The following publication E. K. F. Dang, R. W. P. Luk and J. Allan, "A Principled Approach Using Fuzzy Set Theory for Passage-Based Document Retrieval," in IEEE Transactions on Fuzzy Systems, vol. 29, no. 7, pp. 1967-1977, July 2021 is available at https://doi.org/10.1109/TFUZZ.2020.2990110

A Principled Approach Using Fuzzy Set Theory for Passage-based Document Retrieval

Edward Kai Fung Dang, Robert Wing Pong Luk, and James Allan

***Abstract*—We present a novel principled approach to passage-based (document) retrieval using fuzzy set theory. The approach formulates passage score combination according to general relevance decision principles. By operationalizing these principles using aggregation operators of fuzzy set theory, our approach justifies the common heuristics of taking the maximum constituent passage score as the overall document score. Experiments show that this heuristics is only the near best, with some fuzzy set aggregation operators stipulated in our approach being better methods. The significance of our principled approach is the applicability of many passage score combination methods, potentially bringing further performance enhancement. Experiments on several Text REtrieval Conference (TREC) collections demonstrate that our approach performs significantly better than document-based retrieval. While recent works in the literature mostly employ document-based rather than passage-based retrieval due to the common conception that document length normalization solves the problem of varying document lengths, our results show that document length normalization alone is not sufficient, especially in pseudo-relevance feedback retrieval.**

***Index Terms*—Fuzzy information retrieval system, Principled passage-based retrieval, Fuzzy aggregation, t-conorms, Generalized mean, Performance evaluation.**

I. Introduction

WEB search engines such as Google and Bing serve millions of requests per day. One problem for web search is the large variation of the length of webpages ('documents'), which affects the occurrence statistics of query terms in documents, thereby affecting the prediction of document relevance. One way to tackle this issue is passage-based retrieval [2], [3] because documents are divided into passages of more uniform length. The overall ranking score of a document is obtained by combining the scores of its constituent passages. The existing paradigm for combining passage scores is the use of heuristics (e.g. taking the maximum or average [2]–[5]), without a proper understanding of why these methods are effective. We tackle the question of how to achieve effective passage-based retrieval by a fuzzy set theoretical framework that specifies how passage scores should be combined. The motivation of this framework is that it can provide a theoretical understanding of some common heuristics and also indicate how better methods may be found, so as to further enhance retrieval effectiveness.

In our fuzzy set theoretical framework, passage score combination is governed by certain principles, namely the Disjunctive Relevance Decision (DRD) and Aggregated Relevance (AR) principles [6], which are formulated based on the evaluation policy of Text REtrieval Conference (TREC) [7]. These principles are operationalized by aggregation operators [8], [9] of fuzzy set theory. The effectiveness of our methodology is demonstrated by extensive retrieval experiments on several TREC test collections of diverse document types. Our approach justifies theoretically the common heuristic of taking the maximum passage score to be the overall document score, as the max operator belongs to the class of t-conorm operators [9], [10] which follow the DRD and AR principles. We find the Dombi t-conorm (DombiD) operator and Generalized Mean (GMean) to yield the best retrieval effectiveness, while the max operator is a near best. We also find that DombiD and GMean yield better results than other past methods, including averaging passage scores and a probabilistic approach that assumes passages to be independent. Fig. 1 illustrates the calculation of the overall document ranking score in document-based and passage-based retrieval with an example query.

An alternative way to tackle the problem of varying document lengths is by incorporating document length normalization [11] in the retrieval models, such as the successful BM25 [12] and PL2 [13], [14]. This approach is the current prevailing paradigm, with TREC participants and recent works in the literature mostly employing document-based retrieval that uses models with built-in document length normalization, rather than passage-based retrieval. Thus we compare our passage-based retrieval against document-based retrieval using the PL2 and BM25 models with built-in normalization. Our results demonstrate our passage-based retrieval performs significantly better than the document-based retrieval baselines, especially in pseudo-relevance feedback (PRF). An intuition is that in passage-based retrieval, query expansion (QE) terms of PRF may be selected within a more focused region near the given query terms within a document. This means that passage-based retrieval enables cleaner QE terms to be selected, while in document-based PRF more noise terms may be chosen especially in long documents. Thus, we show that document-length normalization alone is not sufficient to solve the problem of long document lengths, unlike passage-based retrieval.

The contributions and significance of our work are as follows. First, we establish a principled fuzzy set theoretical framework for passage-based retrieval. It signifies a novel real life application of fuzzy set theory. Second, we introduce a

E.K.F. Dang and R.W.P. Luk are with the Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Hong Kong. E-mail: {cskfdang, csrluk}@comp.polyu.edu.hk.

J. Allan is with Center for Intelligent Information Retrieval, College of Information and Computer Sciences, University of Massachusetts, Amherst. Email: allan@cs.umass.edu.

This work is an extension to our workshop paper [1].

Manuscript received May 15, 2019. This work was supported in part by the grant DGRF P0014024 and the Center for Intelligent Information Retrieval.

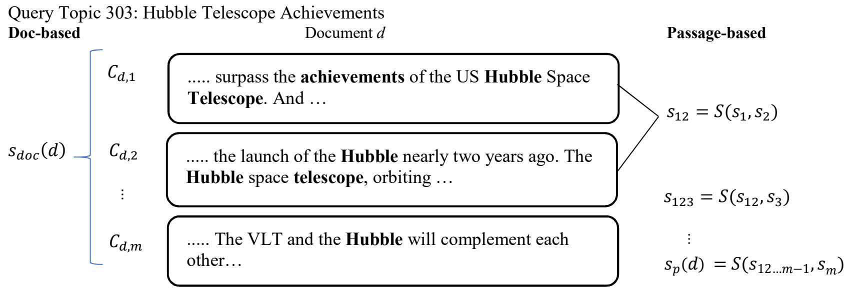

Fig. 1. Illustration of document-based and passage-based retrieval. For the document d with constituent passages $C_{d,1}, \cdots, C_{d,m}$, document-based retrieval returns a ranking score $s_{doc}(d)$ based on query term statistics with the whole document as a single unit, while in passage-based retrieval the overall score $s_p(d)$ is obtained by combining scores of the passages, $\{s_i\}$, such as by using a fuzzy set aggregation operator $S(\cdot)$, e.g. a t-conorm.

novel methodology of designing an application system based on principles that are formulated according to the specific evaluation policy. While we apply this methodology to passage-based retrieval in this work, the methodology is general and may be adopted in other applications, including those that utilize fuzzy set theory. Third, we formulate the principles that govern passage score combination, and show that these principles may be operationalized by aggregation operators of fuzzy set theory. Fourth, our approach justifies theoretically the common heuristic of taking the maximum passage score to be the overall document score, as the max operator is a t-conorm operator that conforms to our formulated DRD and AR principles. Fifth, our empirical study determined which t-conorm operators give good performance in passage-based retrieval. Sixth, our experiments show that the generalized mean is effective for passage score combination. In fact, the framework enables many aggregation operators of fuzzy set theory to be applicable in passage score combination, with the potential of finding other fuzzy set aggregation operators that may be even more effective in passage-based retrieval. Seventh, while recent works in the literature mostly use document-based retrieval due to the common conception that document-length normalization can solve the problem of varying document lengths, our results show that document-length normalization is not sufficient, especially in PRF retrieval, with passage-based retrieval performing significantly better. Last, our approach of a principled fuzzy set theoretical framework is general as the passage scores can be defined by any retrieval model, as exemplified by BM25 and PL2, and may include future effective models.

The rest of the article is organized as follows. Section II provides some background material. Our approach is described in Section III. Section IV presents our experimental environment and results. Section V is a conclusion of the current study.

II. Background

As we need to operationalize the relevance decision principles of our approach, which specify how relevance is aggregated, we first review aggregation via fuzzy set theory in Section II-A. Since our current study focuses on passage-based retrieval, a brief review of this area is provided in Section II-B.

