PASSAGE-BASED RETRIEVAL USING PARAMETERIZED FUZZY SET OPERATORS

K.Y. Kong[1], Robert W.P. Luk[1], W. Lam[2], K.S. Ho[1] and F.L. Chung[1]

[1]Department of Computing, The Hong Kong Polytechnic University
[2]Department of Systems Engineering and Engineering Management
Chinese University of Hong Kong

Abstract

We examined the use of parameterized fuzzy set operators in passage-based retrieval because these operators generalized some of the previous work in passage-based retrieval. Here, we examined the impact of the parameter values on the retrieval performance. The use of these fuzzy operators is justified on the basis of three principles, which are tested by comparing the retrieval effectiveness of different fuzzy operators for different principles. We used the TREC-6 English ad hoc retrieval test collection for evaluation and concluded that different fuzzy operators are useful for different retrieval contexts or requirements.

Keywords: information retrieval, fuzzy set theory, passage retrieval.

1. INTRODUCTION

Passage-based retrieval [1-6] has been investigated quite extensively. Its advantages include mitigating the impact of document length on retrieval effectiveness, as well as facilitating question answering retrieval tasks. In passage-based retrieval, the similarity score of each passage is obtained and these scores are combined to formulate the document similarity score. In the past, the passage scores have been combined using simple averages, as well as taking the maximum. The suitability of a particular score combination is usually determined experimentally. Here, we formulate three principles of relevance decisions which can be applied to passage-based retrieval. Each principle specifies certain desirable algebraic properties (called axioms) that justify the use of particular fuzzy set operators for combining passage scores together. We then test which principle is the most or least applicable in the context of passage-based (English) ad hoc retrieval by evaluating the retrieval effectiveness of different fuzzy set operators. The application of fuzzy set operators to combine passage scores justified on the basis of the relevance decision principles is novel because past fuzzy information retrieval is focused in developing fuzzy retrieval model [8], novel fuzzy query languages [9] and fuzzy similarity scores [10].

2. RELEVANCE DECISION PRINCIPLES

Suppose that there is a hypothetical user who makes relevance decisions for every occurrence of the concept related to the information need of the user. The user detects where there are likely concepts that relate to his/her information need (called the topic) in the document. For each occurrence of a concept related to the topic, the user decides based on information read from the context whether that part of the document contains any information needed by the user. If so, then the document is labeled as relevant. Otherwise, the next occurrence of a concept related to the topic is examined. Alternatively, the user accumulated the evidence in his/her mind and makes the final relevance decision after reading the document. This process (Figure 1) is similar to an evaluator making decisions as to whether the document is relevant by examining keywords in context (KWIC) [11].

In passage-based retrieval, the evidence is obtained from a passage and the passage scores are combined by the evidence combining function *C(.)*, which can be based on three different principles. First, a document contains many places where relevance judgment is made. In many information retrieval applications or evaluations, a document is considered relevant if there is one judgment that certain part of the document is relevant in order to save manual effort. Therefore, we formulated the following:

Disjunctive Relevance Decision (DRD) Principle: If the user considers a particular occurrence of a particular concept with the associated context in a document is relevant, then the entire document is relevant.

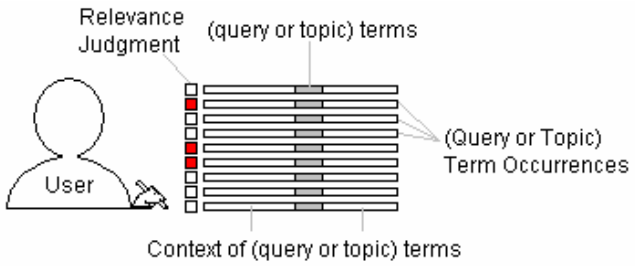

Figure-1. A model of the process of making human judgment of document relevance.

