A Retrieval Model Incorporating Hypertext Links

W.B. Croft and H. Turtle
Department of Computer and Information Science
University of Massachusetts
Amherst, MA 01003

COINS Technical Report 89-84

A Retrieval Model Incorporating Hypertext Links

W.B. Croft and H. Turtle*
Department of Computer and Information Science
University of Massachusetts, Amherst, MA. 01003

August 10, 1989

Abstract

Formal models of retrieval provide the basis for retrieving hypertext nodes in response to explicit queries. This retrieval facility can be integrated with the typical browsing facility for effective access to large databases. In this paper we describe a retrieval model developed for bibliographic information retrieval and show how hypertext links can be incorporated. The model treats retrieval as a form of inference in which multiple sources of evidence, including hypertext links, are combined and used to rank the retrieved nodes. Implementation aspects of the model are also discussed.

1 Introduction

The field of Information Retrieval (IR) has focused on the development and evaluation of retrieval models for text documents, such as those found in bibliographic databases [16,13,1]. These retrieval models specify strategies for evaluating documents with respect to a given query, typically resulting in a ranked output. Hypertext researchers, on the other hand, have emphasized flexible organizations of multimedia "nodes" through connections made with user-specified links, and interfaces that facilitate browsing in this network of links. A number of approaches to the integration of query-based retrieval strategies and browsing in hypertext networks have been proposed. The I^{3}R system [6,22] and a medical handbook system described by Frisse [7], for example, use query-based retrieval strategies to form a ranked list of candidate "starting points" for hypertext browsing. The I^{3}R system also uses feedback during browsing to modify the initial query and locate additional starting points. The important issue from an IR perspective is the choice of a retrieval model, and consequently a retrieval strategy, for hypertext. This choice will have an impact on the effectiveness of retrieval and on the system implementation. A retrieval model can also provide a more formal specification of the meaning of some hypertext links.

It may appear that the highly connected structure of a hypertext database and the variety of link types that may be used to make those connections distinguishes it from a

*On leave from OCLC Online Computer Library Centre

conventional text database. Network structures are, in fact, not new to IR and retrieval models have been proposed that use automatically and manually generated links between documents and the concepts or terms that are used to represent their content. *Document clustering*, for example, is a retrieval model based on links between documents that are automatically generated by comparing similarities of content [23]. *Citations* are another form of link between documents that has been studied [18]. Links between terms can be derived from *term clustering* [19,12] or the information in a manually-generated thesaurus [8]. Network representations that attempt to integrate this information have been developed [10,6], but these have been used primarily for browsing. Recent studies have shown that retrieval effectiveness may be improved by combining the evidence represented by the different link types in a network representation [4], and our approach has been to develop a retrieval model to formalize this approach. In this paper, we describe a model based on Bayesian inference networks [11] that appears to capture the major aspects of previous probabilistic IR models [13] and can easily be extended to include hypertext links. The possibility of such a model was also referred to by Frisse [7].

In the next section of the paper, we will illustrate how multiple sources of evidence can be used to answer a query and how various types of hypertext links can be regarded as forms of evidence. This will be done using a deductive database representation of the problem. We will then describe a non-deductive or plausible inference approach to retrieval. In particular, we will describe how a Bayesian inference network can be constructed from the information in a hypertext database and how this is used to answer queries. In the final section of the paper we discuss alternative approaches and implementation issues.

Throughout this paper, we shall be discussing the effectiveness of retrieval techniques. In contrast to measures of retrieval efficiency based on space and time requirements, measures of effectiveness are based on how successful a technique is in locating the relevant documents for a query. The usual effectiveness measures are *recall* and *precision*, which are, respectively, the proportion of relevant documents that are retrieved and the proportion of retrieved documents that are relevant. These measures and a variety of others are described extensively in the IR literature [13].