A. Aggregation via Fuzzy Set Theory

To operationalize our principles, the tools provided by fuzzy set theory are well suited for the following reasons. First, relevance is generally represented by a numeric value between 0 and 1, which may be mapped to the membership value of a fuzzy variable. Second, a large number of aggregation operators in fuzzy set theory have been studied and applied in various fields such as decision making, expert systems, etc. [15], so that these operators may be tested for aggregating relevance in our current problem. Examples of aggregation operators of fuzzy set theory include mean operators [16], triangular norm (t-norm) and its dual t-conorm [9], [10], ordered weighted averaging operator [17]–[19] as applied in [20] and [21], induced ordered weighted averaging operator [22], fuzzy weighted averaging operator [23], [24], fuzzy ordered weight averaging [25], fuzzy hybrid averaging operator [25], geometric-mean averaging operator [26], continuous ordered weighted geometric averaging operators [27], [28], Bonferroni mean operators [29], linguistic aggregation operator [30], linguistic power aggregation operators [31], trapezoidal fuzzy power aggregation operators [32], hesitant fuzzy power aggregation operators [33], hesitant fuzzy linguistic weighted aggregation operators [34], etc. However, while many aggregation operators are known, there is little past work that i) analyzes the different operators, ii) recommends which operator to use, and iii) compares the effectiveness of the use of them, in retrieval applications. Because of the large number of possible aggregation operators, it is not feasible to test all of them in this work, nor is it the aim of this work to find the best operator. Therefore, in this article we report experiments of passage-based retrieval with a selection of aggregation operators to combine relevance scores, including the t-norm, t-conorm and generalized mean (GMean), which is an averaging aggregation operator. These aggregation operators are selected because our previous work found that using the t-conorm or GMean is effective for retrieval [1]. The GMean has also been applied successfully in the Extended Boolean Model [35] of information retrieval (IR). Furthermore, the commonly used $\max(\cdot)$ function for passage-based retrieval (e.g. [2], [3]) is a special case of t-conorms and GMean, so that we expect appropriate choices of t-conorm and GMean to perform at least as good as the $\max(\cdot)$ function. Overall, the various fuzzy aggregation operators cover different semantics, with the t-conorms and t-norms corresponding to the quantifiers ‘exists’ and 'forall', respectively, and the GMean being somewhere in between these extremes.

We briefly r eview t he u se o f f uzzy s et t heory i n I R in the past. In IR, a fuzzy set model may be constructed by applying fuzzy set operators to the query term weights to obtain the overall membership value ν of a document. In conventional fuzzy set theory, the operators $\max(\cdot)$, $\min(\cdot)$ and 1-ν represent disjunction (or union), conjunction (or intersection) and negation, respectively [36], [37]. However, a problem is that the min and max operators may generate document rankings that do not agree with human intuition [38]. A reason is that with the min and max operators, the resulting membership value is governed by only one operand regardless of the value of the other operands, an issue known as 'single operand dependency' [39], [40]. To tackle this issue, [39] and [40] suggested a class of 'positively compensatory operators' for the fuzzy set model of IR. These operators $\theta(x, y)$ are idempotent, i.e. $\theta(x, x) = x$, and have the property $\min(x, y) < \theta(x, y) < \max(x, y)$ for $x \neq y$. Thus, the positively compensatory operators overcome the single operand dependency issue of min and max. However, [41] showed that binary operators that are positively compensatory suffer from the weakness that they cannot be associative, i.e. $\theta(\theta(x, y), z) \neq \theta(x, \theta(y, z))$. This poses the problem that logically identical queries may give different document similarity scores, such as the queries (q_1 AND q_2) AND q_3 and q_1 AND (q_2 AND q_3). Therefore, [41] pointed out the necessity of using n-ary operators such as the p-norm. The p-norm belongs to a class of aggregation operators in fuzzy set theory called mean operators, $M(\cdot)$ (e.g. [16]), which return a value between the minimum and maximum of its operands and is thus positively compensatory. In fact, the p-norm corresponds to the Extended Boolean Model [35].

In the Extended Boolean Model, disjunction and conjunction have a geometric interpretation. In n-dimensional space (as in the vector space model), for disjunction of m query terms having weights s_i with $i = 1, \cdots, m$, the document score is given by the distance away from the origin, which is the most undesirable point. The distance from the origin is given by a *normalized* p-norm:

$$s(\mathrm{OR}) = \sqrt[p]{\frac{1}{m}\sum_{i=1}^{m}(s_i)^p}, \tag{1}$$

with $p \geq 1$. The extended Boolean disjunction of Eq. (1) is actually the generalized mean (GMean). The extended Boolean conjunction is related to the least distance from the most desirable point, which is (1, ..., 1):

$$s(\mathrm{AND}) = 1 - \sqrt[p]{\frac{1}{m}\sum_{i=1}^{m}(1 - s_i)^p}. \tag{2}$$

B. Passage-based Retrieval

An important consideration in passage-based retrieval is how passages are defined. Callan [2] considered three classes of passage types: discourse, semantic and window. Discourse passages are based on textual discourse units, e.g. sentences, paragraphs and sections. Semantic passages are based on the subject or content of the text. Window passages are based on a specified number of words, without regard to the logical structure of the documents. Some past works found that among these different types of passages, window passages are more effective than the others e.g. [2], [3]. In our current study we adopt the definition of passages as fixed-size text windows.

For window passages, both non-overlapping passages [42], [43] and overlapping passages [2], [3], [44], [45] have been studied in IR. Half-overlapping passages deal with the concern that relevant information may be split across two passages. Liu & Croft [46] found that half-overlapped passages are at least as effective and more efficient than other pre-defined passages. Some works have studied arbitrary passages, which are defined at query time and may start at any point. For example, Kaszkiel & Zobel [42] studied overlapping arbitrary passages. Na et al. [47] studied completely arbitrary passages, with no restriction on both the starting location and the size of each passage.

Passage-based retrieval has been studied for various retrieval models, such as the vector space model [3] and the language model [5], [46], [48]–[51]. Liu & Croft [46] examined passage retrieval in the language modeling framework and found passage retrieval to be effective compared with document-based retrieval, for either a simple language model [52] or a relevance model [53]. Various methods of combining passage scores to yield overall document score have been investigated. A commonly used heuristic is to equate the document ranking score to the maximum score of the constituent passages [3], [46], [54]. Another approach is taking the sum of scores of top-ranked passages [4]. Bendersky & Kurland [5], [50] introduced a probabilistic passage-based language model. They derived a document ranking function as an interpolation of a document-based score and a passage-based score, with the interpolation factor being related to an estimated measure of document homogeneity. This is an interpolation between the document scores and the max function of the passage scores. Since our method performs better than document scores (Table V) and using the max function (Table IV), we expect our method to be better than probabilistic passage-based language model of [5], [50] as the interpolation always lies between the document retrieval and the max of passage scores. Thus, we do not study this method in this work.

III. Our approach

This section presents our approach in detail. Section III-A discusses the relevance decision principles of our theoretical framework. The mathematical formulation and operationalization of these principles via fuzzy set theory are presented in Sections III-B and III-C. Section III-D describes some variations that we have adopted for passage-based retrieval.

A. Relevance decision principles

A longstanding golden standard in information retrieval evaluation is provided by TREC test collections with given sets of queries and lists of known relevant documents judged by human experts. The TREC evaluation policy for ad-hoc retrieval is that a document is regarded as relevant if any part of it is relevant [7]. Previously [1] we introduced a description of the retrieval process as mimicking a human scanning through a document, making judgments of the relevance of individual component parts within the document, and coming up with an overall judgment of the relevance of the document. Here we propose that to attain good retrieval performance, the judgments of relevance of the document components should be combined based on some general principles that are formulated according to the TREC evaluation policy. In particular, we have proposed two such relevance decision principles [1]:

> **Aggregated Relevance (AR) Principle:** The relevance of a document to a topic is higher if there is stronger accumulated evidence of relevance in the document components.

