Int. J. Man–Machine Studies (1983) **19**, 45–56

Fuzzy sets and generalized Boolean retrieval systems

DONALD H. KRAFT AND DUNCAN A. BUELL

Department of Computer Science, Louisiana State University, Baton Rouge, Lousiana 70803, U.S.A.

Substantial work has been done on the application of fuzzy subset theory to information retrieval. Boolean query processing has been generalized to allow for weights to be attached to individual terms, in either the document indexing or the query representation, or both. Problems with the generalized Boolean lattice structure have been noted, and an alternative approach using query thresholds and appropriate document evaluation functions has been suggested.

Problems remain unsolved, however. Criteria generated for the query processing mechanism are inconsistent. The exact functional form and appropriate parameters for the query processing mechanism must be specified. Moreover, the generalized Boolean query model must be reconciled with the vector space approach, suggested new lattice structures for weighted retrieval, and probabilistic retrieval models. Finally, proper retrieval evaluation mechanisms reflecting the fuzzy nature of retrieval are needed.

1. Introduction

We can view information retrieval systems in many ways. Figure 1, taken from Deutsch & Kraft (1974), illustrates our point of view. Documents are identified, acquired, indexed, and stored. The indexing represents a determination of bibliographic information (author, title, etc.) and an assignment of index terms (words or phrases) to specify

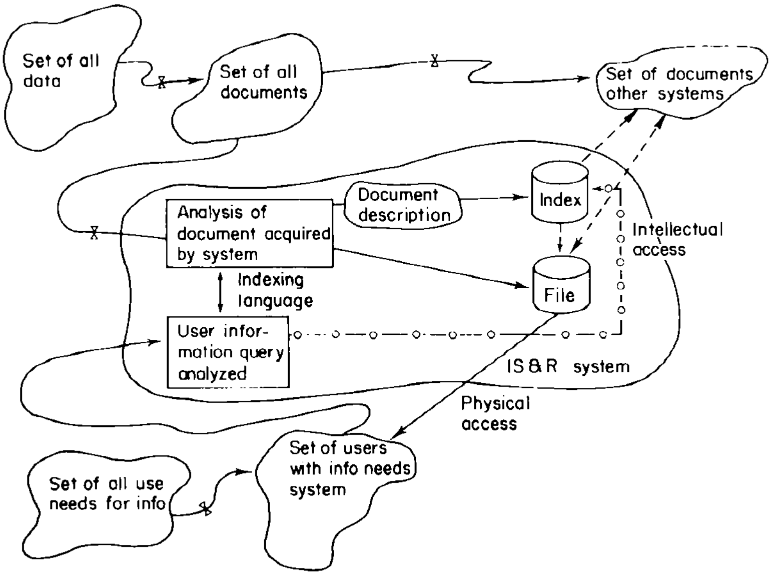
Set of all data
Set of all documents
Set of documents other systems
Document description
Index
Analysis of document acquired by system
Intellectual access
Indexing language
File
User information query analyzed
IS&R system
Physical access
Set of users with info needs system
Set of all use needs for info

FIG 1. Model of an information storage and retrieval system.

0020-7373/83/070045 + 12$03.00/0

the nature of the subject content of the document. Users with queries specifying a need for such documents are analyzed and assigned terms in a similar manner. The index is then scanned to determine which documents are to be deemed relevant to a given user query; this is called intellectual access. Physical access is concerned with the actual presentation of desired documents to a user.

The problem can be modeled as a statistical decision theory problem (Bookstein & Cooper, 1976; Knorz, 1982; Kraft, 1973; Salton & Yu, 1982). However, the basic assumptions of the model include the notion of a document's being either totally relevant or totally nonrelevant. It is not possible in this framework to consider a document's being partially relevant. Some researchers have incorporated a probabilistic approach (Cooper & Huizinga, 1982; Maron & Kuhns, 1960; Robertson, Maron & Cooper, 1982), while others have suggested simple weighting schemes (Angione, 1975; Bookstein & Cooper, 1976; Doszkocs, 1982; Koll, 1978; Noreault, Koll & McGill, 1977; Salton, 1975*b*; Salton & Waldenstein, 1978; Salton & Wu, 1980; Wu & Salton, 1981). Most of the probablistic approaches are concerned with the indexing problem, which is concerned with the assignment of index terms to documents, rather than the related problem of which documents to retrieve in response to a query.

