An evaluation of novelty and diversity based on fuzzy logic

Simone Santini[*]
Escuela Politécnica Superior
Universidad Autónoma de Madrid
simone.santini@uam.es

Pablo Castells
Escuela Politécnica Superior
Universidad Autónoma de Madrid
pablo.castells@uam.es

ABSTRACT

Information retrieval systems are based on an estimation or prediction of the *relevance* of documents for certain topics associated to a query or, in the case of recommendation systems, for a certain user profile.

Most systems use a graded relevance estimation (a.k.a. relevance status value), that is, a real value $r(d, \tau) \in [0, 1]$ for the relevance of document d with respect to topic τ. In retrieval systems based on the Probability Ranking Principle [9], this value has a probabilistic interpretation, that is, $r(d, \tau)$ is equivalent (in rank) to the probability that a user will consider the document relevant. We contend in this paper for an alternative interpretation, where the value $r(d, \tau)$ is considered as the fuzzy truth value of the statement "d is relevant for τ". We develop and evaluate two measures that determine the quality of a result set in terms of *diversity* and *novelty* based on this fuzzy interpretation.

1. INTRODUCTION

Information Retrieval (IR) theory and systems revolve around the core –and ill-defined– notion of *relevance*. IR models, methods, evaluation and –if we may use the term– philosophy are concerned with the estimation, prediction, assessment, evaluation, formalization, and understanding of relevance. In a simple and generic formulation of a retrieval system (the one we shall use in this paper), we have a set $\mathcal{T}$ of *topics* of interest, a data base $\mathcal{D}$ of documents, and a function r over $\mathcal{T} \times \mathcal{D}$, where $r(d, \tau)$ represents the *relevance status* of the document d for topic τ. In the Boolean IR model, relevance takes values in $\{0, 1\}$ or, more in general, in a set isomorphic to the boolean data type $\mathbf{2}$; $r(d, \tau) = \text{true}$ if document d is relevant for topic τ, while $r(d, \tau) = \text{false}$ if document d is not. This crude characterization has often proved insufficient: many algorithms and methods require a finer notion of the relevance of documents than simply declaring them relevant or not relevant. For this reason, IR systems usually work with a *graded* relevance $r(d, \tau) \in [0, 1]$.

How are we to interpret graded relevance? What is the precise meaning of a statement such as $r(d, \tau) = 0.8$? This important semantic question is generally overlooked, mostly because in standard systems the way we interpret relevance does not make all that difference. IR systems return documents sorted by their relevance status value, and under any reasonable interpretation of r, it is always the case that a document with $r(d, \tau) = 0.8$ is more "desirable" than a document with $r(d, \tau) = 0.2$, and should be returned in a higher position. This being the case, who cares what $r(d, \tau) = 0.8$ really means? The issue, however, is quite important in more recent systems that deal with *diversity* and *novelty* [10, 1, 3]. In these cases, relevance status values are used in objective functions for retrieval result diversification, and ground truth relevance values are used as arguments in diversity-oriented IR quality metrics. Here, it is not just a matter of which documents are more relevant than others, but of which are the appropriate tools to manipulate relevance values. These tools depend on the way such relevance values are interpreted.

One common interpretation of relevance is *probabilistic* [9, 11, 1, 12]. In this interpretation, the value $r(d, \tau)$ represents –or is rank-equivalent to– the probability that a user will consider d relevant for topic τ. This identification has important consequences, as it entails that the appropriate machinery for manipulating relevance is Bayesian (e.g. multiplication for independent events, the Bayes theorem for conditional probabilities, etc.). As an alternative to the probabilistic interpretation, we explore a *fuzzy* (graded truth) interpretation of relevance, lifting the binary relevance assumption. Our motivation rests on the difference between uncertainty (caused by incomplete information) and fuzzyness (which is a characteristic of linguistic descriptions such as *relevant*).