Table 1: Desirable disjunction axioms

Axiom Name	Axiom Conditions
Boundary Condition	*C(0,0) = 0; C(1,0)=C(1,0)=C(1,1)=1*
Commutative	*C(a,b) = C(b,a)*
Monotonic	*a £ a' and b £ b' Þ C(a,b) £ C(a',b')*
Associative	*C(C(a,b),c) = C(a,C(b,c))*

Even a single user may have different levels of relevance judgment, i.e. documents that are highly relevant and those that are barely relevant. Here, we make a conjecture that the degree of relevance of a document depends on the number of pieces of relevant information that the document contains. If there is more relevant information in the document then the relevance of a document is higher. We can summarize this in the second principle:

Aggregate Relevance (AR) Principle : if the user finds more occurrences of the concepts related to

the information need, then the relevance of the document to the user's information need is higher.

Table 2: Desirable aggregation axioms

Axiom Name	Axiom Conditions
Boundary Condition	$C(0,..,0)=0; C(1,..,1)=1;$
Monotonic Non-Decreasing	$\forall i \le n, a_i \ge b_i \Rightarrow C(a_1,\ldots, a_n) \ge C(b_1,\ldots, b_n)$

The AR principle is directly relevant to information retrieval because it can be used to justify the use of certain models to combine evidence. Table 2 shows the axioms of functions that comply with the AR principle. The boundary conditions are different from the DRD principle, which are less restrictive than those for the DRD principle. The aggregation function may not be separated unlike functions that comply with the DRD principle. Note that functions that comply with the DRD principle automatically also comply with the AR principle.

Finally, the third principle is as follows:

Conjunctive Relevance Decision (CRD) Principle: If the user considers all occurrences of any concepts related to the user information need with the associated contexts in a document are all relevant, then the entire document is relevant.

The desirable axioms for the CRD principle are shown in Table 3. Note that any passage which gives a zero score would result in a combined zero score due to the boundary condition axiom (i.e., $C(0,1) = 0$). This is not really applicable in the context of ad hoc retrieval. There are several ways to deal with this problem. One approach is to perform a pooled estimate such that the passage score is always larger than zero, similar to language models [12]. The other approach taken here is to ignore those passages that have passage scores equal to zero. Effectively, we are not following the CRD principle but the AR principle.

Table 3: Desirable conjunction axioms

Axiom Name	Axiom Conditions
Boundary Condition	$C(0,0)=C(1,0)=C(1,0)=0; C(1,1)=1$
Commutative	$C(a,b) = C(b,a)$
Monotonic	$a \le a' \text{ and } b \le b' \Rightarrow C(a,b) \le C(a',b')$
Associative	$C(C(a,b),c) = C(a,C(b,c))$

The three principles can be related by their desirable axioms. Specifically, the axioms for the CRD principle are the same as those for the DRD principle, except that the boundary condition axioms are different for CRD and DRD principles. Since the function can be applied recursively, the function can be separated. Similar to functions that comply with the DRD principle, functions that comply with the CRD principle also complies with the AR principle. However, there are functions that comply with the AR principle but not the CRD and DRD principles and these functions:

a. may have different boundary conditions or
b. may not be able to be applied recursively or
c. may not be commutative or
d. may not be associative.

Figure 2 shows a Venn diagram of the set of functions that comply with these principles.

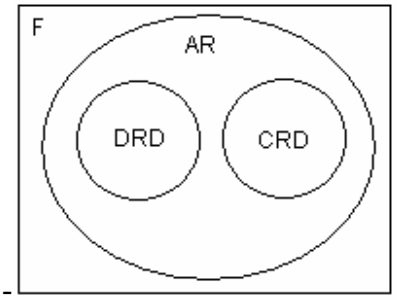

Figure-2. The set of functions that complied with different principles and *F* is the universe of all functions.

Developing retrieval models should comply with one or more of these principles by requiring the evidence combination function to exhibit the relevant axiomatic properties, thereby providing an epistemological basis of the use of these mathematical functions, apart from pragmatic significance and maintaining consistency with evaluation practices. For existing retrieval models which are developed without these principles stated explicitly, we can not demand that every evidence combination function must comply fully with those properties (or axioms) imposed by the relevant principles.

