AD-A235 015

2

REPORT DOCUMENTATION PAGE

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION/DOWNGRADING SCHEDULE

3. DISTRIBUTION/AVAILABILITY OF REPORT
Approved for public release; distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S)
AFOSR-[illegible]

6a. NAME OF PERFORMING ORGANIZATION
Univ. of Massachusetts

6b. OFFICE SYMBOL (If applicable)

7a. NAME OF MONITORING ORGANIZATION

6c. ADDRESS (City, State and ZIP Code)
Amherst, MA 01003

7b. ADDRESS (City, State and ZIP Code)
AFOSR/NM
Bldg 410
Bolling AFB DC 20332-6448

8a. NAME OF FUNDING/SPONSORING ORGANIZATION
AFOSR/NM

8b. OFFICE SYMBOL (If applicable)
NM

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
AFOSR-90-0110

8c. ADDRESS (City, State and ZIP Code)
Bld 410
Bolling AFB DC 20332-6448

10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO
61102F	2304	A7	

11. TITLE (Include Security Classification) (non-classified)
Retrieval Using Plausible Inference

12. PERSONAL AUTHOR(S)
W. Bruce Croft

13a. TYPE OF REPORT
Final Technical

13b. TIME COVERED
FROM 90/01/01 TO 90/12/31

14. DATE OF REPORT (Yr., Mo., Day)
91/04/04

15. PAGE COUNT
18

16. SUPPLEMENTARY NOTATION
Grant #AFOSR-90-0110 Purchase Request #FQ8671-9000661

17. COSATI CODES

FIELD	GROUP	SUB. GR.

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Retrieval is a crucial function in information systems, but the algorithms used in many systems are not effective in locating the required information. For example, text-based systems are an important class of information system where both the query specification and the stored information will be incomplete and uncertain. In this situation, simple models of retrieval based on deductive inference are not adequate. We have proposed a new computational framework that views retrieval as a process of combining multiple sources of evidence, and have carried out a number of experiments in the domain of text retrieval. The experiments show that significant retrieval effectiveness improvements are possible. We have also made progress in the area of text representation using natural language processing, which provides additional evidence for the retrieval model.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. □ DTIC USERS □

21. ABSTRACT SECURITY CLASSIFICATION
UNCLASSIFIED

22a. NAME OF RESPONSIBLE INDIVIDUAL
ABRAHAM WAKSMAN

22b. TELEPHONE NUMBER (Include Area Code)
202-767-5028

22c. OFFICE SYMBOL
NM

91 4 25 047

Retrieval using Plausible Inference

Final Report to AFOSR from

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

April 4, 1991

Abstract

Retrieval is a crucial function in information systems, but the algorithms used in many systems are not effective in locating the required information. For example, text-based systems are an important class of information system where both the query specification and the stored information will be incomplete and uncertain. In this situation, simple models of retrieval based on deductive inference are not adequate. We have proposed a new computational framework that views retrieval as a process of combining multiple sources of evidence, and have carried out a number of experiments in the domain of text retrieval. The experiments show that significant retrieval effectiveness improvements are possible. We have also made progress in the area of text representation using natural language processing, which provides additional evidence for the retrieval model.

1 The Problem

One of the primary functions in many applications is the retrieval of objects that satisfy criteria specified in a user's query. Examples of objects that may be retrieved in this way include natural language documents or parts of documents (Belkin and Croft, 1987; Croft and Thompson, 1987), multimedia objects containing graphics, image and voice as well as text (Weyer and Borning, 1985; Croft, 1987), and fragments of encylopedic knowledge bases for AI applications (Lenat et al, 1986). In many cases, these applications cannot be handled using conventional database technology (Date, 1986) because of the complexity of the objects being stored and because the process of determining if a query criterion is satisfied may involve inference. One approach to this problem would be to represent objects and knowledge about objects in a deductive database system (Gallaire, Minker and Nicolas, 1984). In such a system, a query can be expressed in the form $q = \{X | W(X)\}$ where X is a vector of domain variables and $W(X)$ is a formula in which X are the only free variables. Retrieval involves finding all instances of X for which $W(X)$ can be proved. In other words, for a query q, retrieve X where $KB \models W(X)$. The knowledge base, KB, includes descriptions of objects (extensional data), rules (intensional data), and basic axioms. The main issue in implementing deductive database systems is designing efficient inference methods.