> **Disjunctive Relevance Decision (DRD) Principle:** A document is judged relevant to a topic if any part of the document is relevant to the topic.

It is apparent that the DRD principle conforms to the TREC evaluation policy in making an overall judgment of relevance of a document based on the judgment of its components. The AR Principle is also appropriate because TREC provides the relevance of judged documents in grades (assigning a score of 0, 1 or 2 to each document, with 0 indicating non-relevance and 2 indicating strong relevance). It is conceivable that a higher degree of relevance is supported by finding more indication of relevance from different parts of a document. The AR and DRD principles are independent and may be applied either separately or together in a judgment of relevance.

For passage-based retrieval, the passages within a document may be taken as the components in the above description. In order to show the effectiveness of combining passage scores based on the AR and DRD Principles, we also perform retrieval experiments (Section IV-C) following an alternative:

> **Conjunctive Relevance Decision (CRD) Supposition:** A document is judged relevant only if all component parts are relevant.

It is clear that CRD contradicts with DRD and does not follow the TREC evaluation policy. Here we call CRD a 'supposition' to distinguish it from the 'principles' that conform to TREC. The CRD is unrealistic and not appropriate because information related to the desired topic may appear only in some parts of a document rather than distributed across every component part. By testing the CRD, we confirm our conjecture that good performance is obtained by following principles that conform to the evaluation policy.

B. Mathematical formulation

For retrieval, the relevance decision principles (Section III-A) need to be operationalized in a mathematical framework. A variable $R(C_{d,i})$ denotes the relevance of $C_{d,i}$, which is the ith component in the document d. For example, $R(C_{d,i})$ may be a Boolean variable, with the values of 1 and 0 indicating relevance and non-relevance, respectively. Alternatively, $R(C_{d,i})$ may be a real number in the range [0,1], as in the fuzzy set framework that is described in Section III-C below. In passage-based retrieval, the component $C_{d,i}$ may be a passage in the document d. For a document that consists of n passages $C_{d,1}, \cdots, C_{d,n}$, the variable $\hat{R}(d)$ specifying the overall relevance of the whole document d is calculated by combining $R(C_{d,i})$ with $i = 1, \cdots, n$ via a function $\theta(\cdot)$ that conforms to the chosen relevance decision principle. The general formula for the overall relevance judgment of d is:

$$\hat{R}(d) = \theta(R(C_{d,1}), ..., R(C_{d,n})), \tag{3}$$

with $\theta(\cdot)$ being a n-ary function. In order to define appropriate forms of the function $\theta(\cdot)$, we first discuss some algebraic properties of $\theta(\cdot)$ that conform to the relevance decision principles. These properties are summarized in Table I.

TABLE I
PROPERTIES OF SCORE COMBINATION FUNCTIONS CONFORMING TO THE RELEVANCE DECISION PRINCIPLES

	Aggregated relevance (AR)
boundary condition	$\theta(0, \cdots, 0) = 0$ $\theta(1, \cdots, 1) = 1$
monotonicity	$\forall i \le n, a_i \le a'_i \Rightarrow \theta(a_1, \cdots, a_n) \le \theta(a'_1, \cdots, a'_n)$
	Disjunctive relevance decision (DRD)
boundary condition	$\theta(0, 0) = 0$ $\theta(1, 0) = \theta(0, 1) = \theta(1, 1) = 1$
commutativity	$\theta(a_1, a_2) = \theta(a_2, a_1)$
associativity	$\theta(\theta(a_1, a_2), a_3) = \theta(a_1, \theta(a_2, a_3))$

With the AR principle, the overall evidence of relevance of a document is based on accumulated evidence of the component parts. Therefore for AR, a basic property of $\theta(\cdot)$ is monotonicity, such that if any component part $a_i = R(C_{d,i})$ is ascribed a higher degree of relevance a'_i (i.e. $a'_i > a_i$), then the overall relevance is increased (Table I). The function $\theta(\cdot)$ also satisfies certain boundary conditions. First, if none of the component parts is relevant, the document is judged non-relevant, i.e. $\hat{R}(d) = \theta(0, \cdots, 0) = 0$. Second, if every component part is relevant, the maximum value $\hat{R}(d) = \theta(1, \cdots, 1) = 1$ is attained.

As the DRD (Table I) and CRD (Table II) reflect strong boundary conditions similar to the logical disjunction and conjunction, it is convenient to associate them with a binary function $\theta(a, b)$ for score combination. For DRD, if any component part is relevant ($R(C_{d,i}) = 1$), then the overall score is relevant, i.e. $\theta(1, 0) = \theta(0, 1) = \theta(1, 1) = 1$. For CRD, if any component part is non-relevant ($R(C_{d,i}) = 0$), then the overall score is non-relevant, i.e. $\theta(0, 0) = \theta(1, 0) = \theta(0, 1) = 0$. The binary function $\theta(\cdot)$ satisfies commutativity and associativity properties, as necessitated by the logical requirement that the same overall score is obtained regardless of the order in which the scores of the component parts are combined. With the associativity property, the overall score for combining more than two component parts can be obtained by applying the binary function in a chain. It should be noted that DRD and CRD do not require $\theta(\cdot)$ to be monotonic, unlike AR.

C. Combination of Passage Scores via Fuzzy Set Theory

In a fuzzy set theoretic approach for passage-based retrieval, the general n-ary functions that combine passage scores return

TABLE II
PROPERTIES OF SCORE COMBINATION FUNCTIONS CONFORMING TO THE CRD SUPPOSITION

Conjunctive relevance decision (CRD)	
boundary condition	$\theta(0,0)=\theta(1,0)=\theta(0,1)=0$ $\theta(1,1)=1$
commutativity	$\theta(a_1,a_2)=\theta(a_2,a_1)$
associativity	$\theta(\theta(a_1,a_2),a_3)=\theta(a_1,\theta(a_2,a_3))$

a real number in the range [0,1], i.e. $\theta(a_1,...,a_n):\mathbb{R}[0,1]^n \rightarrow \mathbb{R}[0,1]$. These functions need to satisfy the basic algebraic properties conforming to the chosen relevance decision principle (Table I). Here we discuss several possible choices provided by fuzzy set theory (Sections III-C1, III-C2) and the normalization issue in using the functions (Section III-C3).

1) t-conorm and t-norm operators: In fuzzy set theory, the generalization of the logical disjunction and conjunction is given by two types of binary operators, namely t-conorms $S(a,b)$ and t-norms $T(a,b)$, respectively [55]. For a and b in the range [0,1], $T(a,b)$ is also in the range [0,1] and the dual t-conorm is given by $S(a,b)=1-T(1-a,1-b)$. The duality means that $T(a,b)$ and $S(a,b)$ satisfy De Morgan's law: $1-T(a,b)=S(1-a,1-b)$, with $1-a$ corresponding to negation in fuzzy set theory. The t-conorm and t-norm operators differ by their boundary conditions: $S(a,0)=a$, $S(a,1)=1$, and $T(a,0)=0$, $T(a,1)=a$, respectively. The boundary condition $S(a,1)=1$ of the *t-conorm* conforms to the DRD principle that if any component passage is relevant (score equals to 1), then the document is considered to be relevant. As for the *t-norm*, the boundary condition $T(a,0)=0$ conforms to the CRD supposition which implies that if any component passage is non-relevant (score equals zero), then the document is considered as non-relevant. Furthermore, the t-conorm and t-norm operators satisfy monotonicity, commutativity and associativity [55]. By satisfying these properties, the t-conorm and t-norm operators also conform to DRD and CRD respectively. Therefore, the t-conorm (related to the disjunctive 'exists' semantic) and t-norm (related to the conjunctive 'forall' semantic) are suitable candidates as the function $\theta(\cdot)$ to operationalize DRD and CRD, respectively. Note that the monotonicity property of the t-conorm and t-norm is not required for DRD and CRD. This additional property means that the t-conorm and t-norm conform to the AR principle as well, besides the DRD principle and CRD supposition respectively. While the t-norms and t-conorms are defined as binary operators, they can be applied recursively to combine the scores of multiple passages by their associativity (see Fig. 1).