Bookstein & Cooper (1976) show that there are four basic elements in an information storage and retrieval system:

1. a set of records (surrogates for documents), containing bibliographic information (author, title, publisher, etc.), locational information (classification number, shelf or library indicator), and index terms;
2. a set of user query descriptions, including index terms and relationships among those terms;
3. a set of retrieval status values, a simply ordered set of values in a given interval, any of which can indicate the relevance of a document to a query; and
4. a retrieval function that maps a given record and given query into a retrieval status value.

The main ingredient in any document retrieval system is the indexing language, the language allowing the system to communicate between document and query representations. This indexing language has three parts, which are:

(1) syntax, a set of rules for generating appropriate index terms (the vocabulary) from the natural language in which author, user, and information retriever communicate;

(2) semantics, a set of rules for specifying term meanings for application as part of query and document representations; and

(3) structure, a set of rules to indicate term relationships.

This issue of term relationships is a crucial one, for it is an on-going problem for retrieval systems. There are several classes of related terms, including:

(1) synonyms, terms whose meaning is essentially identical (e.g. descriptions of subsets as crisp, ordinary, or nonfuzzy);

(2) near-synonyms, terms that are similar but for which there are shades of difference in meaning (e.g. operations research and management science);

(3) related terms, terms that are somewhat similar but for which there is substantial difference in meaning (e.g. computer science and data processing); and

(4) hierarchical terms, terms that are related in that one (narrower) term describes a concept that is, more-or-less, a subset of a concept described by a second (broader) term (e.g. consider the hierarchy as indicated by the terms: animal, mammal, canine, dog, collie, Lassie).

These relationships are a sticky problem for many reasons. For example, the relationship between computer engineering and computer science is often a matter of campus politics as well as taxonomy. The reader may decide the relationships among the terms mathematics, set theory, fuzzy subset theory, and multi-valued logic. Moreover, the issues of how to assign index terms to an intellectual concept is often a problem (is optimization to be called "mathematical programming"?). Finally, it is not always easy to decide where a user wishes to be in the hierarchy (for a query on abdominal surgery, a user may sometimes want to expand the search by looking for narrower terms such as stomach surgery, and at other times may wish to see only documents on this general topic without the specifics of any organ).

2. Fuzzy retrieval

It is obvious that the concept of a fuzzy subset can be applied to the document retrieval situation. Many researchers have considered exactly this problem (Negoita, 1973*a*, *b*, *c*, *d*; Negoita & Flondor, 1976; Radecki, 1976, 1977, 1979; Sachs, 1976; Tahani, 1976; Yager, 1980), although this application of fuzzy subset theory has not been without its critics (Bookstein, 1978; Robertson, 1978).

Suppose we have

$$\begin{aligned} \mathrm{D} &= \text{set of documents,} \\ \mathrm{T} &= \text{set of index terms, and} \\ \mathrm{F} &= \text{the membership function.} \end{aligned}$$

Then

$$\mathrm{F}: \mathrm{D} \times \mathrm{T} \rightarrow [0, 1].$$

This means that an indexer can analyze a given document and assign terms with a weight in the interval [0, 1]. Each weight can be viewed as a membership function, indicating the extent to which the document in question is in the subset of documents in the total collection (U, the universal set) that are "about" the concept represented by the term in question. This differs from, but is not in contradiction with, Maron's view of "aboutness" (Maron, 1977). If all weights are either zero or one, we have the traditional (Boolean) indexing; on the other hand, if some weights are in the open interval (0, 1), then we have fuzzy indexing.

The probabilistic approach to retrieval is concerned with estimating the probability of relevance of a given document to a query. Often this is done on an individual term basis. The probability of a document's being relevant, given that term *t* in the query has been assigned to the document, is still, however, a determination of relevance with Boolean logic. Either a document is totally relevant or totally nonrelevant. The probabilist merely allows for the concept of relevance to be nondeterministic, so that one can assign a likelihood measure to it for a specific situation. The fuzzy model views relevance as a concept that is not well-defined, so that indexing can reflect shades of "aboutness" or relevance. In some commercial bibliographic databases, indexers are allowed to give some terms an asterisk(*), indicating a stronger sense of "about" than for a term without an asterisk. Thus, a term can be left off a document, included on a document without an asterisk, or included with an asterisk. This multi-valued logic is the beginning of a true application of fuzzy subsets to retrieval.