2 Retrieval as Inference

Retrieval involves a comparison of the things to be retrieved (in this case, hypertext nodes) and a query. During browsing, the query remains implicit and the comparison is done by the searcher. We are primarily concerned here with the situation where an explicit query exists and the comparison is done by an automated retrieval mechanism. In simpler systems, such as those based on string matching, the comparison tests equality of strings contained in the query and node representations. In general, however, determining that the query and a node are related will require an inference mechanism. To illustrate this, consider the deductive database (i.e. Prolog-like [2]) representation of a hypertext database shown in Figure 1. The first part of the database specifies facts about the nodes. In particular, the about predicate is used to "index" the nodes. That is, we represent the content of

```
about(node1,concept2).
about(node1,concept23).
about(node1,concept34).
about(node2,concept4).
about(node2,concept34).
about(node3,concept15).
.  .
.  .
synonym(concept2,concept8).
synonym(concept23,concept56).
is_a(concept12,concept3).
nn(concept3,concept12).
nn(concept34,concept17).
.  .
nn(node2,node34).
nn(node14,node12)
cites(node5,node45).
cites(node5,node22).
links_to(node2,node41).
links_to(node20,node26).
links_to(node2,node3).
part_of(node5,node36).
.  .
.  .
about(Node,ConceptX) :- about(Node,ConceptY),
                        synonym(ConceptX,ConceptY).
about(Node,ConceptX) :- about(Node,ConceptY),
                        nn(ConceptX,ConceptY).
about(Node,ConceptX) :- about(Node,ConceptY),
                        is_a(ConceptX,ConceptY).
.  .
about(NodeX,Concept) :- about(NodeY,Concept),
                        nn(NodeX,NodeY).
about(NodeX,Concept) :- about(NodeY,Concept),
                        cites(NodeY,NodeX).
about(NodeX,Concept) :- about(NodeY,Concept), %hypertext rule 1
                        links_to(NodeY,NodeX).
about(NodeX,Concept) :- about(NodeY,Concept), %hypertext rule 2
                        part_of(NodeX,NodeY).
```

Figure 1: A Hypertext Deductive Database

the node by associating it with particular concepts. This association can be established in a variety of ways, including manual term selection or automatic assignment based on statistical or natural language processing techniques [20,9]. Note that in our example we have used generic concepts such as concept2 rather than concepts actually derived from text. The second part of the database specifies the current links between nodes. These are of various types, including:

- Links between nodes derived by statistical "nearest neighbor" measures - `nn`
- Links between nodes derived from citations in the text - `cites`
- Links between nodes that represent a structural hierarchy - `part_of`
- Other links between nodes specified manually - `links_to`

The last two link types are those found in a typical hypertext database, whereas the other two are more typical of bibliographic retrieval systems. The database also specifies a variety of relationships between concepts, such as those derived from a thesaurus (`synonym`, `is_a`), that may not be represented as links in a hypertext database.

The remainder of the deductive database specifies rules that can be used to derive additional facts about nodes from the stated facts. For example, the rule labelled "hypertext rule 1" can be used by a deductive inference engine to derive facts about connected nodes such as node2 and node41. In this case, we can derive the facts `about(node41,concept4)` and `about(node41,concept34)`. In other words, the meaning of a node is partially derived from the nodes connected to it. We shall refine this "definition" of a hypertext link in the next section. Note that some types of hypertext link may not participate in the retrieval process. A link type that connects the nodes in a particular sequence, for example, might be used only for browsing the database and not for direct retrieval.

A query in a deductive database is of the form $\{N|W(N)\}$, which can be read as "Retrieve all nodes N such that $W(N)$ can be shown to be true in the current database". For simplicity, we are assuming that N is the only free variable in the formula W. In our Prolog-like syntax, an example query might be

```
?- about(N,concept15), about(N,concept34).
```

The only N that satisfies this query is node3, which will be retrieved. Note that the `links_to` rule was used to prove the query is true for this node.

In general, proving the query and thereby retrieving nodes may involve the use of all the rules specified in our example database. Removing a rule associated with a particular link type may result in some nodes not being retrieved. Each of the link types that are used in rules, therefore, can be regarded as a form of evidence that is used by the retrieval mechanism. Effective retrieval requires combination of multiple sources of evidence.