We believe that, in evaluation, typically the DRD and the AR principles are followed by the user. If the evaluator sees a highly relevant piece of information, then the user will probably stop scanning the document immediately to save effort and label the document as relevant. This will realize the DRD principle. If the document contains pieces of evidence that are not highly relevant but aggregated together can be considered as relevant, then the evaluator will probably make decisions similar to a user under the AR principle.

3. FUZZY LOGICAL OPERATORS

According to the DRD principle, we can model the relevance decision of the user for the *k*-th passage in the *i*-th document as the relevance decision function *rd(i,k)* which returns true for relevant and false for irrelevant. Here, we made a drastic simplification that the user only looked at the query terms and not their related terms (e.g. synonyms) or expressions. Therefore, the Boolean expression for $r(d_i)$ is:

$$r(d_i) \leftarrow \bigvee_k rd(i,k)$$

This view of finding relevant documents can also be applied to the standard Boolean retrieval model or the generalized weighted Boolean retrieval model by examining whether information at the *k*-th passage would satisfy the information need expressed by the (generalized weighted) Boolean expression. In general, the expresiveness of the Boolean expression need not be sacrificied because any Boolean expression can be put into the disjunctive normal form (DNF). In genearl, the expressiveness of the information need not be compromised by all the relevance decision principles (i.e., DRD, CRD and AR). Similarly, the CRD principle requires:

$$r(d_i) \leftarrow \underset{k}{\wedge} rd(i,k)$$

Although the CRD principle does not sacrifice expressiveness because any Boolean expression can be put into the conjunctive normal form (CNF), this form is not desirable because CNF combines disjunctions of individual pieces of information.

3.1 Fuzzy Set Version

In general, $r(d_i)$ is the membership of the i-th document that belongs to the set of relevant documents using the evidence combination function $C(.)$ as follows:

$$r(d_i) \leftarrow C(\{rm(i,k)\})$$

where $rm(i,k)$ is the relevance measure of the k-th passage in the i-th document that is relevant to the query. Here, $rm(i,k)$ needs to be normalized between zero and one, instead of the two logical values (i.e., true and false) and it needs not be a σ-algebra.

By the DRD principle, the combination function becomes the Boolean expression using the generalized fuzzy union operation as follows:

$$C(\{rm\{i,k\}\}) = \bigcup_k rm(i,k)$$

Table 4: Example fuzzy set union and intersection operators.

Name and Reference	Union	Intersection
Yager [14]	$\min\left\{1, \sqrt[a]{\sum_k rm(i,k)^a}\right\}$	$1 - \min\left\{1, \sqrt[a]{\sum_k rm(i,k)^a}\right\}$
Dombi [15]	$\frac{1}{1 + \sqrt[-a]{\sum_k \left(\frac{1}{rm(i,k)} - 1\right)^{-a}}}$	$\frac{1}{1 + \sqrt[a]{\sum_k \left(\frac{1}{rm(i,k)} - 1\right)^{a}}}$

Some fuzzy union operators are shown in Table 4. The fuzzy union operations satisfy the axioms in Table 1. Note that some mathematical choice to realize the fuzzy union operation can conform to DeMorgan's theorem. Such fuzzy union operators are called dual operators [13] but it is not known whether information retrieval requires these dual operators. Similar to the DRD principle, the Boolean expression based on the CRD principle becomes the following fuzzy set operation:

$$C(\{rm(i,k)\}) = \bigcap_k rm(i,k)$$

Table 4 shows examples of fuzzy intersection operations which satisfy the desirable axioms in Table 3.

3.2 Aggregation Operators

A well-known general aggregation operation is the generalized mean function $h_a(i)$ [16] (also the same as the disjunction of the extended Boolean model [17]) where a is the parameter that controls whether the aggregation is hard or soft. This function satisfies the desirable axioms in Table 2 and it is defined as:

$$h_a(\{rm(i,k)\}) = \left(\frac{\sum_k rm(i,k)^a}{m} \right)^{1/a} \qquad for \quad -1 \le a \le \infty$$

where m is the total number of passages in the i-th document. Here, we used this function as an example of evidence combination that comply with the AR principle, i.e. $C(\{rm(i,k)\}) = h_a(\{rm(i,k)\})$. In general, there are other aggregation operators that can be experimented (e.g. the ordered weighted averaging operators [18]).