We have tested a large number of well-known t-norm and t-conorm operators for the combination of passage scores in our experiments, including: the Dombi, Yager, Schweizer-Sklar, Sugeno-Weber, Hamacher, Frank, probabilistic and max operators [10], [56]. Many of these operators are parameterized, in terms of a free parameter $p \geq 1$. For example, the Dombi t-conorm $S_{Dombi}(a,b)$ is given by:

$$S_{Dombi}(a,b)=\frac{1}{1+\left(\left(\frac{1}{a}-1\right)^{-p}+\left(\frac{1}{b}-1\right)^{-p}\right)^{-1/p}} \quad (4)$$

2) Generalized mean: The generalized mean (GMean) is an averaging aggregation operator of fuzzy set theory that satisfies the monotonicity property as well as the boundary conditions of the AR principle (Table I). For a document d consisting of m passages with *non-zero* passage scores, s_i, the GMean of the passage scores is exactly the normalized p-norm of Eq. (1). As GMean is monotonic non-decreasing, it conforms to the AR principle that a higher overall score is obtained if the passages show more evidence of relevance.

In the GMean, excluding from the sum in Eq. (1) any passage in d that has zero score deals with the problem of long documents that may contain passages without any query term. This implementation of GMean differs from that in our previous work [1] where all passages, including those with zero score, are included in the sum. For $p=1$, the GMean as given by Eq. (1) is equal to the arithmetic mean. For normalized passage scores, in the limit of $p \rightarrow \infty$, GMean becomes $\max(\cdot)$.

Apart from the extended Boolean disjunction (OR) (Eq. (1)), the extended Boolean conjunction (AND) (Eq. (2)) also satisfies the AR boundary condition and monotonicity property. For large values of p in Eq. (2), the extended Boolean conjunction behaves like $\min(\cdot)$, which is a t-norm that conforms to the CRD supposition.

3) Passage score normalization: Because the arguments of the fuzzy operators are in the range [0,1], it is necessary to apply a monotonic mapping on passage scores to convert their values to the [0,1] range. Monotonic mapping of the passage scores means that their order of ranking is preserved. In this study we have used the following passage score normalization:

$$s_i=\gamma\cdot\left(\frac{\tilde{s}_i-MinScore}{MaxScore-MinScore}\right), \quad (5)$$

where $\tilde{s}_i$ is the score of the ith passage, s_i is the normalized score, $MinScore$ and $MaxScore$ are the minimum and maximum score among all retrieved passages respectively, while γ is a constant parameter, with $0<\gamma\leq 1$.

Eq. (5) represents our novel generalized version of the min-max normalization. In this equation, setting the scaling factor γ to a value less than 1 can avoid the score reaching the boundary condition $a=1$. Without this factor, the passage with the maximum score $s_i=MaxScore$ becomes normalized to $s_i=1$, so that by fuzzy disjunction the combined score for its parent document will be 1. Thus, including the γ factor gives the flexibility that the parent document containing the passage with the MaxScore does not necessarily rank first, but its other constituent passages are considered as well. In this study, we have set $\gamma=0.3$.

D. Novel Variations of the IR models

We discuss some variations in our passage-based retrieval, with regard to the retrieval models (Section III-D1) and pseudo-relevance feedback (Section III-D2).

1) Retrieval models: We have tested our approach on two highly effective retrieval models, namely the PL2 [13], [14] and BM25 [12]. Some variations are applied in our implementation of the retrieval models for passage-based retrieval. In general, the ranking formulae of these models (e.g. see [14]) are applied to passages instead of documents. Hence in these formulae, the term frequency tf is the count of terms in each passage, and the document length becomes passage length. The PL2 model (Eq. (3) of [14]) contains a parameter $\lambda = F(q)/N$ = frequency of query term q in the collection / number of documents in the collection. We have also tested the passage-based version of the parameter, with N being the total number of passages instead of documents in the collection. We find t hat o ur p assage-based v ariations t o p erform b etter than the document-based versions.

2) Pseudo-relevance Feedback (PRF): PRF is an established method to enhance retrieval effectiveness. Typically PRF involves query expansion (QE), whereby after an initial retrieval with the original query, a number of terms are automatically extracted from the top ranked N_{PRF} documents and added to the query for a second retrieval. For PRF in passage-based retrieval, we assume the top N_{PRF} passages rather than documents as being relevant. Query expansion terms are selected from these top ranked passages, as in [44]. Selecting query expansion terms from passages rather than whole documents is expected to be more effective, as they are more likely to be terms appearing near the original query terms and thus related to the desired topic.

For QE term selection, terms in the top ranked N_{PRF} documents/passages of the initial retrieval are assigned a score according to a term scoring function (e.g. [57]). Generally a total of N_{QE} terms having the highest scores are included in the vector of QE terms, $\vec{Q}_{QE}$. An expanded query vector $\vec{Q}_{PRF}$ is obtained by mixing the original query $\vec{Q}$ and $\vec{Q}_{QE}$:

$$\vec{Q}_{PRF} = \alpha_m \frac{\vec{Q}}{|\vec{Q}|} + (1 - \alpha_m) \frac{\vec{Q}_{QE}}{|\vec{Q}_{QE}|}, \tag{6}$$

where $|\cdot|$ is the total number of terms and α_m is a mixing factor with a value between 0 and 1. A second retrieval, i.e. a re-retrieval [58] is performed by calculating new scores for all passages in the collection with the expanded query $\vec{Q}_{PRF}$. In our experiments, the parameters α_m and N_{QE} are determined by calibration to yield the highest performance metric.

We use the following QE term scoring function (PRF-1):

$$score_{PRF-1}(w, d) = tf(w, d) \times \log_{10}(N/df(w)), \tag{7}$$

where $tf(w, d)$ is the term frequency of the word w in document d, $df(w)$ is the document frequency of w (i.e. the number of documents in the collection that contains w) and N is the number of documents in the collection.

IV. Experiments

This section presents our experiments of passage-based document retrieval. The general experiment setup is described in Section IV-A. In Section IV-B, we briefly discuss the parameterization of the aggregation operators in passage score combination. Section IV-C contains the the main experimental results of the evaluation of our approach with a comparison against various baselines of document-based retrieval.

TABLE III
Some statistics of the test collections and the corresponding sets of topics used in our experiments

	Disks 4&5	WT10g		GOV2		
TREC	6	9	10	T-2004	T-2005	T-2006
Topics	301-350	451-500	501-550	701-750	751-800	801-850
Doc. types	news; reports	webpages		webpages		
No. of doc.	556,075	1,692,096		25,205,179		
Av.doc.len.	165.7	190.7		210.7		
Size (GB)	3.27	10		426		

Note: TREC-9 and TREC-10 are run on the WT10g collection; Terabyte tracks of 2004, 2005 and 2006 (denoted by T-2004, T-2005 and T-2006, respectively), are run on GOV2. Average document length (Av.doc.len.) is measured after stopword removal.