The critical issue for us, however, is the *effectiveness* of retrieval. By this, we mean how well the system does at locating objects that are judged relevant by the user. This has been a central focus of the research in information retrieval (IR) and a number of evaluation measures and methodologies have been developed (Van Rijsbergen, 1979). Less than perfect retrieval is the result of people viewing objects not retrieved by the system as being relevant and viewing some objects that are retrieved (i.e. satisfy the query criteria) as not relevant. As we do retrieval experiments, we quickly realize that the usual situation is that the query specification, the object descriptions, and the rules in the knowledge base are incomplete and often errorful. In these situations involving uncertain information, *deductive inference does not provide effective retrieval.* Instead, retrieval must be implemented as a process of plausible inference or evidential reasoning. The classic example of this problem in IR is the common use of Boolean query formulations and string matching techniques in many commercial systems. It has been shown in a number of experiments that techniques based on probabilistic models are much more effective. There has also been significant evidence for the evidential nature of retrieval in that searches based on different aspects of a text object's representation (e.g. full text, citations, keywords) have been shown to retrieve different sets of relevant objects (for example, Katzer et al, 1982: Croft et al, 1989).

In this project, we have made significant progress in the following areas:

- We have developed a retrieval model based on Bayesian inference networks and have shown that this model subsumes previous models such as the probabilistic model, cluster-based retrieval, Boolean retrieval, and even hypertext (Croft and Turtle, 1989; Turtle and Croft, 1990). We have also described how this model could be used for retrieval of objects with complex structure (Croft, Krovetz and Turtle, 1990).

- We have carried out a number of experiments with natural language aspects of text representation such as word senses (Krovetz and Croft, 1989, 1991; Krovetz, 1990a, 1990b), nominal compounds (Gay and Croft, 1990), and syntactic phrases (Lewis, Croft and Bhandaru, 1989; Lewis and Croft, 1990; Lewis, 1990).

- We have carried out experiments with databases of 1-3000 documents that show that our retrieval model can produce significant performance improvements (up to 60% improvement in average precision[1])(Turtle, 1990; Turtle and Croft, 1991b). We have also shown that these techniques can be used efficiently on much larger databases (Turtle, 1990; Turtle and Croft, 1991a).

In the following section, we give an overview of the retrieval model and its relationship to other models. In the third section we describe some of the results in more detail.

2 An Inference Network Model

2.1 Inference Networks

A number of inference techniques developed for use with expert systems can be adapted to the text retrieval problem. The model we present here is based on Bayesian inference networks (Pearl, 1989), and this approach appears to have some advantages, but other inference techniques could be used, for example, RUBRIC's fuzzy set theory (Tong, 1985) or the Dempster-Shafer theory of evidence (Shafer, 1976).

A Bayesian inference network is a directed, acyclic dependency graph (DAG) in which nodes represent propositional variables or constants and edges represent

[1] We assume retrieval effectiveness is measured in terms of *recall* and *precision*, where recall is the proportion of relevant documents for a query that are retrieved, and precision is the proportion of retrieved documents that are relevant. Average figures are produced using a benchmark set of queries and relevance judgments.

dependence relations between propositions. If a proposition represented by a node p "causes" or implies the proposition represented by node q, we draw a directed edge from p to q. The node q contains a matrix (a *link* matrix) that specifies $P(q|p)$ for all possible values of the two variables. When a node has multiple parents, the matrix specifies the dependence of that node on the set of parents and characterizes the dependence relationship between that node and all nodes representing its potential causes. Given a set of prior probabilities for the roots of the DAG, these networks can be used to compute the probability or degree of belief associated with all 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 proposition depends on the degree to which we believe the propositions that might cause it), and a *diagnostic* component in which children provide support for their parents (if our belief in a proposition increases or decreases, so does our belief in its potential causes). The propagation of probabilities through the net can be done using information passed between adjacent nodes.

2.2 A Network for Text Retrieval

The basic retrieval inference network, shown in figure 1, consists of two component networks: a document network and a query network. The document network represents the document collection using multiple document representation schemes. The document network is built once for a collection and its structure does not change during query processing. The query network consists of a single node which represents the user's information need and one or more query representations which express that information need. A query network is built for each information need and is modified during query processing as existing queries are refined or new queries are added in an attempt to better characterize the information need. The document and query networks are joined by links between representation concepts and query concepts. All nodes in the inference network are binary-valued.