Some passage-based retrieval mechanisms obtained the maximum passage score of a document as the document score. This is the same as setting a = ¥ for the generalized mean (i.e., $h_{¥}(.)$). Some passage-based retrieval mechanisms summed [6] the passage scores as the document scores. This is the same as $h_1(i)$ if the number of passages of a document is the same for all documents. In practice, the number of passages per document varies. Even though summing the passage scores is not the same as $h_1(i)$, the function that sums the passage scores satisfies the axioms of functions that comply with the AR principle, provided that after summation the total passage scores for a document is normalized between zero and one.

4. EVALUATION

We used the TREC-6 English ad hoc retrieval test collection (Disk 4 and 5) and topic 301-350 inclusive. The passage score is obtained using the BM11 term weighting function of the 2-Poisson model [19]. These passage scores are normalized by dividing the passage scores over the sum of the passage scores. The passage size is set to 500 words. The inverted index is modified to include passage information for retrieval. The title queries were used for evaluation.

4.1 Comparison between operators

In this experiment, we compare the mean average precision (MAP) of the different fuzzy set operators: union, intersection and aggregation, with respect to different values (between 0 and 100) for the parameter a. We have used the Dombi union and intersection, as well as the generalized mean, as examples for comparison. We have also obtained results for the Yager union and intersection [13] but these results are not reported because they are similar to the Dombi operators. The Dombi union operator achieved better performance than that of the Dombi intersection for $\alpha >$ 5, indicating that the DRD principle is more suitable to achieve good MAP. This is expected that since the DRD principle would formulate a combination score that is more explorative where any passage with a high score in the document will substantially increase the document score. The best

performance of the Dombi union operator is similar to the best performance of the generalized mean operator. Interestingly, the MAP of the generalized mean operator with $a = \infty$ is not the best. The best MAP performance was achieved using $a = 80$ for the generalized mean operator. The standard Euclidean norm (i.e., $a = 2$) does not produce good results.

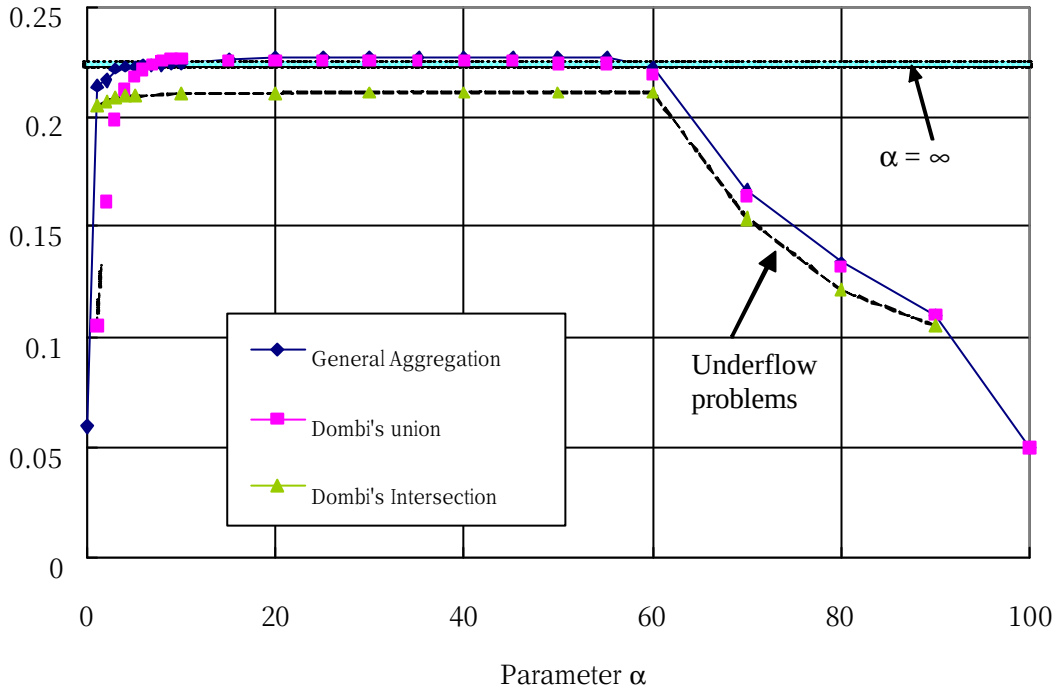

Figure-3. Mean average precision of fuzzy set operators with different (a) parameter values.