We now have

$$\begin{aligned} m_F(t) &= \{<d, F(d,t)> \mid d \text{ in D}\} \\ &= \text{meaning of term } t, t \text{ in T, and} \\ \mathrm{M}_F(t) &= \{m_F(t) \mid t \text{ in T}\} \\ &= \text{set of term meanings for term } t. \end{aligned}$$

For the Boolean operators, we shall use the normal (Bellman & Zadeh, 1970) functions;

$$\begin{aligned} m_F(t \text{ AND } t') &= \{\langle d, \min[\mathrm{F}(d,t), \mathrm{F}(d,t')]\rangle \mid d \text{ in D}\}, \\ m_F(t \text{ OR } t') &= \{\langle d, \max[\mathrm{F}(d,t), \mathrm{F}(d,t')]\rangle \mid d \text{ in D}\}, \\ m_F(\text{NOT } t) &= \{\langle d, 1-\mathrm{F}(d,t)\rangle \mid d \text{ in D}\}. \end{aligned}$$

This preserves the Boolean lattice properties of idempotence, commutativity, associativity, distributivity, absorption, intersection ((U AND A) = A), union ((NULL OR A) = A), involution (NOT(NOT A) = A, NOT U = NULL), and DeMorgan's laws. However, the complementarity rules are not valid; (A AND NOT A) does not generally equal NULL, and (A OR NOT A) does not generally equal U. Note that U implies a fuzzy subset with all membership function values equal to one, NULL implies a fuzzy subset with all membership function values equal to zero, and A is any fuzzy subset generated by the membership functions for a given term.

It is essential that the user be allowed to generalize from single-term queries. Thus, we consider Boolean connectives in the light of the vector lattice described above. Letting T* represent the set of all well-formed Boolean expressions from T, this yields

$$\mathrm{M}_F^*(t) = \{m_F(t) \mid t \text{ in T}^*\}.$$

Buell (1982) discusses this at length, and states that if the image of F is {0, 1}, then we have a Boolean lattice.

3. Generalized queries

Consider the problem of satisfying a query for good mystery novels. Suppose that the user decides that "good" implies either plenty of sex or violence plus a suspenseful plot. All of the above are fuzzy so fuzzy indexing seems reasonable in terms of how much sex is "plenty" or how "suspenseful" is "suspenseful". In fact, even the idea of a novel could be fuzzy (say, versus a novella). However, suppose that the idea of sex or violence was less important than that of suspense in describing this specific user's interests. Perhaps the query could be "weighted" to reflect this inequality.

Thus we see that the generalization of traditional Boolean query processing is more complex than merely fuzzifying the indexing. The query representation can also be "weighted". Moreover, the document relevance evaluation mechanisms can yield other than zero–one retrieval status values. This leads us to four levels of generalizations.

1. Boolean (zero–one) indexing and Boolean queries with non-Boolean retrieval status values (Bookstein & Cooper, 1976; Salton, 1975*a*, 1979).
2. Fuzzy (non zero–one) indexing and Boolean queries with retrieval status values computed using fuzzy subset rules (Maron & Kuhns, 1960; Salton, 1975*a*).
3. Boolean indexing and fuzzy queries with retrieval status values either seen as equivalent to traditional Boolean query processing rules or used to rank output of traditional processing output (Angione, 1975; Koll, 1978; Noreault *et al.*, 1977).
4. Fuzzy indexing and fuzzy queries (weights or thresholds on terms and/or on Boolean sub-expressions) with the retrieval status value being calculated by some general function (Bookstein, 1980*a*; Buell, 1982; Buell & Kraft, 1981*c*, *d*).

The last generalization, the most general, raises some interesting issues. In essence, we have the following:

$$a: \mathrm{T} \rightarrow [0, 1] = \text{term ``weight''},$$

and

$$g(\mathrm{F}, a): [0, 1] \times [0, 1] \rightarrow [0, 1],$$

where a is a function generating a "weight" for a term and g is the document evaluation in terms of its "aboutness" regarding the concepts represented by the term t in question.

Now we must consider the form of g, and determine how g will be employed in Boolean queries. We suggest that the MIN, MAX, and One-Minus operators be used for AND, OR, and NOT, respectively. For g, there are several possibilities, including

$$g = \mathrm{F}{*}a \quad \text{and} \quad g = \mathrm{F}{**}a.$$

However, there are problems with these forms for g. For example, suppose that we are to AND two terms t and with weights $a = 0{\cdot}8$ and $a' = 0{\cdot}0001$, and we use the multiplication for g and use MIN for the AND. Then, the term t' which has a low weight indicating that we do not much care if the document is or is not about the concepts of term t', dominating the MIN, producing just the opposite effect. Moreover, the exponential form for g produces problems as F and a both approach zero.