The endorsement of fuzzyness over uncertainty entails a different choice of manipulation instruments. We shall use a version of fuzzy logic to express formally the statement that a set of result $\mathcal{R}$ is *novel* (has no redundancies) and *diverse* (covers all the topics of interest). The fuzzy interpretation of the relevance will transform these statements too into fuzzy formulas, so that for each set of results $\mathcal{R}$ we shall be able to give the degree of truth of he statement $\mathcal{R}$ *is novel and diverse* and, consequently, to pick the set for which the statement is most true.

2. THE SEMANTIC OF RELEVANCE

As we have mentioned in the introduction, relevance is ofen given in the form of a real number, generally as $r(d, \tau) \in [0, 1]$. The obvious question to ask (one, as we shall see, that bears quite strongly on the form that the systems should

[*]This work was supported by the *Ministerio de Educación y Ciencia under the grant N. MEC TIN2008-06566-C04-02,* Information Retrieval on different media based on multidimensional models: relevance, novelty, personalization and context.

Copyright is held by the authors. Workshop on Novelty and Diversity in Recommender Systems (DiveRS 2011), held in conjunction with ACM RecSys 2011. October 23, 2011, Chicago, Illinois, USA.
.

take) is: what is the interpretation that we should give to this value?

The most common interpretation of this vaue that is given in information retrieval is probabilistic, that is: *the value $r(d, \tau)$ represents the probability that a user will consider document d relevant for topic τ.* The probabilistic framework entails that we are dealing with a situation in classical logic subject to uncertainty due to limited information. That is, the underlying model is still that of documents that either completely relevant or completely irrelevant (that is, reelvance can be described within the framework of Boolean propositional logic), but we do not have enough information to make a determination [5].

We explore here an alternative logical framework for the question of relevance to be posed. In reality, the documents are given and known completely, so (within the limits of the modeling technques used) instead of modeling the uncertainty in the determination of relevance, one may consider the relevance of a document for a certain topic as a *fuzzy* truth value. This corresponds to the most natural longuistic description that one might give of a document. One doesn't just describe a document as relevant or not relevant: one would rather say that a document is *not very* relevant, *somewhat* relevant, *very* relevant, and so on. These linguistic qualifiers are appropriately modeled with graded truth values rather than with formalisms that deal with uncertainty.

A good example of the difference between the two is given in [2]. Imagine a bottle of water locked in a pantry, so that we can't see it. We know that the bottle is either full or empty, but we have no information about which is which. We can model this situation of uncertainty by saying that with probability 0.5 the bottle is full. Even if we don't know which is which, the bottle is still either completely full or completely empty. The situation is the opposite if *we can see* the bottle and the bottle is half full. In this case, we have complete information: there is no uncertainty involved, and all observers will agree that the bottle is half full. We say in this case that the statement "the bottle is full" has a *truth value* of 0.5; we have fuzzyness, but no uncertainty.

Relevance assessment can be dealt with analogously: the values $r(d, \tau)$ do not model uncertainty (since, as we have said, we have complete information about the documents), but the fuzzyness of the statement *document d is relevant for topic τ*. They are not probabilities, but degrees of truth. The assumption of graded truth entails that the right formalism to use is that of fuzzy logic, to which we shall give a brief introduction in the next section.

3. FUZZY LOGIC AND BL-ALGEBRA

There are several approaches to develop a fuzzy logic. One can start with the basic connective and an involutive negation [4], or define the operations based on a suitable t-norm. The latter approach, which we shall follow here, is based mainly on [7, 6].

Definition 3.1. *A (continuous)* t-norm *is a continuous function $* : [0,1]^2 \rightarrow [0,1]$ such that, for all $x, y, x \in [0,1]$*

$$\begin{array}{lrcll} \textit{i)} & x * y & = & y * x & \textit{(commutativity)} \\ \textit{ii)} & (x * y) * z & = & x * (y * z) & \textit{(associativity)} \\ \textit{iii)} & x \le y & \Rightarrow & x * z \le y * z & \textit{(left monotony)} \\ \textit{iv)} & x \le y & \Rightarrow & z * x \le z * y & \textit{(right monotony)} \\ \textit{v)} & 1 * x & = & x & \\ \textit{vi)} & 0 * x & = & 0 & \end{array} \tag{1}$$

(Note that property iv is redundant, as it is a consequence of commutativity and left monotony.)