For $\alpha > 60$, the MAPs of the Dombi operators fall with increasing values of a. We found that this is due to the arithmetic underflow problem. This problem is solved by combining the scores of the passages of the top 1000 ranked documents instead of all the passages. Figure 4 shows the MAP performance of the Dombi and generalized mean operators. The best MAP performance of using passage of the top 1000 ranked documents was similar to the best MAP performance of the passages of all the retrieved documents. We conclude that this technique is effective to avoid underflow.

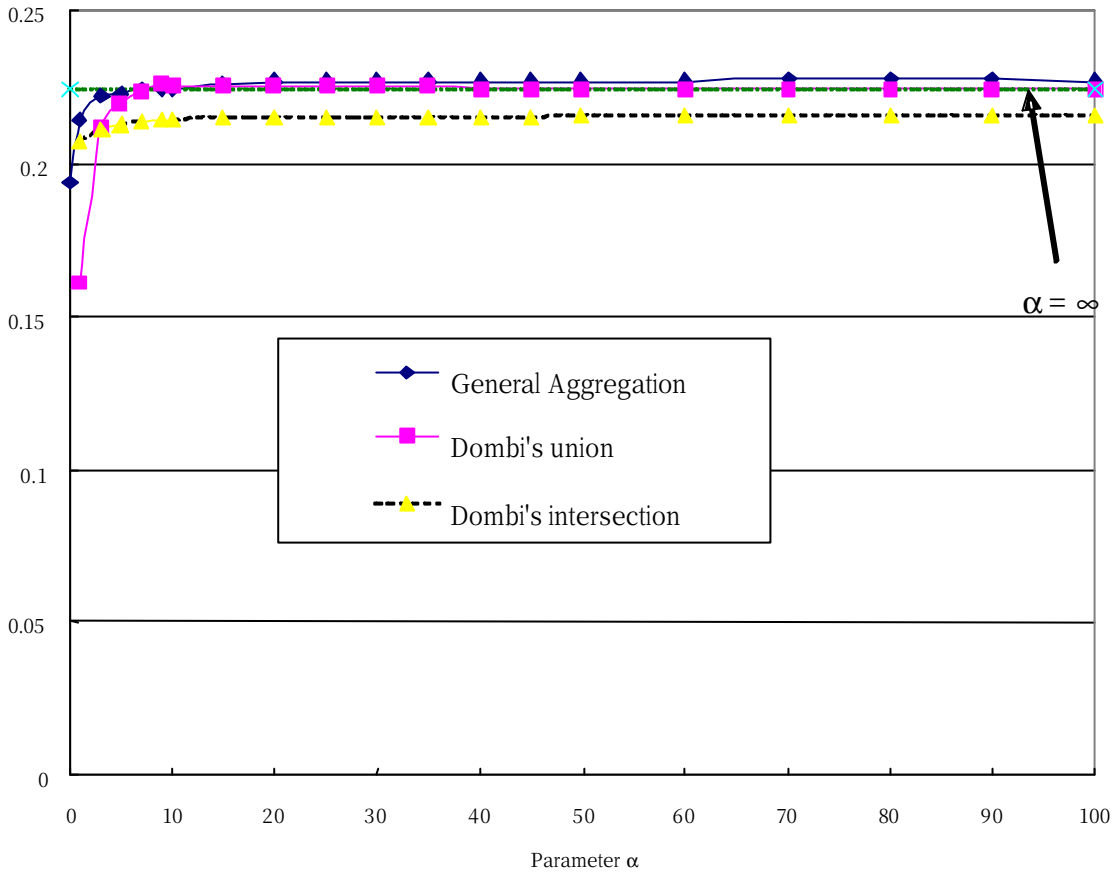

Figure-4. Mean average precision of fuzzy set operators with different (α) parameter values using passages of the top 1000 ranked documents.

