# A RULE-BASED APPROACH TO INFORMATION RETRIEVAL: SOME RESULTS AND COMMENTS

Richard M. Tong, Daniel G. Shapiro,
Brian P. McCune and Jeffrey S. Dean

Advanced Information & Decision Systems
201 San Antonio Circle, Suite 286
Mountain View, CA 94040, USA.

## ABSTRACT

This paper is a report of our early efforts to use a rule-based approach in the information retrieval task. We have developed a prototype system that allows the user to specify his or her retrieval concept as a hierarchy of sub-concepts which are then implemented as a set of production rules. The paper contains a brief description of the system and some of the preliminary testing we have done. In particular, we make some observations on the need for an appropriate language for expressing conceptual queries, and on the interactions between rule formulation and uncertainty representation.

## I THE INFORMATION RETRIEVAL PROBLEM

Existing approaches to textual information retrieval suffer from problems of precision and recall, understandability, and scope of applicability. Boolean keyword retrieval systems (such as Lockheed's DIALOG) operate at a lexical level, and hence ignore much of the available information that is syntactic, semantic, or contextual. The underlying reasoning behind the responses of statistical retrieval systems [4] is difficult to explain to a user in an understandable and intuitive way, and systems that rely on a semantic understanding [5] must severely restrict the style and content of the natural language in the documents.

In the near future, large on-line document repositories will be made available via computer networks to relatively naive computer users. In this context, it is important that future retrieval systems possess the following attributes:

(1) Queries should be posed at the user's own conceptual level, using his or her vocabulary of concepts, and without requiring complex programming.

(2) The system should be able to provide partial matching of queries to documents, thereby acknowledging the inherent imprecision in the concept of a relevant document.

(3) The number of documents retrieved should be dependent upon the needs of the user (e.g., uses for the documents, time constraints on reading them).

(4) A logical, understandable, and intuitive explanation of why each document was retrieved should be available.

(5) The user should be able to easily experiment with and revise the conceptual queries, in order to handle changing interests or disagreement with previous system performance.

(6) Conceptual queries should be easily stored for periodic use by their author and for sharing with other users.

## II A RULE-BASED APPROACH

In our efforts to address the issues raised above, we have created a prototype knowledge-based information retrieval system called RUBRIC (**RU**le **B**ased **R**etrieval of **I**nformation by **C**omputer), in which queries are represented as a set of logical production rules [2].

The rules define a hierarchy of retrieval topics (or concepts) and subtopics. By naming a single topic, the user automatically invokes a goal oriented search of the tree defined by all of the subtopics that are used to define that topic. (i.e., a search process similar to that used in MYCIN [7]). The lowest-level subtopics are defined in terms of pattern expressions in a Text Reference Language, which allows keywords, positional contexts, and simple syntactic and semantic notions. The context functions restrict the pattern matching to occur in some specified syntactic context. So for example, one can specify that two patterns are of interest only if they occur in the same sentence or paragraph. Contexts can be made "fuzzy", giving RUBRIC the ability to find patterns that are "almost" within the same sentence or paragraph.

Our current implementation supports a variety of features including a simple explanation facility, variable thresholding and clustering of documents, one-level thesauri, and stem extraction on stories and queries.

### A. A Novel Rule Format

As in most other rule based systems, each rule in the query definition may have a user-defined heuristic weight which represents the degree to which the occurrence of the antecedent supports the occurrence of the consequent. That is the user can write rules of the form:

IF "the story is about topic A"

THEN "there is evidence to degree $\alpha$ that it is also about topic B"

However, in contrast to other systems we also provide an extended rule format, which enables the user to incorporate auxiliary (or contextual) evidence into the query.

Auxiliary evidence is evidence that by itself neither confirms or disconfirms our hypothesis, but which may decrease (or increase) our belief if seen in conjunction with some primary evidence. The syntax of such a rule is:

if A then C to degree $\alpha$
but if also B then C to degree $\beta$

where if $\alpha$ is greater than $\beta$ then B is disconfirming auxiliary evidence, and if $\alpha$ is less then $\beta$ then B is confirming auxiliary evidence. This has the effect of interpolating between $\alpha$ and $\beta$ depending upon the truth of the auxiliary clause B. Thus we might have a rule of the kind:

IF
"the story contains the literal string ´bomb´"

THEN
"it is about an <u>explosive</u> <u>device</u> with degree 0.6"

BUT IF
"it also mentions a <u>boxing</u> <u>match</u>"

THEN
"reduce the strength of the conclusion to 0.3"

Here we see the concept of disconfirming evidence in operation; notice that by itself "being about a <u>boxing</u> <u>match</u>" is not evidence that can be used to support or deny the conclusion we are trying to establish. (c.f., MYCIN which uses a concept of directly disconfirming evidence).

