

Recall that the RL consists of two parts, a program and a dictionary. The dictionary contains two classes of words, operands and verbs. Insofar as the system is concerned, a sentence is an ordered set of the dictionary terms it contains. The sentence

THE POST-TRIAL APPLICATION OF ECS OFTEN DISRUPTS MEMORY

will be stored as

POST-TRIAL EC DISRUPT MEMORY[4]

assuming that "THE", "APPLICATION", "OF", "OFTEN" are not stored in the dictionary.

An input may contain any number of operand names, but at most one verb.

Each verb is a member of one class of verbs, which is further divided into "positive" and "negative" terms. Thus the "produce-destroy" class of verbs consists of the positive terms

"produce, facilitate, improve"

and the negative terms

"hinder, disrupt, destroy".

We will refer to a verb as having a sign, in addition to having a class. If the special word NOT precedes a verb in an input sentence, the sign of the verb is changed. This does not correspond exactly with English usage, since "not produce" is not equivalent to "destroy". So far this has not proven a problem.

Operands are related to each other by set inclusion. We have seen examples of this already, "strychnine" is included in the set "convulsant". It is possible to include an operand in two sets which are not included the one within the other. Electroconvulsive-shock (ECS) is included in the sets "Traumas" and "Electric-Shock" in the physiological dictionary. Somewhat to our surprise, however, we have thus far found few such terms.

Each operand is represented internally by a code which permits an immediate match between operands to determine their lowest common intersection. The coded operand name is a string of symbols, where the first symbol is the name of the "most general" class to which the operand belongs, the first and second symbol the name of the second most general class, etc. Thus if A were the code for animal, then AM might be the code for mammal, AMR the code for rat, and AMR6 and AMR7 the code for, say, Tryon S1 and Tryon S3 rats. By a simple masking operation, we can detect that the common denominator, as it were of AMR6 and AMR7 is AMR, the code for "rat." Similarly, the names of all mammals can readily be retrieved (specialization), by selecting all codes which have AM as their first two symbols. To handle the case in which a given term is a member of two sets which are not nested, as in the ECS example, the operand name is assigned two codes. It would be inconvenient if very many operands had to have multiple codes, but, as noted, so far this has not happened very often. (We are sure situations can be constructed in which this would give trouble, but the pragmatic question is how often such situations occur.)

Internally, sentences are coded by their verb class and sign, and by the set of left-hand and right-hand operands which they contain. The example sentence reduces to the equivalent of

verb = disrupt

left = EC, post-trial

right = memory

Sentences are stored under their verb-classes, using a tree form of information organization. When a fact is read as above, a search is made of the file of facts already stored whose verb is in the same class as that of the input fact. The object is to find stored facts which can be generalized or specialized to a metafact which is itself a generalization or specialization of the input fact. The processes of generalization are

1. Removal of one or more operands from either the subject or the object, or
2. Replacement of an operand by an operand which includes it.

The process of specialization is the replacement of an operand by one of its subsets. So, for instance, if

cat feline animal,

rat rodent animal,

some of the constructed facts to which the fact "black cats eat rodents" could be transformed by the above processes are

cats eat rodents

black felines eat rats

cats eat animals

black animals eat animals

If a metafact can be found to which both the input fact and a stored fact can be transformed, then the system announces an agreement or disagreement depending on whether the verbs of the matched facts agree in sign. If the word "NOT" appears in a sentence, the sign of the verb is changed to determine agreement or disagreement.

User Options--The Control Section

By use of the control section special retrievals may be ordered. The two most obvious are that a user may store a fact without receiving retrievals, or assert a fact and obtain retrievals without having the input fact stored. The first option is useful in reading in large banks of commonly accepted statements, as would be done when a RL system was initialized. The second option is useful if one wants to "see how an idea stacks up against the literature". For example, by stating the hypothesis of a proposed experiment, one would hopefully obtain a retrieval of related literature.

The"Buzzword" and "Keyword" sections permit more sophisticated control. If all the operands on both the left and right sides of a metafact are in the list of buzzwords, no retrieval will be output with that metafact. This is often useful in suppressing a large number of irrelevant retrievals. For example, POST-TRIAL is a subset of TIME-REFERENCE, a term which includes operands such as DAYS, WEEKS, SECONDS, etc. Obviously there will be times when generalizations should not be based only upon common references to any time at all.

Designation of operands as keywords has the opposite effect. If a keyword is designated, all facts retrieved must share a generalization or specialization of the keyword with the input fact.