2.2.1 Document network

The document network consists of document nodes (d_i's), text representation nodes (t_j's), and concept representation nodes (r_k's). Each document node represents an actual document in the collection and corresponds to the event that a specific

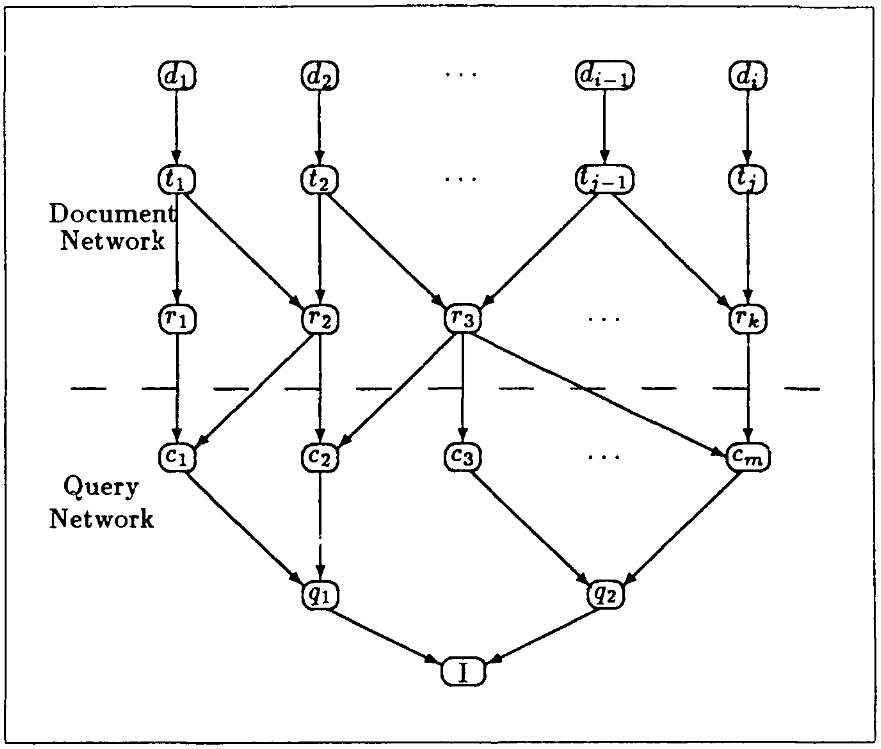

Figure 1: Basic document inference network

document has been observed.

Document nodes correspond to abstract documents rather than their physical representations. A text representation node or text node corresponds to a specific text representation of a document. A text node corresponds to the event that a text representation has been observed.

The content representation nodes or representation nodes can be divided into several subsets, each corresponding to a single representation technique that has been applied to the document texts. For example, if a collection has been indexed using automatic phrase extraction and manually assigned index terms, then the set of representation nodes will consist of two distinct subsets with disjoint domains. Thus, if the phrase "information retrieval" has been extracted and "information retrieval" has been manually assigned as an index term, then two representation nodes with distinct meanings will be created. One corresponds to the event that "information retrieval" has been automatically extracted from a subset of the collection, the second corresponds to the event that "information retrieval" has been manually assigned to a (presumably distinct) subset of the collection. We represent the assignment of a specific representation concept to a document by a directed arc to the representation node from each text node corresponding to a document to which the concept has been assigned.

In principle, the number of representation schemes is unlimited. In addition to phrase extraction and manually assigned terms we will use representations based on natural language processing and automatic keyword extraction. For any real document collection, however, the number of representations used will be fixed and relatively small.

Each document node has a prior probability associated with it that describes the probability of observing that document; this prior probability will generally be set to 1/(collection size) and will be small for typical collections. Each text node contains a specification of its dependence upon its parent; by assumption, this dependence is complete, a text node is observed ($t_i = true$) exactly when its parent document is observed ($d_i = true$).

2.2.2 Query network

The query network is an "inverted" DAG with a single leaf that corresponds to the event that an information need is met and multiple roots that correspond to the concepts that express the information need. As shown in figure 1, a set of intermediate query nodes may be used in cases where multiple queries are used to express the information need. These nodes are a representation convenience; it is always possible to eliminate them by increasing the complexity of the distribution specified at the node representing the information need.

In general, the user's information need is internal to the user and is not precisely understood. We attempt to make the meaning of an information need explicit by expressing it in the form of one or more queries that have a formal interpretation. It is unlikely that any of these queries will correspond precisely to the information need, but some will better characterize the information need than others and several query specifications taken together may be a better representation than any of the individual queries.