4.2 Precision

Table 5 shows the interpolated recall-precision performance for comparison between the (near) best cases of the fuzzy set operators. Even though we have shown that the Dombi union was performing better than the Dombi intersection in Figure 3 and 4, the Dombi intersection was able to perform better than the Dombi union for the top ranked documents. Therefore, the CRD principle may be applicable to the cases where the user wants to find one or two highly relevant items instead of exploring for better coverage of relevant documents. The aggregation operator performs consistently better than document retrieval except when the recall level is low.

Table 5: Interpolated Recall-Precision Performance Comparison of the Best Cases. The percentages are the differences in performance with respect to the corresponding performance of document retrieval.

Recall Level	Document Retrieval	Aggregation Operation $h_a(.)$				Dombi's union (a = 40)		Dombi's intersection (a = 40)	
		$a = \infty$		a = 80					
Interpolated Recall - Precision Averages:									
0.00	0.672	0.667	-0.7%	0.687	2.17%	0.661	-1.59%	0.688	2.31%
0.10	0.485	0.476	-1.9%	0.511	5.34%	0.501	3.24%	0.496	2.14%
0.20	0.403	0.372	-7.8%	0.407	0.84%	0.394	-2.38%	0.394	-2.28%
0.30	0.287	0.281	-2.2%	0.310	8.00%	0.309	7.62%	0.273	-4.98%
0.40	0.237	0.234	-1.4%	0.249	5.07%	0.250	5.32%	0.224	-5.53%
0.50	0.189	0.194	2.6%	0.208	10.17%	0.211	11.7%	0.174	-7.95%
0.60	0.13	0.156	20.2%	0.153	17.84%	0.155	19.5%	0.139	7.26%
0.70	0.094	0.114	22.22%	0.115	22.86%	0.114	21.5%	0.084	-10.3%
0.80	0.052	0.045	-12.0%	0.052	0.00%	0.051	-0.39%	0.047	-9.11%
0.90	0.043	0.04	-8.9%	0.042	-3.04%	0.042	-2.10%	0.038	-12.4%
1.00	0.032	0.028	-11.8%	0.030	-8.10%	0.030	-8.10%	0.027	-17.1%
Average precision(non-interpolated) for all rel docs(averaged over queries)									
MAP	0.215	0.2156	0.14%	0.228	5.85%	0.226	4.97%	0.2113	-1.86%

4.3 Efficiency Issues

Table 6 shows the storage demand of the inverted index for document retrieval, as well as for passage-based retrieval. The index consists of three components: the postings, the dictionary and the (document) extension information (e.g. file name). Passage length and other related information is added as the extension information. For the passage size of 500 words each, the relative storage is about 112% of the storage for the corresponding inverted index for document retrieval. The increase in storage is not considered to be substantial.

Table 6: Storage efficiency for document and passage-based indexing.

Level Storage	Document Original (Mbytes)	Passage		
		Original (Mbytes)	Relative	Additional
Postings	755	848	112%	12%
Dictionary	78	86	110%	10%
Extension	68	76	112%	12%
Total	901	1010	112%	12%

Table 7 shows the retrieval time per query. The additional retrieval time is about 60% to 70% for passage-based retrieval compared with document retrieval. The differences in retrieval between

different fuzzy set operators are not too substantial. Therefore, we conclude that with a passage size of at most 500 words, the additional storage overhead and the retrieval time for passage-based retrieval compared with that of document retrieval are not too substantial for the TREC-6 data.

Table 7: Retrieval time comparison between different fuzzy sets operations.