## III SOME EXPERIMENTS

A methodological advantage of working with the information retrieval problem is that we always know independently of RUBRIC whether or not the stories in the database are of interest. This makes it possible to conduct a variety of interesting experiments. We report on just two that we performed as a preliminary investigation of the validity of the RUBRIC model of information retrieval. First, we look at the improvements that can be achieved over a conventional Boolean keyword approach, then second, we explore the effects of using different calculi for propagating the uncertainty values within the system.

### A. Experimental Method

As an experimental database for testing the retrieval properties of RUBRIC we have used a selection of thirty stories taken from the Reuters News Service. Our basic experimental procedure is thus to rate the stories in the database by inspection (i.e., define a subjective ground truth), define a query, apply the query to the database, and then compare the rating produced by RUBRIC with the <u>a priori</u> rating.

RUBRIC´s basic task is to assign a weight to each story in the database. This weight is the truth of the statement "this story is relevant to the query", with its value being determined by propagating the uncertainty values through the tree defined by the rule-based query. This makes the assessment of performance somewhat complicated, since we are interested in the properties of the ordering, both in absolute terms (i.e., the truth values returned) and with reference to the ordering that we determined beforehand.

For the purposes of this discussion, however, we can concentrate on two basic measures. Both of these are based on the idea of using a selection threshold to partition the ordered stories so that those above it are "relevant" and those below it are "irrelevant". In the first we lower the threshold until we include all those deemed <u>a priori</u> relevant, and then count the number of unwanted stories that are also selected (denoted $N_F$). In the second we raise the threshold until we exclude all irrelevant stories, and then count the number of relevant ones that are not selected (denoted $N_M$). The first definition therefore gives us an insight into the system´s ability to reject unwanted stories (precision), whereas second gives us insight into the system´s ability to select relevant stories (recall).

### B. A Comparison with Boolean Retrieval

First we selected as a retrieval concept "Violent acts of terrorism", and then constructed an appropriate rule-based query. This is summarized in Figure 1 where we make extensive use of our extended rule format (indicated in the figure by the use of "Modifier" sub-trees). Application of this query to the story database results in the story profile shown in Figure 2. Notice that for presentation purposes the stories are ordered such that those determined to be <u>a priori</u> relevant are to the left of the figure and are further subdivided into definitely relevant and marginally relevant. In this case the ground truth defines nine definitely relevant stories, four marginally relevant ones, and seventeen that are not relevant. In Figure 2 each story rating assigned by RUBRIC is represented by a number in the interval [0,1]. A perfect profile would be one that gave the relevant stories a high rating, the marginal ones an intermediate rating, and the non-relevant stories a low rating. The performance scores for this output are:

<u>Precision</u>:
$N_F$ = 1 when we ensure that $N_M$=0, and

<u>Recall</u>:
$N_M$ = 5 when we ensure that $N_F$=0

This is excellent performance, being marred only by the selection of story (25) which, although it contains many of the elements of a terrorist article, is actually a description of an unsuccessful bomb disposal attempt. The lowest rated relevant story, (26), is one about the kidnapping and shooting of a minor political figure in Guatemala. In our ground truth this was given a marginal rating.