Thus, Waller & Kraft (1979) generated a list of criteria for the function g. This list is reproduced in Appendix A. The problem with the AND as described above is that g must be a non-increasing function of a. Bookstein (1980*a*) suggests an alternative function for g, where $g = \mathrm{F}{*}a$ unless the term is to be ANDed, $g = \mathrm{F}/a$ (rounded down to one if $\mathrm{F}/a > 1$) otherwise. However, this violates separability, the Waller–Kraft criterion that a document is to be evaluated first along each term separately, then combined via the Boolean logic of the query. The importance of separability is demonstrated by Buell (1982), who also shows that the Waller–Kraft criteria (what Buell calls a "wish list") cannot be simultaneously satisfied.

4. Thresholds

Buell & Kraft (1981*d*) have suggested a different approach to attempt to resolve this dilemma. They suggest that the variable a represent a threshold, rather than a weight.

Thus, a term t with a threshold in a query implies that a document is evaluated by considering whether or not that document's membership function F for term t exceeds that threshold. Let P and Q be two functions of the threshold a that give g its form. This leads to the following:

$$\begin{aligned} g &= \mathrm{P}*(\mathrm{F}/a) && \text{if } \mathrm{F}>a, \\ g &= \mathrm{P}+\mathrm{Q}*(\mathrm{F}-a)/(1-a) && \text{if } \mathrm{F}<a. \end{aligned}$$

Buell & Kraft generate several criteria analogous to the Waller–Kraft criteria for the forms that P and Q must take. One example of a proper threshold function is $\mathrm{P}=(1+a)/2$ and $\mathrm{Q}=a(1-a)/2$.

These criteria, which specify necessary conditions for a proper document evaluation mechanism, are:

1. $\mathrm{dP}/\mathrm{d}a>0$ (g increases with a for constant F/a if $\mathrm{F}<a$),
2. $\mathrm{dP}/\mathrm{d}a+\mathrm{dQ}/\mathrm{d}a>0$ (g increases with a for constant $(\mathrm{F}-a)/(1-a)$ if $\mathrm{F}>a$),
3. $\mathrm{d}g/\mathrm{d}a>0$ for constant F,
4. $\mathrm{d}g/\mathrm{dF}>0$ for constant $a>0$
5. $g=0$ for $\mathrm{F}=a=0$, (boundary condition),
6. $g=1$ for $\mathrm{F}=a=1$, (boundary condition), and
7. $\lim_{\mathrm{F}\to 0} g$ $=\mathrm{P(O)}>0$.

The function form implies that one is given some partial credit (F/a) for the membership function's coming close but not exceeding the threshold p. This credit is weighted by an increasing function of the threshold. This implies that as the threshold increases, a given percentage of partial is given more weight. Moreover, partial credit is given for exceeding the threshold, if the excess is not by the maximum amount, with this credit again being weighted by an increasing function of the threshold so that the partial credit is greater for a higher threshold.

5. Other approaches

The fuzzy subset approach is an interesting one that can potentially yield many interesting and important results for use in information retrieval systems. Buell & Kraft (1982) have begun to produce a prototype retrieval system, known as LIARS, that does fuzzy set retrieval. However, other approaches exist and they should be examined to see their relation to the fuzzy model.

Cooper & Maron (1978) and Roberson *et al.* (1982) are concerned with a probabilistic model. The relationship of this model to the fuzzy subset model has already been discussed in this paper. Cooper & Maron (1978) are more concerned with a utility theory approach to indexing and retrieval, looking at expected utility in a model closely related to the probability model. Moreover, Cooper & Huizinga (1982) use relevance probabilities as query weights in a non-Boolean, discrete indexing model of retrieval. They generate these weights via the maximum entropy principle, which involves the joint distribution of term occurrences on relevant documents.

Salton has done considerable work (Salton, 1975*a*, *b*; Salton & Waldenstein, 1978; Salton & Wong, 1978; Salton & Wu, 1980; Salton & Yu, 1982; Wu & Salton, 1981;

Yu & Salton, 1971) on a vector space model of retrieval. Here, index terms are weighted but queries are not Boolean and are usually discrete. Documents in the vicinity (we avoid the use of the term "neighborhood" as the requirements of the topological definition for that term are not always satisfied) of a query are clustered hierarchically and retrieved. Usually, the documents are represented as a vector of term weights which exist in the unit hypercube. The unweighted queries are represented as corner points on the same hypercube. "Distance" measures are calculated to generate a list of documents "near" the query in this vector space. Often, these measures consist of the dot product of the query vector with a document vector normalized by division. The possible normalization mechanisms include Dice's coefficient, Jaccard's coefficient, the overlap coefficient, and the cosine coefficient. The latter one is used by Salton, and normalizes the space to the unit hypersphere.