Retrieval Type	Retrieval Time per query (s)	Relative Retrieval Time
Document retrieval	4.62	100%
Aggregation operation (a is infinity)	7.45	161%
Union operation (a = 40)	7.60	165%
Intersection operation (a = 40)	7.48	162%

5. CONCLUSION AND FUTURE WORK

We conclude that the different parameterized fuzzy set operators are useful in different contexts. If the user is explorative, the fuzzy union operators are appropriate. If the user wants only one or two highly relevant items, the fuzzy intersection operators may be more appropriate. However, the fuzzy intersection operators need to be modified so that its boundary condition property can be avoided (e.g. ignoring passages with zero scores). In effect, these fuzzy intersection operator behave more like an aggregation operator. For the best overall results, the generalized mean aggregation operator appears to be a good choice. We are evaluating with (a) more Fuzzy set operators, (b) different query types, (c) pseudo relevance feedback and (d) passage size and types.

ACKNOWLEDGEMENTS

This work is supported by the Hong Kong Polytechnic University Grant No. A-PE36.

REFERENCES

1. G. Salton, J. Allan, and C. Buckley, "Approaches to passage retrieval in full text information systems". *ACM SIGIR 93*, pp. 49-56, 1993.
2. M. Kaszkiel, and J. Zobel, "Passage retrieval revisited". *ACM SIGIR 97*, pp. 178-185, 1997.
3. J.P. Callan, "Passage-level evidence in document retrieval". *ACM SIGIR 94*, pp. 302-310, 1994.
4. M.A. Hearst, and C. Plaunt, "Subtopic structuring for full-length document access". *ACM SIGIR 93*, pp. 59-68, 1993.
5. M. Kaszjiel, J. Zobel, and R. Sacks-Davis, "Efficient passage ranking for document database". ACM Transactions on Information Systems, 17(4): 406-439, 1999.
6. R. Wilkinson, "Effective retrieval of structured documents". ACM SIGIR 94, pp. 311-317, 1994.
7. J. Kupiec, Jan Pedersen, and F. Chen, "A trainable document summarizer", ACM SIGIR 95, pp. 68-73, 1995.

8. D.H. Kraft, G. Bordogna, and G. Pasi, "Fuzzy set techniques in information retrieval," in J.C. Bezdek, D. Didier, and H. Prade, (eds.), *Fuzzy Sets in Approximate Reasoning and Information Systems*, vol. 3, The Handbook of Fuzzy Sets Series, Norwell, MA: Kluwer Academic Publishers, 1999.
9. Bordogna, and G. Pasi, "A fuzzy linguistic approach generalizing Boolean IR: a model and its evaluation", *Journal of the American Society for Information Science* 44(2): pp. 70-82, 1993.
10. S.E. Robertson, "On the nature of fuzzy: a diatribe" *Journal of the American Society for Information Science* 29: pp. 304-307, 1978.
11. F. Song, and W.B. Croft, "A general language model for information retrieval". *ACM CIKM 99*, pp. 316-321, 1999.
12. D. Dubois, and H. Prade, "A review of fuzzy set aggregation connectives". *Information Sciences*, vol. 36, pp. 85-121, 1985.
13. R.R. Yager, "On a general class of fuzzy connectives". *Fuzzy Sets and Systems*, vol. 4, pp. 235-242, 1980.
14. J. Dombi, "A general class of fuzzy operators, the DeMorgan class of fuzzy operators and fuzziness measures induced by fuzzy operators". *Fuzzy Sets and Systems*, vol. 8, pp. 149-163, 1982.
15. H. Dychkoff, and W. Pedrycz, "Generalized mean as a model of compensative connectives". *Fuzzy Sets and Systems*, vol. 14, pp. 143-154, 1984.
16. E. Fox, S. Betrabet, and M. Koushik, "Extended Boolean models", In W. Frakes and R. Baeza-Yates, *Information Retrieval: Data Structures and Algorithms*, NJ: PHI, pp. 393-418, 1992
17. G. Bordogna, and G. Pasi, "Linguistic aggregation operators in fuzzy information retrieval". International Journal of Intelligent Systems, vol. 10, pp. 233-248, 1995.
18. S. E. Robertson, and S. Walker, "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval". *ACM SIGIR 94*, pp. 232-241, 1994.