Definition 3.2. *A BL-algebra is an algebra*

$$\mathbf{L} = ([0,1], \cap, \cup, *, \Rightarrow, 0, 1) \tag{2}$$

where

i) $([0,1], \cap, \cup, 0, 1)$ *is a lattice with least element* 0 *and largest element* 1*;*

ii) $(L, *, 1)$ *is a commutative semigroup, where $*$ is a t-norm;*

iii) *for all x, y, z:*

a) $z \le (x \Rightarrow y)$ *iff* $x * z \le y$*;*

b) $x \cap y = x * (x \Rightarrow y)$*;*

c) $x \cup y = ((x \Rightarrow y) \Rightarrow y) \cap ((y \Rightarrow x) \Rightarrow x)$*;*

d) $(x \Rightarrow y) \cup (y \Rightarrow x) = 1$.

Property a and the continuity of $*$ imply that $\Rightarrow$ is the *residual* of $*$ [7]:

$$x \Rightarrow y = \sup\{z | z * x \le y\} \tag{3}$$

that is, that $\mathbf{L}$ is a residuated lattice. Property b and continuity imply that $x \cap y = \min\{x, y\}$, while property c and continuity imply that $x \cup y = \max\{x, y\}$.

The syntax of the fuzzy logic is based on two operators: the *strong conjunction* $\sqcap$ and the implication $\rightarrow$, as well as the constant $\bar{0}$. Formulas are composed of propositional variables, the constant, and these operators. Well formed formulas are defined recursively: propositional variabless and $\bar{0}$ are well formed formulas; if ϕ and ψ are well formed formulas then

$$\phi \sqcup \psi \quad \phi \rightarrow \psi \quad (\phi) \tag{4}$$

are as well. Nothing else is a well formed formula. Let W be the set of well formed formulas. An *evaluation function* assigns a value $e(x)$ to each propositional variable x and extends to a function $e : W \rightarrow [0,1]$ through the definition

$$\begin{aligned} e(\bar{0}) &= 0 \\ e(x \sqcap y) &= e(x) * e(y) \\ e(x \rightarrow y) &= e(x) \Rightarrow e(y) \end{aligned} \tag{5}$$

Further connectives are defined as:

$$\begin{array}{rcll} \phi \wedge \psi & \text{is} & \phi \sqcap (\phi \rightarrow \psi) & \text{(conjunction)} \\ \phi \vee \psi & \text{is} & ((\phi \rightarrow \psi) \rightarrow \psi) \wedge ((\psi \rightarrow \phi) \rightarrow \phi) & \text{(disjunction)} \\ \neg\phi & \text{is} & \phi \rightarrow \bar{0} & \text{(negation)} \\ \phi \equiv \psi & \text{is} & (\phi \rightarrow \psi) \sqcap (\psi \rightarrow \phi) & \end{array} \tag{6}$$

A formula ϕ is a tautology if $e(\phi) = 1$ for each evaluation function e. Based on this syntax and the algebraic semantics, different logic systems can be obtained by selecting different axioms. Here we shall use the standard axioms of [7]. The deduction rule is modus ponens. One consequence of the use of certain t-norms, which constitutes a problem in our case, is that the negation might degenerate into a two-values function, that is, with the residual of many t-norms we have

$$(x \Rightarrow 0) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

to avoid this, we requires that the Lukasiewicz axiom be true, namely that $\neg\neg\phi = \phi$. This constraints us to make use of the Lukasiewicz norm

$$x * y = \max\{0, x + y - 1\} \tag{8}$$

a choice that gives us

$$\begin{aligned} x \Rightarrow y &= \begin{cases} 1 & \text{if } x < y \\ 1 + y - x & \text{otherwise} \end{cases} \\ \neg x &= 1 - x \end{aligned} \tag{9}$$

Finally, we efine the *true* constant $\bar{1} = \neg\bar{0} = \bar{0} \to \bar{0}$.