Is deductive retrieval an appropriate retrieval model for hypertext? There is, in fact, considerable evidence that such a model would have poor effectiveness [17]. The major problem lies in the uncertainty associated with natural language. This uncertainty affects

all aspects of our deductive database example, including the facts (is concept2 an accurate description of node1?), the rules (some rules may be more certain than others), and even the query (some parts of the query may be more important than others). A retrieval model must take this uncertainty into account to produce effective results. For this reason, we are looking at retrieval models based on *plausible* or non-deductive inference. In the next section we shall describe one such model that incorporates uncertainty and allows combination of multiple sources of evidence.

3 Bayesian Inference Networks

A number of approaches to plausible inference have been proposed [21]. These include heuristic certainty measures, non-standard logics, Bayesian inference networks [11], the Dempster-Shafer model of evidential reasoning, and symbolic representations [3]. We have chosen to use Bayesian inference networks, primarily because information retrieval research has had considerable success with simple probabilistic models [13,1] and because Bayesian networks are similar to the representations used in some advanced IR systems (e.g. I^3R).

A Bayesian inference network is a directed, acyclic dependency graph in which nodes represent propositional variables or constants and arcs represent dependence relations between propositions (see [11] for a complete treatment of Bayesian nets). For hypertext applications, the roots of the dependency graph are hypertext nodes; interior nodes and leaves represent concepts. If a proposition represented by a node p directly implies the proposition represented by node q, we draw a directed arc from p to q. If-then rules in Bayesian nets are interpreted as conditional probabilities, that is, a rule $A \rightarrow B$ is interpreted as a probability $P(B|A)$. The arc connecting A with B is labelled with a matrix that specifies $P(B|A)$ for all possible combinations of values of the two nodes. The set of matrices labelling arcs pointing to a node characterize the dependence relationship between that node and the nodes representing propositions naming it as a consequence. For the class of rule bases we are interested in, these link matrices are fixed by the rule base and do not change during query processing. Given a set of prior probabilities for the roots of the DAG, these compiled networks can be used to compute the probability or degree of belief associated with the remaining nodes.

Different restrictions on the topology of the network and assumptions about the way in which the connected nodes interact lead to different schemes for combining probabilities. In general, these schemes have two components which operate independently: a predictive component in which parent nodes provide support for their children (the degree to which we believe a consequent depends on the degree to which we believe the antecedents of the rules naming it), and a diagnotic component in which children provide support for their parents (if our belief in a proposition increases or decreases, so does our belief in the antecedents that name the proposition as a consequent). The propagation of probabilities through the net can be done using information passed between adjacent nodes.

To see how a Bayesian network relates to the underlying probabilistic model, consider the simple network shown in Figure 2 which represents a database where there are no