The roots of the query network are the primitive concepts used to express the information need. A single query concept node may have several representation concept nodes as parents. A query concept node contains a specification of the probabilistic dependence of the query concept on its set of parent representation concepts. The query concept nodes define the mapping between the concepts used to represent the document collection and the concepts comprising the queries. In the simplest case, the query concepts are constrained to be the same as the representation concepts and each query concept has exactly one parent representation node.

In a slightly more complex example, the query concept "information retrieval" may have as parents both the node corresponding to "information retrieval" as a phrase and the node corresponding to "information retrieval" as a manually assigned term. As we add new forms of content representation to the document network and allow the use of query concepts that do not explicitly appear in any document representation, the number of parents associated with a single query concept will increase. In many ways, a query concept is similar to a representation concept that is derived from other representation concepts and in some cases it will be useful to "promote" a query concept to a representation concept. For example, suppose that a researcher is looking for information on a recently developed process that is not explicitly identified in any existing representation scheme. The researcher is sufficiently motivated, however, to work with the retrieval system to describe how this new concept might be inferred from other representation concepts. If this new concept definition is of general interest, it can be added to the collection of representation concepts.

The attachment of the query concept nodes to the document network has no effect on the structure of the document network. None of the existing links need change and none of the conditional probability specifications stored in the nodes are modified.

A query node represents a distinct query form and corresponds to the event that the query is satisfied. Each query node contains a specification of the dependence of the query on the query concepts comprising it. It is worth noting that the form of the link matrix is largely determined by the type of query.

The single leaf representing the information need corresponds to thc event that an information need is met. In general, we cannot predict with certainty whether a user's information need will be met by an arbitrary document collection. The query network is intended to capture the way in which meeting the user's information need depends on documents and their representations. Moreover, the query network is intended to allow us to combine information from multiple document representations and to combine queries of different types to form a single, formally justified estimate of the probability that the user's information need is met. If the inference network correctly characterizes the dependence of the information need on the collection, the computed probability provides a good estimate.

2.3 Use of the inference network

The inference network we have described is intended to capture all of the significant probabilistic dependencies among the variables represented by nodes in the document and query networks. Given the prior probabilities associated with the documents (roots) and the conditional probabilities associated with the interior nodes, we can compute the posterior probability or belief associated with each node in the network. Further, if the value of any variable represented in the network becomes known we can use the network to recompute the probabilities associated with all remaining nodes based on this "evidence."

The network, taken as a whole, represents the dependence of a user's information need on the documents in a collection where the dependence is mediated by document and query representations. When the query network is first built and attached to the document network we compute the belief associated with each node in the query network. The initial value at the node representing the information need is the probability that the information need is met given that no specific document in the collection has been observed and all documents are equally likely (or unlikely). If we now observe a single document d_i and attach evidence to the network asserting $d_i = true$ we can compute a new belief for every node in the network given $d_i = true$. In particular, we can compute the probability that the information need is met given that d_i has been observed in the collection. We can now remove this evidence and instead assert that some d_j, $i \neq j$ has been observed. By repeating this process we can compute the probability that the information need is met given each document in the collection and rank the documents accordingly.

The document network is built once for a given collection. Given one or more queries, we then build a query network that attempts to characterize the dependence of the information need on the collection. If the ranking produced by the initial query network is inadequate, we must add additional information to the query network or refine its structure to better characterize the meaning of the queries. This process is quite similar to the relevance feedback mechanisms used in current retrieval models.

3 Extensions to the basic model

The basic model is limited in at least two respects. First, we have assumed that evidence about a variable establishes its value with certainty. Second, we have represented only a limited number of dependencies between variables. In this section we will see that these limitations can be removed.

3.1 Uncertain evidence and feedback

The only use of evidence in the basic model is to assert that a document representation has been observed ($d_i = true$). During query processing we assert each document true and rank documents based on the probability that the information need is met. Since we do not assert that the remaining documents are *false*, they continue to contribute to the belief that the information need is met so that, while we instantiate documents in isolation, the resulting probability is dependent upon both the instantiated document and some subset of the uninstantiated documents in the collection. In real document collections, the prior probability associated with each document is small and only a small portion of the representation concepts will bear on the information need, so the contribution of these uninstantiated documents will generally be small compared to the contribution of the instantiated document.