A. Experiment Setup

Retrieval experiments are performed in our own retrieval system on several TREC test collections (Table III) with given sets of queries and documents. The TREC collections constitute a longstanding golden standard in information retrieval evaluation. The TREC-6 (Disks 4&5) collection consists of texts of a wide range of sizes, including newsfeeds which may be short, as well as very long congress reports. The presence of both long and short texts makes this collection suitable for testing passage-based retrieval. GOV2 and WT10g are collections of webpages which do not include spam. The collection sizes range from 3G bytes for TREC6 to about 0.5 Terabytes of GOV2, thus showing the scalability of our approach and suggesting the possibility to scale to further larger sizes. If the max operator is used or in the case of a retrieval model with normalized ranking score (e.g. BM25), passage score normalization (Eq. (5)) can be skipped and not much additional processing is incurred, so that fuzzy passage-based retrieval should be scalable. For very large datasets most search engines in practice scale out, distributing the data over more servers rather than scale up, so that the retrieval time can be traded off by using more servers.

We perform both an initial retrieval with the original queries, and pseudo-relevance feedback (PRF) with query expansion. For the initial retrieval we use title queries, each with about 2 to 3 query terms on average (e.g. [43]), as such query lengths are typically used in web searches [59]. We first use the 50 title queries of Terabyte Track 2006 on the GOV2 collection as the training set to calibrate the various retrieval model parameters and the parameters of PRF. Calibration is performed by a grid search for each of the parameters [43]. Using the calibrated parameters, retrieval is performed on the queries of Terabyte 2004 and Terabyte 2005 tracks on the GOV2 collection, as well as the queries of TREC-9 and TREC-10 on the WT10g collection for testing. Due to the different nature of the TREC-6 collection, a separate calibration is performed.

We have adopted the above methodology of calibration, instead of cross-validation, for the following reasons. First, our method is a more realistic approach for practical retrieval systems, which are generally calibrated beforehand for retrieval with any unseen query, as cross-validation is not possible. Second, our methodology uses less training data than cross-validation, so that it should be a stronger test of the effectiveness of a retrieval method. Third, our methodology allows the retrieval performance of both the training set and testing sets of queries to be examined, thus enabling checking whether good performance of a trained model can generalize to good results in testing using fixed parameters.

In this study, retrieval effectiveness is measured by the standard metric Mean Average Precision (MAP), which is a composite measure of precision and recall. The MAP is chosen, as commonly used in the literature, because it is a robust measure with values that past research has found to be difficult to improve. Following the common TREC evaluation environment, all the MAP values are calculated for a ranked list of 1000 documents retrieved for each query. MAP has a value between 0 and 1, with 1 indicating perfect retrieval.

We have performed passage-based retrieval using the successful PL2 and BM25 models, comparing the results with the baselines of these models in document-based retrieval. We have checked that our baseline values are comparable as those in the literature. For example for an initial retrieval with 100 queries (Topics 451-550) on WT10g, our document-based BM25 yields a MAP value of 0.2084, while [60] obtains a value of 0.2055 using BM25 with two-fold cross validation.

For passage-based retrieval, we have used window passages of fixed sizes, as these were found to be effective in past studies [2], [3]. We have tested both non-overlapping and half-overlapping passages, and found that they generally attain similar retrieval performance for the various retrieval models. Therefore in this article, we only present results based on non-overlapping passages, which are defined by contiguous windows. The first passage starts from the first word of a document. Stopword removal and Porter's stemming [61] are applied to the documents before the passages are defined. In our experiments, we have tested passage sizes in the range of 150 words to 750 words. The retrieval results presented in Tables IV and V correspond to the best MAP values obtained with the tested passage sizes.

B. p Parameter in Aggregation Operators

We examine the combination of passage scores using various aggregation operators of fuzzy set theory (Section III-C), i.e. t-conorms, t-norms and generalized mean (GMean). From the definition of the Dombi t-conorm in Eq. (4), $S_{Dombi}(a, b)$ approximates the $\max(a, b)$ function as the parameter $p \to \infty$. Hence, the common method of taking the maximum passage score in passage-based retrieval [2], [3] is equivalent to setting a large p value in the Dombi t-conorm. We have examined the sensitivity to the p parameter in passage-based retrieval by testing with the the PL2 model using Dombi t-conorm for passage score aggregation. With all passage sizes, for both an initial retrieval and PRF retrieval, MAP is generally low for p smaller than about 6. For large passage sizes (say above 450 words), MAP increases with larger values of p, with MAP saturating for $p > 12$. As the Dombi t-conorm approximates the maximum operator at large p values, for large passage sizes the Dombi t-conorm performs similarly as the maximum operator. However at small passage sizes (say 250 words or below), a peak in MAP occurs at a value of p typically between 6 and 12. Generally with small passage sizes, using the Dombi t-conorm operator with a free p parameter can potentially give better retrieval results than using the max function.

For GMean (Eq. (1)), the special case $p = 1$ is the simple arithmetic mean. Therefore, allowing variable p values mean that our approach is more general and covers the method of combining passage scores by the arithmetic mean, which is an approach considered by others [5]. We find that generally the best results are obtained with larger values of p (about 24), so that using the GMean is able to perform better than ranking documents according to the average passage score.

C. Evaluation of Our Approach

This section describes the experimental evaluation of our passage-based document retrieval. In Section IV-C1 we examine the performance of passage score combination by various fuzzy set aggregation operators, based on retrieval with the PL2 model. We then test the generality of the approach by experiments with another retrieval model, namely the BM25 (Section IV-C2). In Section IV-C3 we compare our passage-based retrieval results with the PL2 and BM25 models in document-based retrieval.

1) Comparison of aggregation operators: We examine the combination of passage scores using several types of fuzzy set aggregation operators. In particular, the generalized mean (GMean) (Eq. (1)) operationalizes the AR principle because it satisfies the monotonicity property and boundary conditions of AR (Table I). As discussed in Section III-C1, t-conorm and t-norm operators conform to the DRD and CRD respectively. In addition, the t-conorm and t-norm conform to the AR principle on account of their monotonicity.

We have tested various t-conorm and t-norm operators mentioned in Section III-C1. We find that using any of the t-norm operators categorically yields poor retrieval results. This suggests following the CRD supposition is not effective, as it is inconsistent with the TREC evaluation policy for ad-hoc retrieval that regards a document as relevant if any part of it is relevant [7]. Due to a lack of space, we omit a detailed report of results for t-norms here.

Both the AR and DRD principles are consistent with the TREC evaluation policy. Therefore we compare the effectiveness of passage score combination using t-conorm, which conforms to AR and DRD, and GMean, which conforms to AR. The results of retrieval by the PL2 model using these operators for passage score combination is summarized in Table IV, which shows the best MAP values obtained for each track, with the corresponding passage size.

Table IV shows that for both the initial retrieval and the PRF retrieval on all sets of queries and test collections, the highest MAP is obtained with either the Dombi t-conorm (DombiD) operator or the GMean. We find by the randomization test [62] that the difference between DombiD and GMean is not statistically significant in the majority of cases. Therefore, it is difficult to conclude whether using Dombi t-conorm or the

TABLE IV
COMPARISON OF RETRIEVAL PERFORMANCE BASED ON VARIOUS WAYS OF PASSAGE SCORE COMBINATION, FOR (A) THE INITIAL RETRIEVAL, AND (B) PSEUDO-RELEVANCE FEEDBACK.