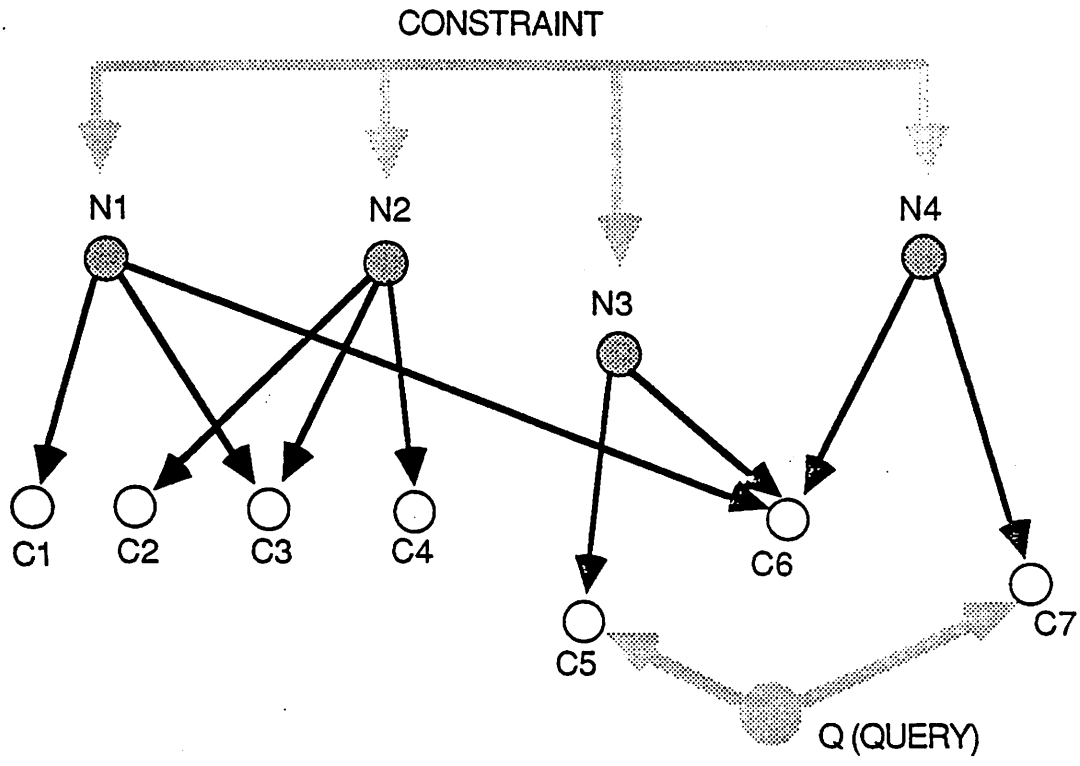

Figure 2: A Simple Bayesian Network

connections between the hypertext nodes (i.e. simple document retrieval). In this network, the nodes labelled N1 to N4 represent propositions that a hypertext node has a particular content. The specific content is that of the corresponding hypertext node. The nodes labelled C1 to C7 represent propositions that a hypertext node is represented (indexed) by a particular concept. The directed links from an N node to a C node represent the conditional probability $P(C|N)$, which is the probability that a particular concept represents a hypertext node given that the node has a particular content.

The query node Q represents a conjunction of C propositions. During retrieval, a constraint is applied to the network that restricts a single N node to be initially true. The network is then used to compute the probability of Q being true given this particular content (or hypertext node). This constraint is applied to each N node in turn (conceptually, at least) and the end result is a ranking of hypertext nodes according to the values of $P(Q)$. Note that retrieval, as in the deductive database example, is implemented as a form of inference. In the particular forms of Bayesian inference networks used here, we could in fact regard $P(Q)$ as the probability that Q could be proved true.

We can also describe this use of the network in terms of an explicit probabilistic model. It can be shown that, given certain assumptions, the optimum retrieval effectiveness will be obtained by ranking nodes (or text documents) according to estimates of $P(Q|N)$, which is

the probability that the query is true given a particular hypertext node [15]. This *Probability Ranking Principle* is usually expressed in terms of probability of *relevance*, but we believe that this formulation is actually more precise. The probability $P(Q|N)$ can be tranformed into an expression that involves the indexing concepts using standard probability theory as follows:

If we use $R_i = (C_1, C_2, ..C_m)$ to refer to a particular indexed representation of a node, or in other words an assignment of concepts to a node (C_i is true when assigned), then

$$P(Q|N) = \sum_i P(Q|R_i, N)P(R_i|N)$$

and since Q and N are conditionally independent with respect to R,

$$P(Q|N) = \sum_i P(Q|R_i)P(R_i|N)$$

Now we apply Bayes' theorem,

$$P(Q|N) = \sum_i \frac{P(R_i|Q)P(Q)}{P(R_i)} P(R_i|N)$$

At this point, we can use various approximations to estimate these probabilities. The most common assumption is independence, e.g.