Evidence is attached to a node a in a Bayesian network by creating a new *evidence* node b as a child of a. This new node b then passes a likelihood vector (both components of a likelihood ratio) to a. The evidential support for a is then the product of the likelihood vectors from b and any other children. Since evidence is expressed in terms of likelihood we are not restricted to the values *true* and *false* (the vectors $(0, 1)$ and $(1, 0)$, respectively) but need only specify the likelihood of $a = true$ and $a = false$ given the evidence summarized at b. As a result, evidence can be used as a weight associated with a node. For example, if we attach confirming (disconfirming) evidence to a representation node it raises (lowers) the belief in all documents containing it, in all query concepts and queries that use it, and in the information need. The effect of the evidence is to bias the node so that the positive (negative) belief component passed to parents or children is amplified and the negative (positive) component is attenuated. If the evidence entirely confirms or disconfirms the node, then it blocks the flow of belief/evidence entirely – essentially it infinitely amplifies one component and attenuates the other so that the belief/evidence passed on is independent of the support received from parents or other children.

A side effect of using evidence in this way is that it establishes a coupling between documents containing the representation concept. When we instantiate one document, belief in all other documents containing the same representation concept will be reduced. This effect is probably not significant for a single term, but if two documents had similar indexing and all common terms had evidence attached, the coupling could be pronounced.

One potential use for this kind of weight is to implement a form of feedback. If, as a result of relevance feedback, a query, query concept, or representation concept is found to be more or less important than others, its effect on the propagation of belief through the network can be altered by attaching evidence. Frisse and Cousins (1989) use this approach to implement feedback in a hierarchy of index terms associated with a hypertext medical handbook.

In principle, the link matrix associated with a representation concept contains the probability that that concept is true given any set of parent beliefs. This probability depends both on the descriptive quality of the term and on the specific parent documents that are instantiated. In practice, we cannot store the matrix for nodes that have more than a few parents. Instead, we store the indexing weight associated with each parent, the term weight associated with the representation concept, and a function that computes the desired probability based on these weights.

In some cases we can manipulate the function used to compute the conditional probability in order to adapt the behavior of the network. This approach could be used as an alternative to evidence when implementing feedback. The two approaches are fundamentally different. When using evidence, the original probability distribution defined by the network is always maintained. Manipulating the combining function will generally alter the probability distribution. Manipulating the combining function will not generally be useful in the document network where the distribution models the statistical and semantic relations in the collection and its representation, but it may be useful in the query network where the dependence relations are much less constrained. The document network is largely fixed by the collection and our choice of representations; during query processing we attempt to build a network that "correctly" characterizes the dependence of the information need on that collection.

3.2 Additional dependencies

In the basic model, we assume that there are no dependencies between documents, between texts, between representation concepts, between query concepts, or between queries. While independence assumptions like these are not uncommon in retrieval models, it is widely recognized that the assumptions are unrealistic; there are a number of both statistical and logical dependencies between representation concepts and between documents. In particular, we would like to incorporate term and document clustering and would like to represent citation links between documents and thesaurus relationships between terms.

The basic mechanism for representing these dependencies is unchanged, we identify the set of nodes upon which a given node depends and characterize the probability associated with each node conditioned on its immediate parents. When adding these new links, however, we must be careful to preserve the acyclic nature of the inference network. Bayesian inference networks cannot represent cyclic dependencies, in effect evidence attached to any node in the cycle would continually propagate through the network and repeatedly reinforce the original node. In the basic model, no cycles are possible since nodes are only linked to node types that are lower in the DAG. The introduction of these "horizontal" dependencies makes cycles possible.

Document and term clustering. A variety of clustering techniques have been developed for information retrieval (van Rijsbergen, 1979). These may be loosely categorized as *document* clustering techniques which attempt to divide the collection into (possibly overlapping) subsets which are similar and *term* clustering techniques which attempt to identify subsets of representation concepts with similar usage or meaning. Clustering techniques differ widely in the document or term attributes considered, the definition of a similarity or dissimilarity measure, and the structure of the resulting classification. Term clustering techniques represent one kind of automatically-built thesaurus in which terms contained in a cluster are, in some sense, synonymous; clusters may be organized in a hierarchy to represent broader and narrower classifications. Representation of these thesaurus-like relationships will be discussed shortly.