Salton's method does not allow for Boolean logic in the query structure, which is lacking in general in the works of those using the vector space approach. This means that the terms are implicitly ANDed together, which is very limiting. Moreover, the cosine law does not take into account the magnitude of the vectors, only the angle between the query vector and a document vector. This makes sense for Salton's model, since, as stated previously, vectors are normalized to the hypersphere. However, this ignores "weighted" queries which might be represented as vectors inside the unit hypercube. Moreover, the differences between situations where the query vector has greater magnitude than the vector representing the document in question and situations where the reverse is true have been suppressed. These differences are precisely what the fuzzy threshold approach emphasizes, while still allowing for Boolean logic in the query.

Salton (Salton 1975*a*, *b* Salton & Waldenstein, 1978; Salton & Wu, 1980; Wu & Salton, 1981; Yu & Salton, 1971) suggests that the index term weights could be generated by looking at the relative frequency with which each term is used in the title and/or abstract of a document. The possible use of stemming and synonyms is also considered and tested. Moreover, the query weights, if used, can be modified dynamically by having the user examine a tentative list of retrieved documents and test each one for relevance. Clearly, terms frequently on relevant documents should be given more weight, and terms frequently on nonrelevant documents should be given less weight. Thus, the classifier is trained, to put things in the terminology of pattern recognition.

However, this model still considers relevance as a discrete, rather than a fuzzy, phenomenon. Also, it is not evident that the frequency calculations, or the probabilities of Maron, Cooper, etc., could not be used as the fuzzy membership functions and/or the weights or thresholds for the Buell–Kraft model. The implications of doing this still must be fully explored.

Kantor (1981) suggests that the problems of fuzzy membership functions and "weights" on query terms are due to the lattice structure itself. He proposes a new structure, in which the function g becomes

$$g = a*\mathrm{F} + (1-a)*\mathrm{V},$$

where V is the membership function for some special element. Kantor states that this membership function value V must be 1/2 for all documents.

6. Performance measures

The output of a fuzzy retrieval system will be a list of all of the documents ranked according to their relevance evaluations. It is certainly possible to use a cut-off, either on the number of documents displayed to the user, or on the retrieval status value. The question arises as to the performance of such a system. How does the result compare to the results from other systems or from some hypothetical notion of perfection? Buell & Kraft (1981*a*, *b*) have considered this problem and state that the idea is to measure the extent to which a document retrieval system fulfills its goals, i.e. the extent to which the system responds to a query by retrieving the same set of documents which would be retrieved by a human expert (perhaps the user himself/herself). Current measures include precision (the fraction of retrieved documents that are relevant), recall (the fraction of relevant documents that are retrieved), combinations of recall and precision, and various functions of search length for a given number of relevant records (Kraft & Bookstein, 1978). However, these measures were designed for discrete Boolean retrieval systems.

What is needed is a fuzzy concept of retrieval. Since relevance is seen as a fuzzy concept, why, not retrieval as well? Then performance can be viewed as the similarity between the fuzzy subset of relevant documents and the fuzzy subset of retrieved documents. The retrieval status value can be viewed as a membership function for the notion of retrieval. The membership function for relevance must be specified by the human expert, but with the notion of fuzziness preserved. Then recall and precision can be calculated in terms of documents being both relevant and retrieved (relative AND retrieved, via MIN). This is still not totally proper, since the result of a fuzzy retrieval system is a ranked list of documents. What is needed is the fuzzy equivalent of Salton's (1975*a*) generalized recall and precision measures for considering rankings. The relationship between precision, recall, search length, and the ranked order of the documents, discussed for discrete system, (Kraft & Bookstein, 1978), must also be determined for fuzzy retrieval systems.