$$P(R_i|Q) = \prod_{j=1}^{m} P(C_j|Q)$$

although we shall see that the inference network can represent dependencies between concepts. Given the independence assumptions, we get

$$P(Q|N) = P(Q) \prod_{j=1}^{m} (\frac{P(C_j|Q)}{P(C_j)} P(C_j|N) + \frac{1 - P(C_j|Q)}{1 - P(C_j)} (1 - P(C_j|N))$$

This expression is the basis of the network shown in Figure 2. When the appropriate estimates for the probabilities in this expression are introduced, a function for ranking nodes is produced. In this case, the ranking function is approximately equivalent to giving each node a score that is the sum of the weights of matching query terms, where the weights depend on the frequency of occurrence of a term in each hypertext node and in the entire collection of hypertext nodes (Salton's *tf.idf* weight [16]).

As more complicated models are introduced, the inference network becomes a more convenient representation for calculation of the probabilities. For hypertext databases, the most important extension to the simple retrieval model is to introduce dependencies that represent the links between hypertext nodes. In terms of the probabilistic model, this means that we introduce probabilities of the form $P(N_j|N_k)$ in the expression

$$P(C_i|N_k) = P(C_i|N_j)P(N_j|N_k)$$

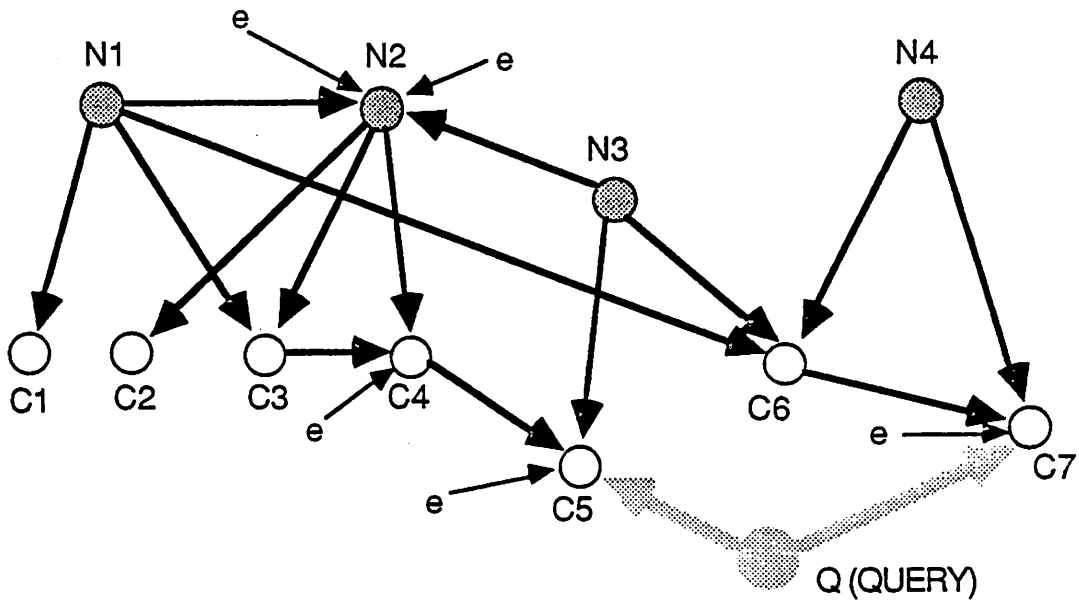

Figure 3: Bayesian Inference Network Incoporating Node and Concept Dependencies.

Intuitively, this means that if hypertext node j is indexed by a particular concept and is linked to hypertext node k, there is some probability that hypertext node k should also be indexed by that concept. This is the non-deductive form of the hypertext link rules in the example deductive database in section 2.

Figure 3 shows the simple Bayesian inference network extended to include dependencies between hypertext nodes and between concepts. The probabilities of the form $P(C_j|C_k)$, which capture the thesaurus and statistical information mentioned in the deductive database example, allow us to calculate $P(Q|N)$ more accurately than the simple model assuming independence. We have introduced *evidence* nodes (those labelled e) in this network to show that the dependencies between hypertext nodes and concepts are of the form $P(N_j|N_k, e)$. The evidence is the facts stored in the database, such as `linked_to(Node1,Node2)`. When several pieces of evidence relate to a conclusion, such as mutiple links between the same hypertext nodes, a model of disjunctive interaction [11] can be used.