Document clustering techniques are generally used to find documents that are similar to a document that is believed relevant under the assumption that similar documents are related to the same queries. Our use of cluster information is somewhat different since we do not retrieve clusters, but we can incorporate the cluster information in the dependence relationships between document texts and representation concepts. In the fragment shown in figure 2, document texts t_1, t_2, and t_3 are indexed using representation concepts r_1, r_2, r_3, and r_4. Documents t_2 and t_3 have been identified as part of cluster c_1; both texts are linked to a cluster node and the cluster node is linked to the representation concepts that define the cluster. The cluster node is similar to a conventional cluster representative. Documents t_1 and t_2 are indexed by the same representation concepts (r_1 and r_2) and, if

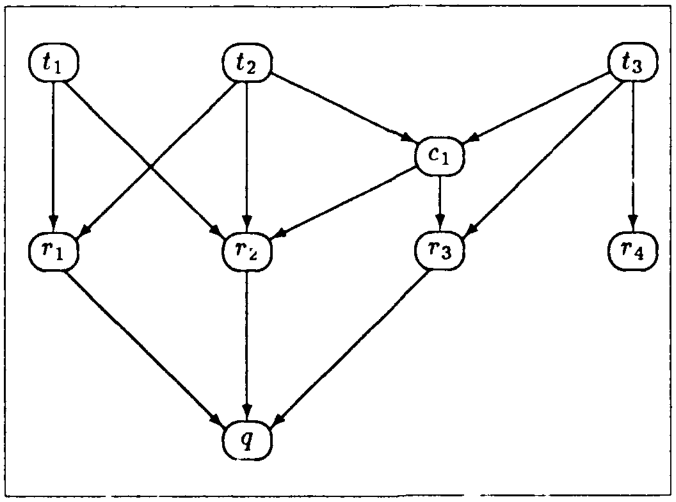

Figure 2: Document clustering model

we assume equivalent conditional probabilities, would be ranked equivalently in the absence of the cluster node. With the addition of the cluster node, however, a new representation concept (r_3) is associated with t_2 by virtue of its cluster membership. Assuming that r_3 contributes positively to the belief in q, t_2 would be ranked higher than t_1. Like query nodes, cluster nodes are a representation convenience, it is always possible to eliminate them by increasing the complexity of the distribution specified at the representation concept nodes.

Citation and nearest neighbor links. A variety of asymmetric relationships between pairs of documents can also be represented. These relationships are similar to clustering in that they use an assumed similarity between documents to expand the set of representation concepts that can be plausibly associated with a text. They differ in that they are ordered relations defined on pairs of documents rather than an unordered, set membership relationship between documents and clusters.

Perhaps the best example of this kind of relationship is the nearest neighbor link in which a document is linked to the document judged to be most similar to the original document. In figure 3 the set of representation concepts associated with document t_1 is expanded by virtue of its nearest neighbor link to document t_2. Note that it is not possible to simultaneously represent t_2 as t_1's nearest neighbor and t_1 as t_2's nearest neighbor since the pair of links would induce a cycle. A second kind of ordered link is based on citations occurring in the text. Citation links may be useful if the type of reference can be determined (e.g., citing a similar work, a peripherally related work, or a work presenting an opposing viewpoint) to allow

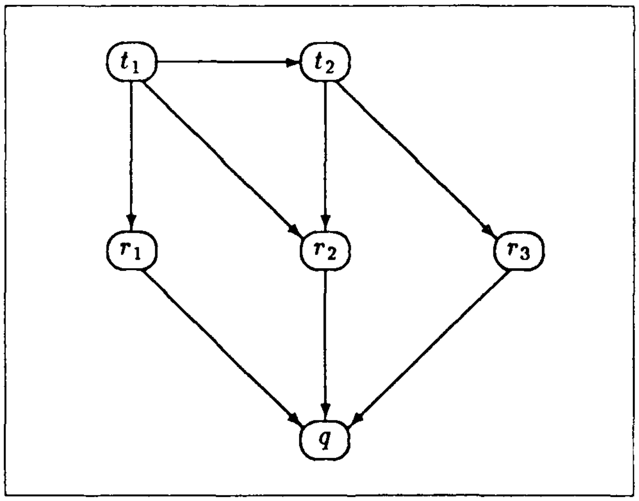

Figure 3: Nearest neighbor link

estimation of the probabilistic dependence between the nodes.

Thesaurus relationships. The structure of these networks provides a natural mechanism to represent probabilistic dependencies between the concepts or terms that describe documents and information needs. These relationships are similar to conventional thesaurus relationships, but include more information. For example, a conventional thesaurus might list "house pet" as a broader term for "dog" and "cat"; the network representation will include a specification of the probability that "house pet" should be assigned given a document containing "dog" or "cat" in isolation, neither term, or both terms.