Dictionary Construction

Our limited experience indicates that in an active experimental field, as the physiological study of memory currently is, dictionary terms are added, drop out, and even change meaning with amazing speed. A related problem is that if one person specifies a dictionary, his colleagues are very likely to have specific suggestions for improvements. While it is always possible to do dictionary maintenance on a separate run, it is also advantageous to be able to alter the dictionary somewhat during input of data. The special advantage of this is that it is a relatively painless operation for the user, and it can be done at the time that the inadequacy of the dictionary is called to his attention.

In the appropriate part of the control section (see Figure 1) the user may introduce new operands and define them in terms of previously established operands. Within a single run an operand may be introduced, and then used to define a new operand, providing that the operations are done in that order. The user may also indicate synonyms for terms in the dictionary. This is a convenient way to deal with adjectives. For example, in physiological psychology, the term "Hippocampal animals" is often used to indicate "Animals with lesions in the hippocampus, a region of the brain". By making the definition

HIPPOCAMPAL = HIPPOCAMPUS

appropriate retrievals will be made. Once again we have violated strict English meaning, but the violation does not disrupt the system too badly.

Updated versions of the current dictionary must be distributed to users frequently.

Current Status and Future Plans

The Reactive Library is now undergoing field tests. In addition, we hope to have an experimental evaluation of it using a body of users attacking a constructed problem with and without the program as an aid.

The initial reaction of users has been mixed, but not unfavorable. As a first test, the assertions concerning memory which are contained in an introductory physiological text (Deutsch and Deutsch, 1965) were input. In general, the relations between statements which the program found were mirrored by the remarks in the textbook. The dictionary was then modified and remodified until the agreement was quite good. This proved to be an effective way of establishing a dictionary.

Subsequent evaluations have been plagued by two sources of trouble, malfunctioning of the computer center's operating system, which is irrelevant, and user unfamiliarity, which is more serious. Some of our users have found it very hard to write simple sentences. Another problem is that people tend to take cryptic notes. The phrase "This affected it" is a classic example. Such a statement completely defeats a sentence-based information retrieval system. We have also found that even within a small, closely knit research group it is surprisingly hard to get agreement on terms, or even to get the same people to use a term consistently from one day to the next. (The mathematical linguist or mathematician may see this as a confirmation of his suspicions of experimental scientists, but we must take people as they are.) In spite of these hazards, which we regard as growing pains, we have every hope that the system will very shortly prove to be quite useful.

The RL is obviously quite a primitive program. In fact, we are surprised it works so well considering its crudity. Certain improvements, such as better procedures for pre-processing suffixes and prefixes, or introducing the ability to define new verbs with an input, are obvious and can be introduced by simple extensions of the program.

Another extension is more subtle. The RL infers metafacts solely by considering generalizations and specializations of individual facts. An ideal RL would also infer metafacts on the basis of the interaction between facts. Thus if "A produces B" and "B destroys C", it seems reasonable to hypothesize that "A inhibits C". Programs which make inferences to answer questions do exist and have been discussed in the open literature and in review papers (cf. Hunt, 1968; Simmons, 1965). These programs are illustrations, rather than working systems. (Cooper's 1963) is a possible exception.) The illustrations usually deal with either a constructed data base or with logical or mathematical statements, where there is good agreement on the rules of inference. Such agreement is not so easy to obtain in an actual experimental field; even qualified experts disagree. If the "A produces B" example is proposed to a physiological psychologist he is apt to say "What else does A produce?" or "How does B destroy C?" This does not mean that inferential fact retrieval is impossible, but it does suggest caution. Like it or not, most scientists probably do not use language with the precision used by a mathematician when he is talking about his subject. It is hard to see how a computer can help with this problem.

1. This research sponsored by the Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, Under AFOSR Grant No. AF-AFOSR-1311-67. Distribution of this document is unlimited.

2. We would also like to acknowledge the remarks of Professor Manfred Kochen, of the University of Michigan, who first pointed out to Earl Hunt that small research groups were active and appropriate users of such a system as the RL.

3. Any term written as A-B is interpreted as a single word. In the example which follows, it helps to know that Wistars, Tryon S3s, and Tryon S1s are types of rat, and that strychnine, picrotoxin, and 1757-IS are convulsant drugs.

4. A word for word copy of the sentence will be stored for output when this sentence is later retrieved, but plays no role in establishing the retrieval. Our present very simple pre-processing routines only strip terminal s's, a useful and cheap trick when dealing with English text. More sophisticated routines for cleaning up different forms of the same word will be used in later versions, following the technique of Stone, et al, (1967).