7. Summary and conclusions

We have reviewed the work that has been done on applying fuzzy subset theory to document retrieval systems. We have seen that this is a generalization of document and query representation and processing. It is accomplished by allowing non-Boolean index weights to be attached to the document and non-Boolean weights or thresholds to be attached to the individual terms in the query representation. There are problems associated with the preservation of the Boolean lattice structure when queries involve Boolean logic. Moreover, there are other retrieval models that do not involve fuzzy subsets. We have tried to show that the fuzzy model is related to these others and that it is a beginning of a true generalization. Moreover, the threshold approach seems to hold much promise and awaits testing for empirical verification of its superiority. Performance measures are being considered that are appropriate to help evaluate a generalized retrieval system.

This research was supported in part by NSF Grant IST-8115146.

References

ANGIONE, P. V. On the equivalence of Boolean and weighted searching based on the convertibility of query forms. *Journal of the American Society for Information Science*, **26**, 112–124.

BELLMAN, R. E. & ZADEH, L. A. (1970). Decision-making in a fuzzy environment. *Management Science*, **17**, B141–B164.

BILLER, H. (1982). On the architecture of a system integrating data base management and information retrieval. *Proceedings of the Fifth International Conference on Information Retrieval*, West Berlin, Germany.

BOLLMANN, P., KONRAD, E. & ZUSE, H. (1982). FAKYR: a method data base system for education and research in information retrieval. *Proceedings of the Fifth International Conference on Information Retrieval*, West Berlin, Germany.

BOOKSTEIN, A. (1978). On the perils of merging Boolean and weighted retrieval systems. *Journal of the American Society for Information Science*, **29**, 156–158.

BOOKSTEIN, A. (1980*a*). Fuzzy requests: an approach to weighted Boolean searches. *Journal of the American Society for Information Science*, **31**, 240–247.

BOOKSTEIN, A. (1980*b*). *Weighted Boolean retrieval.* Unpublished paper, Graduate Library School, University of Chicago, Chicago, Illinois.

BOOKSTEIN, A. (1980*c*). *Decision making as a retrieval problem.* Unpublished paper, Graduate Library School, University of Chicago, Chicago, Illinois.

BOOKSTEIN, A. (1982). Explanation and generalization of vector models in information retrieval. *Proceedings of the Fifth International Conference on Information Retrieval*, West Berlin, Germany.

BOOKSTEIN, A. & COOPER, W. S. (1976). A general mathematical model for information retrieval systems. *Library Quarterly*, **46**, 153–167.

BUCKLES, B. P. & PETRY F. E. (1981). A fuzzy model for relational databases. *Fuzzy Sets and Systems*, **6**.

BUELL, D. (1981). A general model of query processing in information retrieval systems. *Information Processing and Management*, **17**, 249–262.

BUELL, D. (1982). An analysis of some fuzzy subset applications to information retrieval systems. *Fuzzy Sets and Systems*, **7**, 35–42.

BUELL, D. & KRAFT, D. H. (1981*a*). Performance evaluation in a fuzzy retrieval environment. *Proceedings of the Fourth Annual International Conference on Information Retrieval*, Berkeley, California.

BUELL, D. & KRAFT, D. H. (1981*b*). Evaluation of fuzzy retrieval systems. *Proceedings of the 1981 Annual ASIS Meeting*, Washington, D. C.

BUELL, D. & KRAFT, D. H. (1981*c*). A model for a weighted retrieval system. *Journal of the American Society for Information Science*, **32**, 211–216.

BUELL, D. & KRAFT, D. H. (1981*d*). Threshold values and Boolean retrieval Systems. *Information Processing and Management*, **17**, 127–136.

BUELL, D. & KRAFT, D. H. (1982). LIARS—a software environment for testing query processing strategies. *Proceedings of the Fifth International Conference on Information Retrieval*, West Berlin, Germany.

CHAN, F. K. (1973). Document classification through the use of fuzzy relations and determination of significant features. Unpublished *Master's Thesis*, Department of Computer Science, University of Alberta, Canada.

COOPER, W. S. & HUIZINGA, P. (1982). The maximum entropy principle and its application to the design of probabilistic retrieval systems. *Information Technology: Research and Development*, **1**, 99–112.

COOPER, W. S. & MARON, M. E. (1978). Foundations of probabilistic and utility-theoretic indexing. *Journal of the Association for Computing Machinery*, **25**, 67–80.

DEUTSCH, D. & KRAFT, D. H. (1974). A study of an information retrieval performance measure: expected search length as a function of file size and organization. Presented at the *Operations Research Society of America Meeting*, Boston, Massachusetts.

DOSZKOCS, T. E. (1978). AID, an associative interactive dictionary for online searching. *Online Review*, **2**, 163–173.