		Disks 4&5	WT10g		GOV2		
	TREC	6	9	10	T-2004	T-2005	T-2006
(a) Initial retrieval							
DombiD	passage size	450	650	650	650	650	650
	MAP	.2556	.2138	.2197	.2801	**.3372***	.3224
YagerD	passage size	450	650	650	650	650	650
	MAP	.2551	.2150	.2196	.2797	.3364*	.3243
SSD	passage size	450	650	650	650	650	650
	MAP	.2554	.2150	.2196	.2797	.3364*	.3243
max	passage size	450	650	650	650	650	650
	MAP	.2558	.2152	.2179	.2776	.3305	.3222
GMean	passage size	450	650	650	650	650	650
	MAP	**.2559**	**.2166**	**.2327**	**.2810**	.3348	**.3254**
(b) PRF retrieval							
DombiD	passage size	200	550	550	550	550	550
	MAP	**.2937**†	.2161	.2397	**.3171***	**.3872***	.3681
YagerD	passage size	250	450	450	450	450	450
	MAP	.2858	.2232	.2394	.3115	.3748	.3687
SSD	passage size	200	650	650	650	650	650
	MAP	.2804	.2140	.2425	.3081	.3766	.3700
max	passage size	250	650	650	650	650	650
	MAP	.2852	.2090	.2430	.3066	.3724	.3709
GMean	passage size	200	650	650	650	650	650
	MAP	.2834	**.2215**	**.2438**	.3102	.3773	**.3770**

Note: DombiD, YagerD, and SSD denote the Dombi, Yager and Schweizer-Sklar t-conorm operators, respectively. The MAP values correspond to the best passage size as indicated, for retrieval with the PL2 model. Within (a) initial retrieval and (b) PRF retrieval, statistically significant difference (95% confidence level, randomization test) is indicated: over the corresponding retrieval using the max operator (*), and over retrieval using generalized mean (†); absence of these symbols means that the corresponding difference is not statistically significant.

GMean is better. This is consistent with the premise that both DRD and AR principles satisfy the TREC evaluation policy.

Table IV indicates that while better retrieval results are generally obtained with the Dombi t-conorm or GMean, the max function can be a good approximation. We also find by the randomization test that the difference between the max function and the best operator is statistically significant in some cases as indicated in the table, but not always. As the max function does not require the normalization procedure as do the other methods, it is computationally less demanding and can be a suitable choice if time efficiency is a concern.

The GMean corresponds to the extended Boolean disjunction (OR) model (Eq. (1)). Apart from GMean, we have also tested passage score combination by the function corresponding to the extended Boolean conjunction (AND) model, Eq. (2). We found that it yields rather poor retrieval results for both TREC-6 (Topics 301-350) and GOV2 2006 Terabyte track (Topics 801-850), with almost no relevant document retrieved for many queries. For these tracks, extended Boolean conjunction yielded MAP values of 0.0423 and 0.0059 respectively, compared with 0.2559 and 0.3254 respectively obtained with GMean. Thus, extended Boolean disjunction is preferred over extended Boolean conjunction for passage score combination. Since the extended Booloean conjunction behaves close to the $\min(\cdot)$, which is a t-norm conforming to CRD, the result also indicates that following the CRD is not effective.

The retrieval results for some t-conorm operators are included in Table IV. For the other t-conorm operators that we have tested (Sugeno-Weber, Hamacher, Frank and probabilistic [10], [56]), we find the retrieval performance to be rather poor. In particular, the 'probabilistic' t-conorm corresponds to the assumption that passages are independent. The poor result for the probabilistic operator suggests that the independence assumption is incorrect, as two passages within a document containing the same query terms are likely to be related. This shows that the more general approach based on fuzzy set theory with more choices of aggregation operators is better than the simple probabilistic approach.

2) Testing a Different Retrieval Model for Generality: Having demonstrated the effectiveness of fuzzy set aggregation operators in passage-based retrieval using the PL2 model, in order to test the generality of the approach we further perform experiments with another retrieval model, namely the BM25 model. For the selection of query expansion terms in PRF retrieval, we use the same scoring function PRF-1 (Eq. (7)) as used with the PL2 model. The best MAP values obtained with BM25 for each track are shown in Table V. Same as the retrieval with the PL2 model, the best MAP values obtained by BM25 with the Dombi t-conorm (DombiD) and GMean are quite similar, with either yielding the higher MAP value in different tracks. Thus the results for BM25 are also consistent with the DRD and AR principles satisfying the TREC evaluation policy.

3) Comparison with document-based retrieval: We have compared the effectiveness of our passage-based retrieval against the PL2 and BM25 models in document-based retrieval, which are common baselines in the literature (e.g. [63]). The results are summarized in Table V, which shows the best MAP values obtained for the various tracks.

Table V indicates that for both an initial retrieval and PRF retrieval, our passage-based retrieval generally performs better (with numerically higher MAP values) than document-based retrieval. For the initial retrieval (Table V(a)), passage-based retrieval yield MAP values that are about 0.1 to 2 percentage points higher than the corresponding document-based retrieval values. For PRF retrieval (Table V(b)), the better performance of passage-based retrieval is more obvious, with higher MAP values by about 2.5 to 4 percentage points. One reason for the better PRF performance with passage-based retrieval is that with more query terms in PRF, scoring based on passages is more indicative that these query terms occur in a closer neighborhood and are thus related, rather than spread out over the whole document. Another reason is that by selecting query expansion terms from top-ranked passages containing the given query terms, rather than from over the entire document, the selected terms may be nearer to the original query terms and hence more likely to be related to the given query. Thus, noise terms are more likely to be avoided by using passages.

In Table V(a) and (b), the cases where the MAP value is higher than the document-based PL2 and BM25 values with statistical significance are indicated by the subscripts p and b, respectively. For PRF (Table V(b)), our PL2 model in passage-based retrieval using Dombi t-conorm can attain better

TABLE V
COMPARISON OF THE BEST MAP VALUES OBTAINED WITH OUR IMPLEMENTATION OF VARIOUS MODELS

		Disks 4&5	WT10g		GOV2		
	TREC	6	9	10	T-2004	T-2005	T-2006
(a) Initial retrieval							
Our PL2	DombiD	.2556	.2138	$.2197_{pb}$	.2801	**.3372**	$.3224_{pb}$
	GMean	$\mathbf{.2559}_{p}$	**.2166**	$\mathbf{.2327}_{pb}$	**.2810**	.3348	$\mathbf{.3253}_{pb}$
Our BM25	DombiD	.2522	.2141	$.2165_{b}$	.2692	.3355	$.3194_{b}$
	GMean	.2536	.2094	$.2206_{b}$	.2748	.3291	$.3221_{b}$
Doc-based PL2	-	.2494	.2119	.2050	.2676	.3340	.3072
Doc-based BM25	-	.2483	.2157	.2011	.2656	.3279	.3044
(b) PRF retrieval							
Our PL2	DombiD	$\mathbf{.2936}_{pb}$	.2161	$.2397_{pb}$	$\mathbf{.3171}_{p}$	.3872	$.3681_{pb}$
	GMean	.2834	.2215	$.2438_{pb}$	.3102	.3773	$.3770_{pb}$
Our BM25	DombiD	.2918	**.2292**	$\mathbf{.2556}_{pb}$	.3054	.3893	$.3663_{pb}$
	GMean	.2842	.2199	$.2496_{pb}$	.3050	**.3899**	$\mathbf{.3790}_{pb}$
Doc-based PL2	-	.2655	.2078	.2066	.2743	.3676	.3152
Doc-based BM25	-	.2607	.2180	.2112	.2845	.3616	.3180

Note: 'Our PL2' and 'Our BM25' indicate our implemented models in passage-base retrieval. Passage scores are combined by Dombi t-conorm (DombiD) or generalized mean (GMean). p and b denotes statistically significant difference over document-based PL2 and BM25 models, respectively (95% confidence level); absence of these symbols means that the corresponding difference is not statistically significant.

MAP than the corresponding document-based PL2 model with statistical significance, in four out of six of the tested tracks, with TREC-9 (WT10g) and Terabyte-2005 (GOV2) being the exceptions. A reason for different behavior of the two tracks may be that the TREC-9 queries are hard (as seen by the comparatively low MAP of 0.2138 for the initial retrieval) and the Terabyte-2005 queries are easy (with high MAP of 0.3372 for the initial retrieval), so that in both cases it is more difficult to distinguish performance differences. The results confirm the effectiveness of our passage-based retrieval approach.