Figure 1

A typical input to the reactive library

1. REFERENCE

AUTHOR: Lutlges, M. and McGaugh, J.

TITLE: Permanence of Retrograde Amnesia Produced by ECS

JOURNAL: Science, 1967, 156, 408-410

DATE:

2. AUTHOR'S CONCLUSIONS (for reactive library)

   1. Post-trial ECS Produces Amnesia in Passive Avoidance Learning.
   2. ECS Produces Permanent Amnesia
   3.
   4.
   5.
   6.

3. REVIEWER'S AMPLIFICATION AND COMMENTS

   Mice were given electro-convulsive shock after being trained to avoid stepping off a ledge. Retests as much as one month later showed amnesia.

4. NEW DICTIONARY ENTRIES AND CORRECTIONS

1. Permanent C Time-reference	2.	3.
4.	5.	6.

5. REVIEWER'S FEEDBACK CONTROLS:

   BUZZWORDS = Time-reference

   KEYWORDS =

FIGURE 2 SAMPLE OUTPUT

```
****************************************************************************************

INPUT:       THE PRE-TRIAL INJECTION OF PICROTOXIN FACILITATES DISCRIMINATION LEARNING IN WISTARS #D&D, CH 3
STORED AS:   PRETRIAL INJECTION PICROTOXIN FACILITATE DISCRIMINATION LEARNING WISTAR

****************************************************************************************

INPUT:       INTRAPERITONEAL INJECTION OF 1757-IS HINDERS AVOIDANCE LEARNING IN TRYON-S1S  #D&D, CH 3
STORED AS:   IP INJECTION 1757-I HINDER AVOIDANCE LEARNING TRYON-S1

DISAGREEMENT WITH STORED FACT
             THE PRE-TRIAL INJECTION OF PICROTOXIN FACILITATES DISCRIMINATION LEARNING IN WISTARS  #D&D, CH 3
METAFACT ON WHICH DISAGREE:
             INJECTION CONVULSANT HINDER LEARNING RAT

****************************************************************************************

INPUT:       THE POST-TRIAL INJECTION OF STRYCHNINE IMPROVES DISCRIMINATION LEARNING IN TRYON-S3S  #D&D, CH 3
STORED AS:   POST-TRIAL INJECTION STRYCHNINE IMPROVE DISCRIMINATION LEARNING TRYON-S3

AGREEMENT WITH STORED FACT
             THE PRE-TRIAL INJECTION OF PICROTOXIN FACILITATES DISCRIMINATION LEARNING IN WISTARS  #D&D, CH 3
METAFACT ON WHICH AGREE:
             ADMINISTRATION-TIME INJECTION CONVULSANT IMPROVE DISCRIMINATION RAT

DISAGREEMENT WITH STORED FACT
             INTRAPERITONEAL INJECTION OF 1757-IS HINDERS AVOIDANCE LEARNING IN TRYON-S1S  #D&D, CH 3
METAFACT ON WHICH DISAGREE:
             INJECTION CONVULSANT IMPROVE LEARNING RAT
```

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Department of Psychology University of Washington Seattle, Washington 98105	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP

3. REPORT TITLE

THE REACTIVE LIBRARY

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Scientific Interim

5. AUTHOR(S) (First name, middle initial, last name)

Earl B. Hunt and J. Ross Quinlan

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
December 1, 1967	14	

8a. CONTRACT OR GRANT NO. AF-AFOSR-1311-67 b. PROJECT NO. 9779-01 c. 6144501F d. 681313	9a. ORIGINATOR'S REPORT NUMBER(S) 67-1-03 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. DISTRIBUTION STATEMENT

1. This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES TECH, OTHER	12. SPONSORING MILITARY ACTIVITY Air Force Office of Scientific Research 1400 Wilson Boulevard (SRLB) Arlington, Virginia 22209

13. ABSTRACT

An information retrieval program has been written to aid in coordinating literature searches by separate individuals. The program utilizes rudimentary techniques originally developed in question answering and theorem proving programs in order to determine whether a fact input to it is relevant to previously stored facts. The list of relevant facts, together with a citation to the article reporting them, is then printed out. The program is currently being used by a group of physiological psychologists on a field trial basis.

DD FORM 1 NOV 65 1473

Security Classification

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Information retrieval						
Question answering						
Computer						

Security Classification