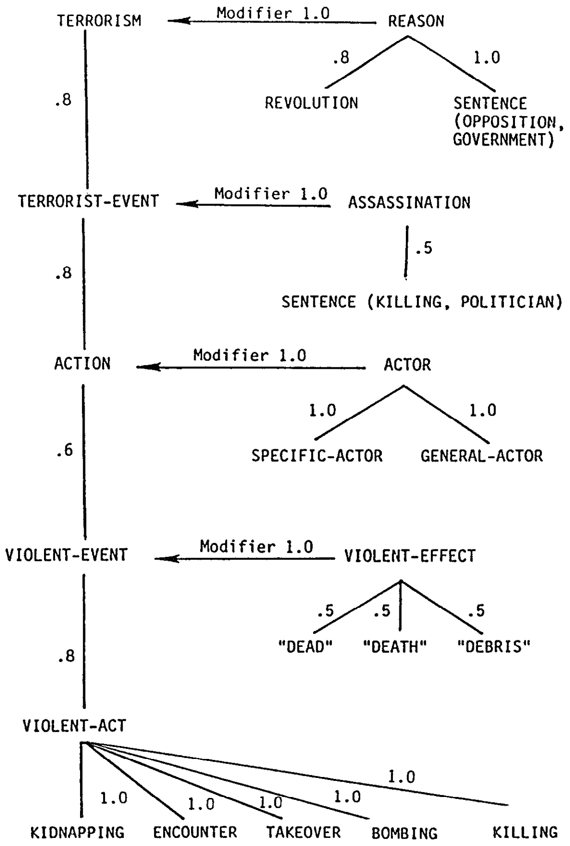

Figure 1. Rule Structure for "Acts of Terrorism"

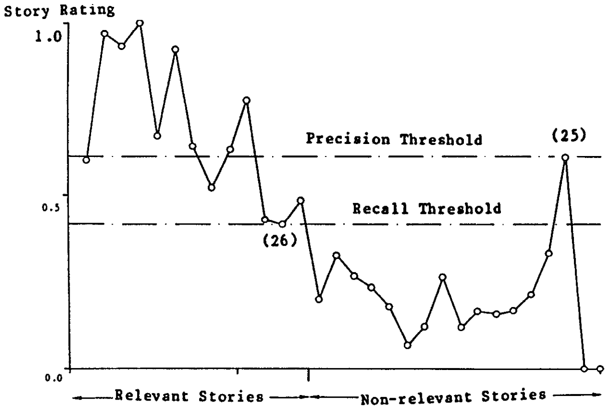

Figure 2. Story Profile from RUBRIC

To compare RUBRIC against a more conventional approach we constructed two Boolean queries using the rule-based paradigm, one of which is shown Figure 3 as an AND/OR tree of sub-concepts. The only difference between the two Boolean queries is that in the first we insist on the conjunction of ACTOR and TERRORIST-EVENT (as shown), whereas in the second we require the disjunction of these concepts.

When we compare the performance of these simulated Boolean queries to the query defined in the extended RUBRIC language we find that the conjunctive form of the Boolean query misses five relevant stories and selects one unimportant story; whereas the disjunctive form selects all the relevant stories, but at the cost of also selecting seven of the irrelevant ones.

While these results represent only a preliminary test, we believe they indicate that the RUBRIC approach allows the user to be more flexible in the specification of his or her query, thereby increasing both precision and recall. A traditional Boolean query tends either to over-constrain or under-constrain the search procedure, giving poor recall or poor precision. We feel that, given equal amounts of effort, RUBRIC allows better models of human retrieval judgement than can be achieved with traditional Boolean mechanisms.

### C. An Experiment with Uncertainty Calculi

Within the literature of expert systems, there has been a debate on the choice of "correct" calculus to represent and manipulate the uncertainty values. Indeed, there have been several attempts to construct a "calculus of uncertainty", some based on the concepts of probability and others on the more general formalisms of mathematical logic (see [6] and [9] for an introduction to some of these). In an attempt to clarify some of these issues, we have conducted a series of experiments in which we have adopted the view that the uncertainty values should be interpreted as representing the partial truth of the associated proposition. That being the case, we can use the formalism of multi-valued logic to define our calculus. Such logics have been studied extensively (see for example [3]), and lend themselves to efficient representation within the RUBRIC framework.

Our experiments consisted of fixing the query ("Acts of terrorism" as before) and changing the uncertainty calculus. There are, of course, a very large number of calculi, but we have concentrated on those in which the AND and OR connectives can be modelled as <u>triangular norms</u> and <u>triangular co-norms</u> respectively [1]. Prototypical examples are "min-max", viz:

$$v(A \text{ \underline{and} } B) = \min [v(A), v(B)]$$

$$v(A \text{ \underline{or} } B) = \max [v(A), v(B)]$$

and "pseudo-bayesian", viz:

$$v(A \text{ \underline{and} } B) = v(A).v(B)$$

$$v(A \text{ \underline{or} } B) = v(A) + v(B) - v(A).v(B)$$

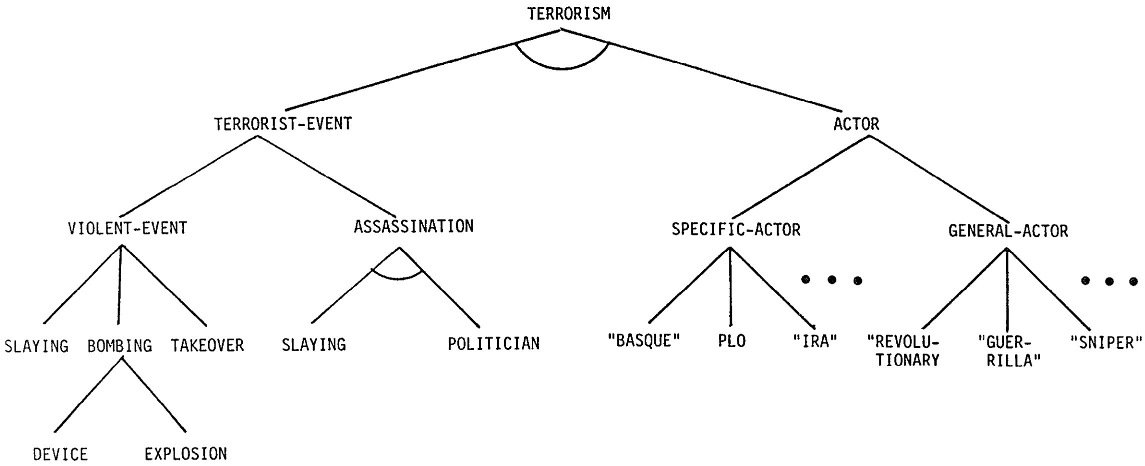

Figure 3. AND/OR Concept Tree for Boolean Query

We also selected a limited number of detachment operators (i.e., the operators used to compute the truth value of the consequent, v(B), given the truth value of the rule, v(A=>B), and the truth value of the antecedent, v(A)) of which a prototypical example is "product".

In all, we tested twenty different calculi and found marked variations in the $N_F$ and $N_M$ scores. Some calculi gave good performance whichever measure we used, some performed well with one but badly on the other, and others did poorly with both. Interestingly, no one calculus seemed to be significantly better than any of the others. We were, however, able to select four that as a group seemed to out-perform the others, and of these, three were based on the prototypical connectives mentioned above. (See [8] for more details.)

Our initial conclusion from this experiment is that a change in calculus can indeed have a marked effect on the interpretation of a query. However, in our search for an understanding of why some calculi did better than others we became aware of the fact that there is an inherent interdependency between rule writing and the uncertainty mechanisms. Thus although some calculi appear to be superior, this may not be because they are better in any absolute sense, but simply because we happened to construct a query that favored them. That is, if in developing a query we unconsciously expect the uncertainty values to behave in certain ways and write rules to exploit this, it would not be surprising to find that the calculus that most closely matched our expectations performed the best. This implies that we need to be more subtle in our investigation, and need to look more deeply into the role that the calculus plays. We address some of these issues in the next section.

## IV KNOWLEDGE AND UNCERTAINTY

Our experience with the development of RUBRIC has given us some valuable insights into the information retrieval process. In particular, we have gained a deeper understanding of the nature of both the knowledge required to describe a query, and the meaning of uncertainty in this context.

To introduce our comments, let us first observe that RUBRIC is not a story understanding system. Rather, it is a system which allows the user to construct a prototypical concept structure for the retrieval topic, and then returns a value that is a measure of the degree to which the story under consideration matches this structure.

To make RUBRIC an effective tool we need to provide the user with a set of appropriate query building constructs. We have mentioned the Text Reference Language and our extension to the standard IF...THEN... rule, but these are not sufficient and we need to consider others. Some of these will be determined by the nature of the particular application, but we believe that it is possible to develop a small set of primitives from which the user can build more elaborate forms.

In our attempts to build experimental queries, we have been struck by the concept of "evidence" and the fact that it is used in several paradigmatic ways. The notion of "auxiliary-evidence" motivated our extended rule format; but we can identify some others. For example, the concept of "weight-of-evidence" applies when we want to express the notion that no single piece of evidence allows us to deduce the occurrence of a topic to any significant degree, but if we have several such pieces then we would like their effect to be cumulative. We can also conceive of a situation in which we need a "cases-of-evidence" construct. That is, we want to use the best of several alternative lines of reasoning, even though each individual path might provide a good indication of the relevance of the story. Yet another form would be "directly-disconfirming-evidence", the occurrence of which reduces our belief in the relevance of a particular story at a global level.