V. CONCLUSION

We have tackled the question of how to achieve effective passage-based retrieval by establishing a novel principled fuzzy set theoretical framework. The importance of a principled framework is that it states the criteria for passage score combination, thus providing a guideline for finding effective methods. This contrasts with the existing paradigm of using heuristics to combine passage scores, without knowing why these methods are effective or how better methods may be found. We also introduce a novel methodology of formulating the appropriate principles according to the specific evaluation requirement, such as the TREC evaluation policy [7] in the current case. While the effectiveness of this methodology is demonstrated for passage-based retrieval, the approach is general and may be adopted for building other applications.

We show that fuzzy set aggregation operators are appropriate to operationalize the relevance decision principles in our theoretical framework for practical retrieval. In particular, the generalized mean (GMean) is consistent with Aggregated Relevance (AR) principle, while the t-conorm operator is consistent with both the Disjunctive Relevance Decision (DRD) and AR principles, which conform to the TREC evaluation policy. Our framework theoretically justifies the common heuristic of taking the maximum passage score as the overall document score, as the maximum operator belongs to the class of t-conorm operators. We find the maximum operator to be only a near best, with better performance being obtained by the Dombi t-conorm (DombiD) or the GMean. While our empirical study demonstrates retrieval effectiveness using a selection of aggregation operators, the significance of our framework is that many other aggregation operators of fuzzy set theory may also be applied, potentially yielding even better performance. This work thus signifies a novel and effective utilization of fuzzy set theory in a widely used application.

While the prevailing paradigm employs retrieval models with built-in document length normalization to solve the problem of varying document lengths, we have shown that this is not sufficient. Our experiments show that our passage-based retrieval performs better than the document-based retrieval baselines, both in an initial retrieval with the given title queries and more obviously in PRF retrieval with an expanded query. In PRF, choosing query expansion terms within passages allows a focused selection of terms more likely to be related to the query, while with document-based retrieval noise terms are more likely to be selected. Thus, better performance is attained by a combination of document length normalization and the passage-based retrieval of our theoretical framework.

Another significance of our framework is its generality, as it can adopt any retrieval model that returns a ranking score to passage-based retrieval. Thus, other researchers can easily adopt our framework to their retrieval systems, no matter what retrieval model is used. This differs from other works, such as [5] which only applies to the language model. While we have demonstrated the framework with the strong PL2 and BM25 models, it may be used with other recent models that show promising results (e.g. [64]). It is of interest to test whether passage-based retrieval using these models, especially in the PRF setting, can further enhance retrieval performance.

REFERENCES

[1] Y. K. Kong, R. W. P. Luk, W. Lam, K. S. Ho, and F. L. Chung, "Passage-based retrieval using parameterized fuzzy set operators," in *ACM SIGIR Workshop on Mathematical/Formal Methods for Information Retrieval*, 2004.

[2] J. P. Callan, "Passage-level evidence in document retrieval," in *Proceedings of the 17th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 1994, pp. 302–310.

[3] M. Kaszkiel and J. Zobel, "Passage retrieval revisited," in *Proceedings of the 20th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 1997, pp. 178–185.

[4] M. A. Hearst and C. Plaunt, "Subtopic structuring for full-length document access," in *Proceedings of the 16th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 1993, pp. 59–68.

[5] M. Bendersky and O. Kurland, "Utilizing passage-based language models for ad hoc document retrieval," *Information retrieval*, vol. 13, pp. 157–187, 2010.

[6] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, "Interpreting tf-idf weights as making relevance decisions," *ACM Transactions on Information Systems*, vol. 26, no. 3, Article 13, pp. 1–37, 2008.

[7] TREC home page, "Data – english relevance judgements, http://trec.nist.gov/data/reljudge_eng.html."

[8] G. Beliako and J. Warren, "Appropriate choice of aggregation operators in fuzzy decision support systems," *IEEE Transactions on Fuzzy Systems*, vol. 9, no. 6, pp. 773–784, 2001.

[9] R. R. Yager, "Noble reinforcement in disjunctive aggregation operators," *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 6, pp. 754–767, 2003.

[10] E. P. Klement, R. Mesiar, and E. Pap, *Triangular Norms*. Kluwer Academic Publishers, 2000.

[11] A. Singhal, C. Buckley, and M. Mitra, "Pivoted document length normalization," in *Proceedings of the 19th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 1996, pp. 21–29.

[12] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval," in *Proceedings of the 17th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 1994, pp. 232–241.

[13] G. Amati and C. J. van Rijsbergen, "Probabilistic models of information retrieval based on measuring the divergence from randomness," *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 357–389, 2002.

[14] B. He and I. Ounis, "Parameter sensitivity in the probabilistic model for ad-hoc retrieval," in *Proceedings of the 16th ACM Conf. on Information and Knowledge Management (CIKM'07)*, 2007, pp. 263–272.

[15] Z. Xu and Q. Da, "An overview of operators for aggregating information," *International Journal of Intelligent Systems*, vol. 18, pp. 953 –969, 2003.

[16] G. Bordogna and G. Pasi, "Personalised indexing and retrieval of heterogeneous structured documents," *Information Retrieval*, vol. 8, pp. 301–318, 2005.

[17] R. R. Yager, "On ordered weighted averaging aggregation operators in multi-criteria decision making," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 18, no. 1, pp. 183–190, 1988.

[18] V. Torra, "Owa operators in data modeling and reidentification," *IEEE Transactions on Fuzzy Systems*, vol. 12, no. 5, pp. 652–660, 2004.

[19] R. Mesiar, L. Šipeky, P. Gupta, and J. LeSheng, "Aggreation of owa operators," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 1, pp. 284–291, 2018.

[20] C. Paice, "Soft evaluation of boolean search queries in information retrieval systems," *Information Technology: Research and Development*, vol. 3, no. 1, pp. 33–42, 1984.

[21] W. G. Waller and H. Kraft, Donald, "A mathematical model of a weighted boolean retrieval system," *Information Processing and Management*, vol. 15, pp. 235–245, 1979.

[22] R. R. Yager and D. P. Filev, "Induced ordered weighted averaging operators," *IEEE Transactions on Systems, Man, and Cybernetics — Part B: Cybernetics*, vol. 29, no. 2, pp. 141–150, 1999.

[23] W. Dong and F. Wong, "Fuzzy weighted averages and implementation of the extension principle," *Fuzzy Sets and Systems*, vol. 21, pp. 183–199, 1987.

[24] T.-S. Liou and M.-J. Wang, "Fuzzy weighted average: an improved algorithm," *Fuzzy Sets and Systems*, vol. 49, pp. 307–315, 1992.

[25] M. Casanovas and J. M. Merigó, "Fuzzy aggregation operators in decision making with dempster-shafer belief structure," *Expert Systems with Applications*, vol. 39, pp. 7138–7149, 2012.

[26] S.-J. Chen and S.-M. Chen, "Fuzzy information retrieval based on geometric-mean averaging operators," *Computers and Mathematics with Applications*, vol. 49, pp. 1213–1231, 2005.

[27] R. R. Yager and Z. Xu, "The continuous ordered weighted geometric operator and its application to decision making," *Fuzzy sets and systems*, vol. 157, pp. 1393–1402, 2006.

[28] P. Wu, S. Liu, L. Zhou, and H. Chen, "A fuzzy group decision making model with trapezoidal fuzzy preference relations based on compatibility measure and cowga operator," *Applied Intelligence*, vol. 48, no. 1, pp. 46–67, 2018.

[29] R. R. Yager, "On generalized bonferroni mean operators for multi-criteria aggregation," *International Journal of Approximate Reasoning*, vol. 50, pp. 1279–1286, 2009.

[30] G. Bordogna and G. Pasi, "Linguistic aggregation operators of selection criteria in fuzzy information retrieval," *International Journal of Intelligent Systems*, vol. 10, pp. 233–248, 1995.