DOSZKOCS, T. E. (1982). From research to application: the CITE natural language information retrieval system. *Proceedings of the Fifth International Conference on Information Retrieval*, West Berlin, Germany.

HANANI, M. Z. (1977). An optimal evaluation of Boolean expressions in an online query system. *Communications of the Association for Computing Machinery*, **20**, 344–346.

HSIAO, D. & HARARY, F. (1970). A formal system for information retrieval from files. *Journal of the Association for Computing Machinery*, **13**, 67–73.

KANTOR, P. B. (1981). The logic of weighted queries. *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-11**, 816–821.

KAUFMANN, A. (1975). *Introduction to the Theory of Fuzzy Subsets*. New York: Academic Press.

KNORZ, G. (1982). A decision theory approach to optimal automatic indexing. *Proceedings of the Fifth International Conference on Information Retrieval*, West Berlin, Germany.

KOLL, M. B. (1978) WEIRD: An approach to concept-based information retrieval. *International Conference on Information Storage and Retrieval Proceedings*, Association for Computing Machinery SIGIR.

KRAFT, D. H. (1973). A decision theory view of the information retrieval situation: an operations research approach. *Journal of the American Society for Information Science*, **24**, 368–376.

KRAFT, D. H. & BOOKSTEIN, A. (1978). Evaluation of information retrieval systems: a decision theory approach. *Journal of the American Society for Information Science*, **29**, 1–40.

KRAFT, D. H. & WALLER, W. G. (1979). Problems in modeling a weighted Boolean retrieval system. *Proceedings, American Society for Information Science Meeting*, Minneapolis, Minnesota.

LANCASTER, F. W. (1979). *Information Retrieval Systems* (2nd Edition). New York: Wiley.

MARON, M. E. (1977). On indexing, retrieval, and the meaning of about. *Journal of the American Society for Information Science*, **28**, 38–43.

MARON, M. E. & KUHNS, J. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery*, **7**, 216–244.

MCGILL, M., SMITH, L. C., DAVIDSON, S. & NOREAULT, T. (1976). "Syracuse Information Retrieval Experiment (SIRE): Design of an On-Line Bibliographic Retrieval System," *ACM SIGIR FORUM*, pp. 37–44.

MELDMAN, M. J., MCLEOD, D. J., PELLICORE, R. J. & SQUIRE, M. (1978). *RISS: A Relational Data Base Management System for Minicomputers*. New York: Van Nostrand Reinhold.

MOULINOUX, C. (1982). MESSIDOR: a distributed information retrieval system. *Proceedings of the Fifth International Conference on Information Retrieval*, West Berlin, Germany.

NAKUMURA, K. & IWAI, S. (1982). Topological fuzzy sets as a quantitative description of analogical inference and its application to question-answering systems for information retrieval. *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-12**, 193–204.

NEGOITA, C. (1973*a*). Linear and non-linear information retrieval. *Studii si Cercetari de Documentare*, 369–381.

NEGOITA, C. (1973*b*). On the decision process in information retrieval. *Studii si Cercetari de Documentare*, 369–381.

NEGOITA, C. (1973*c*). On the notion of relevance in information retrieval. *Kybernetes*, **2**, 161–165.

NEGOITA, C. (1973*d*). On the application of the fuzzy sets separation theorem for automatic classification in information retrieval systems. *Information Science*, **5**, 279–286.

NEGOITA, C. & FLONDOR, P. (1976). On fuzziness in information retrieval. *International Journal on Man–Machine Studies*, **8**, 711–716.

NOREAULT, T., KOLL, M. & MCGILL, M. J. (1977). Automatic ranked output from Boolean searches in SIRE. *Journal of the American Society for Information Science*, **28**, 333–340.

RADECKI, T. (1976). Mathematical model of information retrieval systems based on the concept of fuzzy thesaurus. *Information Processing and Management*, **12**, 313–318.

RADECKI, T. (1977). Mathematical model of time effective information retrieval systems based on the theory of fuzzy sets. *Information Processing and Management*, **13**, 109–116.

RADECKI, T. (1979). Fuzzy set theoretical approach to document retrieval. *Information Processing and Management*, **15**, 247–259.

RADECKI, T. (1982*a*). Similarity measures for Boolean search request formulations. *Journal of the American Society for Information Science*, **33**, 8–17.

RADECKI, T. (1982*b*). Incorporation relevance feedback into Boolean retrieval systems. *Proceedings of the Fifth International Conference on Information Retrieval*, West Berlin, Germany.