Synonyms, related terms, and broader terms can be represented by creating new nodes to represent the synonym or related term class or the broader term and adding the new node as a child to the relevant representation concept node. We will generally prefer to add these nodes as part of the query network since their presence in the document network would represent a computational burden even when not used in a query. Although generally less useful, narrower term relationships can also be represented.

4 Other Results

4.1 Development of the Retrieval Model

As indicated in the previous section, we need to extend the basic model to include dependency information from document clustering, term clustering, hypertext links, and knowledge bases. Rather than using a single modification to accommodate all of these changes, the impact of each form of information on the networks must be considered. Nearest neighbor links, for example, can introduce cycles into the network, but in the case of index terms, this information can be expressed in the form of a dependency tree (Van Rijsbergen, 1979). Hypertext links could be integrated as dependency links between text nodes in the network, but it may be more appropriate to introduce them as evidence attached to the linked nodes. Relevance feedback can also be incorporated in a variety of ways. We have carried out experiments with nearest neighbor links that indicate that additional dependencies can improve effectiveness without a prohibitive efficiency cost.

4.2 Building and Searching Networks

Bayesian inference networks can be computationally expensive to build and maintain. Based on our experience building informal dependency networks in IR, we have developed efficient algorithms for using Bayesian networks in an IR setting. It appears to be possible, for example, to avoid cycles when building a network for IR and we would therefore not have to use an algorithm for breaking cycles.

In the case of the basic form of the networks, and potentially also for the extended form, it is possible to build an inverted file that contains the beliefs generated for each representation concept by each document. This allows very efficient processing of queries. A query network is specified through a user interface and then a ranked list of documents is generated using the inverted file, similar to most current text retrieval systems.

Our analysis shows that networks can be built in $O(t \log t)$ time where t is the number of term occurrences in the collection. Average query processing time is less than one second for our test queries and this time should grow logarithmically with collection size. The network files are roughly twice as large as the original source collection text and will exhibit linear or slightly sublinear growth.

We have tested our programs with networks for the smaller test collections (consisting of 2,000 to 3,000 documents) and have recently modified them to handle databases of hundreds of megabytes.

4.3 Retrieval Experiments

The main results of the retrieval experiments were:

- The basic inference network model offers substantial improvements in retrieval performance compared to the best conventional retrieval models (up to 25% increase in average precision).

- The network interpretation of Boolean queries performs substantially better than conventional Boolean (up to 65% improvement in average precision).

- The use of multiple document representations leads to small performance improvements (8% increase in average precision), but our test collections were not suitable for testing this aspect of the model thoroughly.

- Multiple representations of the information need (queries) significantly improve retrieval performance (20% increase in precision).

- Links between documents can significantly improve retrieval performance.

4.4 Representation based on Natural Language Processing

We expect that one of the major sources of evidence for the retrieval process will come from using natural language processing (NLP) techniques on the text of the documents and queries. Previous experiments with NLP techniques used for text retrieval have not been successful, but we believe that is because the representations produced were not being used appropriately in retrieval strategies. The inference network approach appears to provide the right framework for incorporating evidence from NLP and using it to improve the performance of the (surprisingly effective) word-based representations.

We have been concentrating on the use of word senses and phrases as additional forms of text representations. Our experiments have shown that ambiguity is not a serious problem in text representation, but the correct use of phrases in the inference net model can lead to significant improvements (more than 10% increase in average precision).

5 Summary

In this project, we have described a new approach to retrieval that is based on inference networks. This approach has the potential of significantly improving the performance of information systems that deal with complex and uncertain information, particularly text-based systems. We have carried out experiments that show that retrieval based on inference nets can be very effective and can be done with large databases. We have also identified natural language processing techniques appropriate for the retrieval application and used them as additional sources of evidence. This research addresses a number of fundamental questions about the nature of information retrieval and the effectiveness of natural language processing and domain knowledge for retrieval.

References

[1] Belew, R.K. "Adaptive Information Retrieval: Using a Connectionist Representation to Retrieve and Learn about Documents", Proceedings of ACM SIGIR 89, 1989.

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

[3] Croft, W.B., 1987. "A framework for office document retrieval" *Proceedings of the 2nd IEEE Conference on Office Automation*, 26-32.