[31] L. Zhou and H. Chen, "A generalization of the power aggregation operators for linguistic environment and its application in group decision making," *Knowledge-Based Systems*, vol. 26, pp. 216–224, 2012.

[32] P. Wu, Q. Wu, L. Zhou, H. Chen, and H. Zhou, "A consensus model for group decision making under trapezoidal fuzzy numbers environment," *Neural Computing and Applications*, vol. 31, pp. 377–394, 2019.

[33] Z. Zhang, "Hesistant fuzzy power aggregation operators and their application to multiple attribute group decision making," *Information Sciences*, vol. 234, pp. 150–181, 2013.

[34] P. Wu, L. Zhou, H. Chen, and Z. Tao, "Additive consistency of hesitant fuzzy linguistic preference relation with a new expansion principle for hesitant fuzzy linguistic term sets," *IEEE Transactions on Fuzzy Systems*, vol. 27, no. 4, pp. 716–730, 2019.

[35] G. Salton, E. A. Fox, and H. Wu, "Extended boolean information retrieval," *Communications of the ACM*, vol. 11, no. 26, pp. 1022–1036, 1983.

[36] V. Cross, "Fuzzy information retrieval," *Journal of Intelligent Information System*, no. 3, pp. 29–56, 1994.

[37] C. Meghini, F. Sebastiani, and U. Straccia, "A model of multimedia information retrieval," *Journal of the ACM*, vol. 48, no. 5, pp. 909–970, 2001.

[38] S. E. Robertson, "On the nature of fuzz: A diatribe," *Journal of the American Society for Information Science*, vol. 29, pp. 304–307, 1978.

[39] M. H. Kim, J. H. Lee, and Y. J. Lee, "Analysis of fuzzy operators for high quality information retrieval," *Information Processing Letters*, vol. 46, pp. 251–256, 1993.

[40] J. H. Lee, W. Y. Kim, K. M. Ho, and L. Y. Joon, "On the evaluation of boolean operators in the extended boolean retrieval framework," in *Proceedings of the 16th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 1993, pp. 291–297.

[41] J. H. Lee, "Properties of extended boolean models in information retrieval," in *Proceedings of the 17th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 1994, pp. 182–190.

[42] M. Kaszkiel and J. Zobel, "Effective ranking with arbitrary passages," *Journal of the American Society for Information Science and Technology*, vol. 52, no. 4, pp. 344–364, 2001.

[43] E. K. F. Dang, R. W. P. Luk, J. Allan, K. S. Ho, S. C. F. Chan, K. F. L. Chung, and D. L. Lee, "A new context-dependent term weight computed by boost and discount using relevance information," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2514–2530, 2010.

[44] J. Alan, "Relevance feedback with too much data," in *Proceedings of the 18th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 1995, pp. 337–343.

[45] X. Li and Z. Zhu, "Enhancing relevance models with adaptive passage retrieval," in *Proceedings of the 30th European Conference on Information Retrieval (ECIR 2008)*, 2008, pp. 463–471.

[46] X. Liu and W. B. Croft, "Passage retrieval based on language models," in *Proceedings of the 11th ACM Conf. on Information and Knowledge Management (CIKM'02)*, 2002, pp. 375–382.

[47] S.-H. Na, I.-S. Kang, Y.-H. Lee, and J.-H. Lee, "Completely-arbitrary passage retrieval in language modeling approach," in *Proceedings of the 4th Asia information retrieval symposium (AIRS)*, 2008, pp. 22–33.

[48] Y. Lv and C. X. Zhai, "Positional language models for information retrieval," in *Proceedings of the 32th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2009, pp. 299–306.

[49] ——, "Positional relevance model for pseudo-relevance feedback," in *Proceedings of the 33rd Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2010, pp. 579–586.

[50] M. Bendersky and O. Kurland, "Utilizing passage-based language models for document retrieval," in *Proceedings of the European Conference on Information Retrieval (ECIR 2008)*. Springer Berlin Heidelberg, 2008, pp. 162–174.

[51] E. Krikon, O. Kurland, and M. Bendersky, "Utilizing inter-passage and inter-document similarities for reranking search results," *ACM Transactions on Information Systems*, vol. 29, no. 1, Article 3, pp. 1–28, 2010.

[52] J. M. Ponte and W. B. Croft, "A language modeling approach in information retrieval," in *Proceedings of the 21st Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 1998, pp. 275–281.

[53] V. Lavrenko and W. B. Croft, "Relevance-based language model," in *Proceedings of the 24th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2001, pp. 120–127.

[54] G. Salton, J. Allan, and C. Buckley, "Approaches to passage retrieval in full text information systems," in *Proceedings of the 16th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 1993, pp. 49–58.

[55] W. Pedrycz and F. Gomide, *An Introduction to Fuzzy Sets*. The MIT Press, 1998.

[56] H. J. Zimmermann, *Fuzzy set theory — and its applications*. Kluwer Academic Publishers, 2001.

[57] D. Harman, "Relevance feedback revisited," in *Proceedings of the 15th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 1992, pp. 1–10.

[58] E. K. F. Dang, R. W. P. Luk, and J. Allan, "Fast forward index methods for psuedo-relevance feedback retrieval," *ACM Transactions on Information Systems*, vol. 33, no. 4, Article 19, pp. 1–33, 2015.

[59] A. Spink, B. J. Jansen, D. Wolfram, and T. Saracevic, "From e-sex to e-commerce: Web search changes," *IEEE Computer*, vol. 3, no. 35, pp. 107–109, 2002.

[60] J. Miao, J. X. Huang, and J. Zhao, "Topprf: A probabilistic framework for integrating topic space into pseudo relevance feedback," *ACM Transactions on Information Systems*, vol. 34, no. 4, Article 22, pp. 1–36, 2016.

[61] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[62] M. D. Smucker, J. Allan, and B. Carterette, "A comparison of statistical significance tests for information retrieval evaluation," in *Proceedings of the 16th ACM Conf. on Information and Knowledge Management (CIKM'07)*, 2007, pp. 623–632.

[63] S.-H. Na, "Two-stage document length normalization for information retrieval," *ACM Transactions on Information Systems*, vol. 33, no. 2, Article 8, pp. 1–40, 2015.

[64] J. H. Paik, "A novel tf-idf weighting scheme for effective ranking," in *Proceedings of the 36th Ann. Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2013, pp. 343–352.

Edward K.F. Dang received a BA degree in Physics from the University of Oxford, a Ph.D. degree in Physics from the University of Bristol, and a Ph.D. degree in Computing from The Hong Kong Polytechnic University. His research experience includes topics in condensed matter theory and information retrieval. He is currently a research assistant in the Department of Computing of The Hong Kong Polytechnic University.

Robert W.P. Luk (S'85–M'87–SM'01) received the B.Sc. degree in Electronic Engineering and the Engineering Diploma, from the School of Electronics and Computer Science, University of Southampton, the M.Sc. degree in cognition, computing, and psychology from the Department of Psychology, University of Warwick, and later the Ph.D. degree from the School of Electronics and Computer Science, University of Southampton. He is currently an Associate Professor in the Department of Computing, Hong Kong Polytechnic University. He was a Visiting Research Scholar in the Center for Intelligent Information Retrieval, University of Massachusetts, Amherst. He has participated in the NTCIR and TREC information retrieval evaluation workshops and he has implemented various retrieval models. He is interested mainly in information retrieval and in general science of science.

James Allan received an A.B. degree in Mathematics from Grinnell College, and M.S and Ph.D. degrees in Computer Science from Cornell University. Professor Allan joined the Computer Science Department of the University of Massachusetts Amherst in 1994 and has run through a number of titles, most recently Professor as of 2008. He is currently serving as Chair of the Faculty and is co-Director of the Center for Intelligent Information Retrieval (CIIR). He previously spent five years as a systems programmer/analyst at Dickinson College and has been a partner in a software services company.