The extended model has a significant impact on the hypertext nodes that are retrieved. Consider two hypertext nodes that are ranked equally according to their own contents. If one of these nodes is connected to other nodes and the other is not, then the extended model will rank the connected node higher. This is a desirable property for a retrieval strategy that is designed to find starting points for browsing. The other advantage of the

extended model is that the higher ranked documents will be more likely to satisfy the user's information need, or in other words, the effectiveness of retrieval will improve.

Another extension that can be made to the network is to allow uncertain queries. The easiest way to understand this is to imagine that users express their information needs in the form of text. It will then be possible to derive a range of queries from this text in the same way that different indexed representations of hypertext nodes can be derived from their contents. The network can then be regarded as computing $P(I|N)$, or the probability that a textual information need is satisfied given a particular node. More specifically, we can develop the extended model as follows:

Starting with the expression derived previously and replacing Q with I we get

$$P(I|N) = \sum_i \frac{P(R_i|I)P(I)}{P(R_i)} P(R_i|N)$$

If we now include the intermediate variable Q as was done for the node representations, the expression becomes

$$P(I|N) = \sum_i (\frac{P(I)}{P(R_i)} P(R_i|N) \sum_j P(R_i|Q_j)P(Q_j|I))$$

The probabilities could then be estimated using independence assumptions as before.

The practical use of this extension is to allow users to specify which parts of their query are more important, as was done in the I^{3}R and OFFICER systems [5]. This amounts to allowing them to estimate $P(Q_j|I)$.

4 Implementation Issues

The retrieval model we have described is of theoretical interest since it integrates and extends existing retrieval models, but its practical significance depends upon the degree to which it allows us to improve retrieval in real hypertext applications. In this section we review basic implementation approaches and open research issues.

As described previously, a hypertext network is similar to advanced document databases that have been used in various experimental settings. Existing automated indexing techniques can be used to create the set of concept nodes and link them to hypertext nodes. Similarly, techniques exist for generating nearest neighbor, citation and other types of links.

The links and nodes of the Bayesian inference net can be handled either as new hypertext nodes and links that are normally invisible to the user or as a separate network that is used as an index to the hypertext database. An integrated hypertext/inference network would allow the retrieval engine to make direct use of hypertext links already present, whereas a separate inference network is probably easier to implement with existing hypertext systems. Further, the information associated with a link in the inference network is different than that typically associated with a hypertext link (i.e. a probability matrix).

Other open issues are:

- Topology restrictions. Inference networks cannot contain directed cycles (a node cannot support itself). This restriction limits the class of recursive rules and predicates that can be represented. For example, a rule like

  ```
  about(NodeX,ConceptY) :- about(NodeX,ConceptX), synonym(ConceptX,ConceptY).
  ```

  can be represented because it has a finite extension and need not introduce a cycle, but mutually recursive predicates are not allowed, nor are base predicates that would induce a cycle (e.g., `cites(NodeX,NodeY)` and `cites(NodeY,NodeX)`).

- The networks are computationally simpler when they are singly connected (i.e. when each path from the document node to the query has no shared intermediate nodes). When the paths are not distinct, special techniques are required to augment the basic belief propagation rules. While several techniques are known, we will need experience with real hypertext databases to determine how often they are required, which techniques work best in practice, and what computational burden they impose.

- While most queries handled by conventional retrieval systems can be answered using the precompiled network, some important query forms cannot. In particular, queries which involve evaluable predicates (e.g. `greater`, `less`) or multiple document variables present problems. This suggests that deductive and non-deductive inference will need to be integrated in practical systems.