ROBERTSON, S. (1978). On the nature of Fuzz: a diatribe. *Journal of the American Society for Information Science*, **29**, 304–307.

ROBERTSON, S. E., MARON, M. E. & COOPER, W. S. (1982). Probability of relevance: a unification of two competing models for document retrieval. *Information Techonology: Research and Development*, **1**, 1–21.

SACHS, W. M. (1976). An approach to associative retrieval through the theory of fuzzy sets, *Journal of the American Society for Information Science*, **27**, 85–87.

SALTON, G. (1975*a*). *Dynamic Information and Library Processing*. Englewood Cliffs, New Jersey: Prentice–Hall.

SALTON, G. (1975*b*). *A Theory of Indexing*. Regional Conference Series in Applied Mathematics No. 18, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.

SALTON, G. (1979). Mathematics and information retrieval. *Journal of Documentation*, **35**, 1–29.

SALTON, G. & WALDENSTEIN, R. H. (1978). Term relevance weights in on-line information retrieval. *Information Processing and Management*, **14**.

SALTON, G. & WONG, A. (1978). Generation and search of clustered files. *ACM Transactions on Databsase Systems*, **3**, 321–346.

SALTON, G. & WU, H. (1980). *A term weighting model based on utility theory*. Cornell University Department of Computer Science, Ithaca, New York.

SALTON, G. & YU, C. T. (1982). A comparison of two term dependence models. *Proceedings of the Fifth International Conference on Information Retrieval*, West Berlin, Germany.

SWETS, J. A. (1967). Information retrieval systems. In KOCHEN, M., Ed., *The Growth of Knowledge*. New York:Wiley.

TAHANI, V. (1976). A fuzzy model of document retrieval systems. *Information Processing and Management*, **12**, 177–187.

VAN RIJSBERGEN, C. J. (1979). *Information Retrieval* (2nd edition). London: Butterworths.

WALLER, W. G. & KRAFT, D. H. (1979). A mathematical model of a weighted Boolean retrieval system. *Information Processing and Management*, **15**, 235–245.

WU, H. & SALTON, G. (1981). A comparison of search term weighting: term relevance vs. inverse document frequency. *Technical Report TR 81-457*, Department of Computer Science, Cornell University, Ithaca, New York.

YAGER, R. R. (1980). A logical bibliographic searcher: an application of fuzzy sets. *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-10**, 51–53.

YU, C. T. & SALTON, G. (1971). Effective information retrieval using term accuracy. *Communications of the Association for Computing Machinery*, **20**, 135–142.

ZADEH, L. (1965). Fuzzy sets. *Information and Control*, **8**, 338–353.

Appendix A

WALLER-KRAFT CRITERIA FOR WEIGHTED RETRIEVAL

1. Separability. Each query is processed by evaluating a document along each individual weighted term in the query separately, without regard to the context of the terms (whether they stand alone or appear as part of higher-level expressions). Sub-expressions are evaluated similarly before evaluating the combined subexpressions in the overall Boolean query expression.

2. Generalization. The evaluation mechanisms for weighted Boolean retrieval should resemble in structure the unweighted (traditional) Boolean retrieval evaluation

mechanisms. Moreover, when the weights and membership functions are discrete, the evaluation mechanisms should yield the equivalent results.

3. Boolean self-consistency. Logically equivalent queries should yield identical results for any given document.

4. For queries involving only a simple weighted term:

(a) g should be monotonically increasing as F increases, for positive weights a;

(b) g should be constant as F changes, for weight $a = 0$;

(c) the magnitude of the change in g, as F increases, should itself be monotonically increasing as the weight a increases.

5. For queries consisting of higher-level expressions:

(a) the document evaluation should be monotonically increasing as F increases, for positive weight a and term t not negated;

(b) the evaluation should be constant as F changes, for weight $a = 0$;

(c) the evaluation should be nonincreasing as weight a increases, for term t connected to the query via AND;

(d) the magnitude of the change in the evaluation of the document as the evaluation for a single term increases should itself be monotonically increasing as weight a increases;

(e) the magnitude of the change in the evaluation as F (or a) increases should itself be monotonically increasing as a (or F) increases;

(f) the evaluation should be monotonically decreasing as F increases, for positive weight a and the term negated.

6. For higher level expressions:

(a) the evaluation should be non-increasing if an additional expression is connected to the original query via AND;

(b) the evaluation should be non-decreasing if an additional expression is connected to the original query via OR.