Clearly, the existence of such canonical evidence schemata has implications for the choice of uncertainty calculus that we use. For example, the min-max calculus cannot be used to model the weight-of-evidence form, and the pseudo-probabilistic calculus cannot model the cases-of-evidence form. Similarly, none of the calculi we explored in our second experiment have any direct mechanism for modelling absolute disconfirmation. This leads us to conclude that the choice of calculus is not a decision to be made independently of rule writing, and indeed, it seems to us that we probably need to allow several calculi to co-exist within a given query.

An issue that quickly becomes apparent as more elaborate queries are constructed is the semantics of the relevance values. For example, all scalar-valued calculi, except min-max, have the property that very long chains of reasoning (i.e., in queries that have many levels in the hierarchy of sub-concepts) lead to relevance values which are very small. This is somewhat counter-intuitive; the user who constructs a very complex query to model his or her retrieval concept will in general get lower relevance values than those obtained from something more primitive. While this effect is a direct consequence of allowing non-Boolean reasoning, a user faced with story ratings the maximum of which is 0.2, say, will undoubtedly feel rather disconcerted.

One way to overcome this is by normalization, two forms of which we might consider. First, we could normalize by scaling the retrieved stories so that the most relevant is assigned the value 1.0. This is merely a presentation device (in fact one we used in our first experiment) and is appropriate when the user is only concerned with the relative evaluations of sets of stories. Alternatively, we could compute the maximum value that any story could receive (i.e., the value obtained by setting all the terminal values of the query to 1.0) and divide the actual story values by it. This is more fundamental, since it amounts to a redefinition of the meaning of the uncertainty values attached to the rules.

Finally, we have observed that as queries become more complex there seems to be a reduced sensitivity to both the absolute value of the heuristic weight assigned to rules and to the choice of uncertainty calculus. This suggests that we might not need the precision provided a numerical representation scheme, and we conjecture that a form of symbolic uncertainty calculus may be the way to proceed. Thus rather than evaluate the strength of support that some evidence gives to a hypothesis as 0.8, say, it is more natural, and appropriate, to label it by something like "very strong". We conjecture that as we develop a more expressive language for query construction, there will be decreasing emphasis on uncertainty as a numerical adjunct to rule writing, and a realization that knowledge about uncertainty enters directly into the expression of retrieval concepts.

## V <u>FUTURE DIRECTIONS</u>

In the future we plan to conduct additional experiments to investigate the effectiveness of providing a richer language for writing queries. In particular, we will be looking at the impact of the cases-of-evidence and weight-of-evidence rule forms discussed in the previous section in terms of both their expressiveness and usability within a text reference language. As part of this effort, we will examine the effect of allowing several uncertainty calculi within the same rule-base, playing particular attention to the rule to rule interface questions that will arise. Finally, we will investigate the feasibility of using symbolic rather than numeric representations of uncertainty.

## REFERENCES

[1] Dubois, D. and H. Prade, "A Class of Fuzzy Measures based on Triangular Norms", <u>Int. J. General Systems</u> 8 (1982) 43-61.

[2] McCune, B.P., J.S. Dean, R.M. Tong and D.G. Shapiro, "RUBRIC: A System for Rule-Based Information Retrieval," Final Technical Report, TR-1018-1, Advanced Information & Decision Systems, Mtn. View, CA., 1983.

[3] Rescher, N., <u>Many Valued Logic</u>. McGraw-Hill, New York, 1969.

[4] Salton, G., M.J. McGill, <u>Introduction to Modern Information Retrieval</u>, McGraw-Hill, New York, 1983.

[5] Schank, R.C. and G. DeJong, "Purposive Understanding", In J.E.Hayes, D.Michie and L.I.Mikulich (eds.), <u>Machine Intelligence</u> 1979, chpt. 24.

[6] Shafer, G., <u>A Mathematical Theory of Evidence</u>. Princeton Univ. Press, 1976.

[7] Shortliffe, E.H., <u>Computer Based Medical Consultations: MYCIN</u>. American Elsevier Publishing Co. Inc., 1976.

[8] Tong, R.M. and D.G. Shapiro, "An Experiment with Multiple-Valued Logics in an Expert System", In <u>Proc. IFAC Symp. on Fuzzy Information, Knowledge Representation and Decision Analysis</u>. Marseille, July, 1983.

[9] Zadeh, L.A., "Approximate Reasoning Based on Fuzzy Logic", In <u>Proc. IJCAI-79</u>. Tokyo, August, 1979, pp. 1004-1010.