While the practical utility of inference networks has yet to be established, it is important to note that they represent extensions to current IR practice. We know, for example, that we can successfully represent large collections of text documents and produce effective rankings of those documents based on automatically derived probability estimates. Further work is required to validate this new retrieval model in a hypertext setting and to compare the effectiveness of retrieval techniques based on inference networks to those based on simpler probabilistic models.

Acknowledgments

This work was supported in part by OCLC Online Computer Library Center, by Rome Air Development Center and Air Force Office of Scientific Research under contract F30602-85-C-0008, and by NSF Grant IRI-8814790.

References

[1] Belkin, N. and Croft, W.B., "Retrieval Techniques". *Annual Review of Information Science and Technology*, Edited by M.E. Williams, Elsevier Science Publishers, 22, 110-145, 1987.

[2] Clocksin, W.; Mellish, C. *Programming in Prolog.* Springer-Verlag, New York, 1981.

[3] Cohen, P.R.; Kjeldsen, R. "Information Retrieval by Constrained Spreading Activation in Semantic Networks", *Information Processing and Management*, 23, 255-268, 1987.

[4] Croft, W.B., Lucia, T.J., Cringean, J, and Willett, P. "Retrieving Documents by Plausible Inference: An Experimental Study", *Information Processing and Management*, (in press).

[5] Croft, W.B.; Krovetz, R. "Interactive Retrieval of Office Documents", Proceedings of the ACM Conference on Office Information Systems, 228-235, 1988.

[6] Croft, W. B.; Thompson, R., "I^3R: A New Approach to the Design of Document Retrieval Systems", *Journal of the American Society for Information Science*, 38, 389-404, 1987.

[7] Frisse, M.E., "Searching for Information in a Hypertext Medical Handbook", *Communications of the ACM*, 31, 880-886, 1988.

[8] Humphrey, S.M.; Miller, N.E., "Knowledge-Based Indexing of the Medical Literature: The Indexing Aid Project", *Journal of the American Society for Information Science*, 38, 184-196, 1987.

[9] Lewis, D., Croft, W.B. and Bhandaru, N., "Language-Oriented Information Retrieval", *International Journal of Intelligent Systems*, (in press).

[10] Oddy, R.N. "Information Retrieval Through Man-Machine Dialogue", *Journal of Documentation*, 33, 1-14, 1977.

[11] Pearl, J., *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, California, 1989.

[12] Van Rijsbergen, C.J., "A Theoretical Basis for the Use of Co-Occurrence Data in Information Retrieval", *Journal of Documentation*, 33, 106-119, 1977.

[13] Van Rijsbergen, C.J., *Information Retrieval*. Butterworths, London, 1979.

[14] Van Rijsbergen, C.J.. "A Non-Classical Logic for Information Retrieval". *Computer Journal*, 29, 481-48, 1986.

[15] Robertson, S.E. "The Probability Ranking Principle in IR", *Journal of Documentation*, 33, 294-304, 1977.

[16] Salton, G. and McGill, M., *An Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1983.

[17] Salton, G.; Fox, E.A.; Wu, H. "Extended Boolean Information Retrieval", *Communications of the ACM*, 26, 1022-1036, 1983.

[18] Small, H. "Co-Citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents". *Journal of the American Society for Information Science*, 24, 1973.

[19] Sparck Jones, K. *Automatic Keyword Classification for Information Retrieval.* Butterworths, London, 1970.

[20] Sparck Jones, K. "Automatic Indexing", *Journal of Documentation*, 30, 393-432, 1974.

[21] Spiegelhalter, D.J. "A Statistical View of Uncertainty in Expert Systems", in *Artificial Intelligence and Statistics*, 17-55, Addison-Wesley, Reading, 1986.

[22] Thompson, R. and Croft, W.B., "Support for Browsing in an Intelligent Text Retrieval System", *International Journal of Man-Machine Studies*, 30, 639-668, 1989.

[23] Willett, P., 'Recent Trends in Hierarchic Document Clustering: A Critical Review', *Information Processing and Management*, 24 (5), 577-598, 1988.