[4] Croft, W.B.; Lewis, D., 1987. "An Approach to Natural Language Processing for Document Retrieval", *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 26-32, New Orleans, 1987.

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

[6] Croft, W.B.; Lucia, T.J.; Cringean, J.; Willett, P., 1989. "Retrieving Documents by Plausible Inference: An Experimental Study". *Information Processing and Management*, 25, 599-614, 1989.

[7] W.B. Croft and H. Turtle, "A Retrieval Model Incorporating Hypertext Links", Proceedings of Hypertext 89, 213-224, (1989).

[8] W.B. Croft, R. Krovetz, and H. Turtle, "Interactive Retrieval of Complex Documents", *Information Processing and Management*, 26(5), 593-613, (1990).

[9] Date, C.J., 1986. *An Introduction to Database Systems*, Vol.1, Addison Wesley.

[10] Frisse, M. and Cousins, S.B. "Information Retrieval from Hypertext: Update on the Dynamic Medical Handbook Project", Proceedings of Hypertext 89, 1989.

[11] Gallaire, H.; Minker, J.; Nicols, J., 1984. "Logic and Databases: A Deductive Approach", *ACM Computer Surveys*, 16, 153-186.

[12] L. Gay, W.B. Croft, "Interpreting Nominal Compounds for Information Retrieval", *Information Processing and Management*, 26(1), 21-38, (1990).

[13] Katzer, J.; McGill, M.; Tessier, J.A.; Frakes, W.; Dasgupta, P., 1982. "A Study of the Overlap among Document Representations", *Information Technology*, 1, 261-274.

[14] R. Krovetz and W.B. Croft, "Word Sense Disambiguation Using a Machine-Readable Dictionary", Proceedings of the 12th International Conference on Research and Development in Information Retrieval, 127-136, (1989).

[15] R. Krovetz and W.B. Croft, "Lexical Ambiguity and Information Retrieval", *ACM Transactions on Information Ssystems*, (to appear).

[16] R. Krovetz. "Information Retrieval and Lexical Ambiguity", Proceedings of AAAI Symposium on Text-Based Intelligent Systems, 1990.

[17] R. Krovetz, "Lexical Acquisition and Information Retrieval", in *Lexical Acquisition: Using On-Line Resources to Build a Lexicon*, Uri Zernick (ed), LEA Press, (to appear).

[18] Lenat, Doug; Prakash, Mayank; and Shepherd, Mary. CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks. *AI Magazine,* 6(4), 65-85.

[19] Lewis, D.; Croft, W.B.; Bhandaru, N., 1989. "Language Oriented Information Retrieval", *International Journal of Intelligent Systems*, 4, 285-318, 1989.

[20] Lewis. D. "Text Representation for Text Classification", Proceedings of AAAI Symposium on Text-Based Intelligent Systems, 1990.

[21] D. Lewis and W.B. Croft, "Term Clustering of Syntactic Phrases", Proceedings of the 13th International Conference on Research and Development in Information Retrieval, 385-404, (1990).

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

[23] Van Rijsbergen, C. J., 1979. *Information Retrieval.* Second Edition. Butterworths, London.

[24] Van Rijsbergen, C.J., 1986. "A Non-Classical Logic for Information Retrieval". *Computer Journal*, 29, 481-485.

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

[26] Shafer, G. *A Mathematical Theory of Evidence.* Princeton University Press, 1976.

[27] Tong, R.M. and Shapiro, D. "Experimental Investigations of Uncertainty in a Rule-Based System for Information Retrieval", *International Journal for Man-Machine Studies*, 22, 265-282, 1985.

[28] H. Turtle and W.B. Croft, "Inference Networks for Document Retrieval", Proceedings of the 13th International Conference on Research and Development in Information Retrieval, 1-24, (1990).

[29] H.Turtle and W.B. Croft, "Efficient probabilistic inference for text retrieval", Proceedings of RIAO 3, (to appear).

[30] H. Turtle, *Inference Networks for Document Retrieval*, Ph.D. thesis, Computer and Information Science Department, University of Massachusetts, COINS TR 90-92, (1990).

[31] H. Turtle and W.B. Croft, "Evaluation of an Inference Network-Based Retrieval Model", *ACM Transactions on Information Systems*, (to appear).

[32] Weyer, S.A. and Borning A.H., 1985. "A Prototype Electronic Encyclopedia". *ACM Transactions on Office Information Systems*, 3, 2-21.