THEOREM 3.1. *The following are tautologies in the BL-algebra*

$$\begin{aligned} &\bar{1} \\ \phi &\to \bar{1} \end{aligned} \tag{10}$$

3.1 Quantifiers

In this paper, we shall define the semantics of forumlas only on finite models. In this context, a quantifier can be seen as a mapping from the power set of the set of truth values to truth values. For example, the (classical) quantifier $\forall$, used in an expression like $\forall x.p$, where the model of x is the finite set $X = \{x_1, \ldots, x_n\}$ can be seen as a mapping $\forall : \mathbf{2}^X \to \mathbf{2}$ such that $\forall : \{p_1, \ldots, p_n\} \mapsto$ true only if all the p_n are *true* [13]. While in classical logic there are two quantifiers $(\forall, \exists)$, in fuzzy logic there is an infinite family of quantifiers, which are used to model linguistic expressions such as *many*, *few*, *about ten*, etc. Here we shall conly need quantifiers from the simplest of such family, the so-called *type* $\langle 1 \rangle$ *quantifiers* [8].

We shall define and analyze two families of quantifiers, which we shall call the *strong* and the *weak* family. In the weak family we give independent definitions of the universal and the existential quantifiers, that is, we will not use the classical logic equivalence $\forall x.p \equiv \neg\exists x.\neg p$. This will give us some more freedom to choose the t-norm on which we will base our system, since we will not have to worry too much if the negation degenerates into a binary-valued operation.

Given a finite model $X = \{x_1, \ldots, x_n\}$ and a unary logic function p, the expression $\forall x.p$ is true to the extent that p is true for all the values $x \in X$. This entails the definition:

$$\forall x.p \text{ is } \bigwedge_{i=1}^{n} p(x_i) \tag{11}$$

and

$$e(\forall x.p) = \bigcap_{i=1}^{n} e(p(x_i)) = \min_X e(p(x_i)) \tag{12}$$

The existential quantifier we interpret indipendently as the quantifier that is true to the extent that *at least one* of the propositions $p(x_i)$ is true, that is

$$\exists x.p \text{ is } \bigvee_{i=1}^{n} p(x_i) \tag{13}$$

and

$$e(\exists x.p) = \bigcup_{i=1}^{n} e(p(x_i)) = \max_X e(p(x_i)) \tag{14}$$

In the case of Lukasiewicz logic, in which $\neg\neg\phi = \phi$, the weak quantifiers still have the property that $\exists x.p = \neg\forall x.\neg p$, however this is not true in general.

The second possibility is to define a *strong (universal) quantifier* using the strong conjunction. In this case we have

$$\forall x.p \text{ is } \sqcap_{i=1}^{n} p(x_i) \tag{15}$$

and

$$e(\forall x.p) = e(p(x_1)) * e(p(x_2)) * \cdots * e(p(x_n)) \tag{16}$$

In this case we can't define the existantial quantifier as we have done for the weak case, since we don't have a corresponding *strong* disjunction. Rather, we will resort to the standard idea from classical logic: *there is an* x_i *such that* $p(x_i)$ *is true to the extent that "not for all* x_i *is* $p(x_i)$ *false"*, that is:

$$\exists x.p \equiv \neg\forall x.\neg p \text{ is } \neg\bigwedge_{i=1}^{n} \neg p(x_i) \equiv (\bigwedge_{i=1}^{n} (p(x_i) \to \bar{0})) \to \bar{0} \tag{17}$$

The fuzzy logic that we have introduced is sound with respect to the BL-algebra (every theorem of fuzzy logic is a tautology in the BL-algebra) and the Lukaziewicz logic is complete with respect to the class of MV-algebras, that is, of the algebras such that, for all x, $((x \Rightarrow 0) \Rightarrow 0) = x$. So, we have two ways to prove that a formula is true. We can either derive it from the axioms of fuzzy logic using modus ponens, or we can prove that it is a tautology in the BL-algebra (or in the MV-algebra, in the case of Lukasiewicz logic) based only on the general properties of the evaluation function, and independently of the evaluation of the predicate variables that appear in the formula. The first way is formally more correct, but much more labor-intensive. Since in this paper we shall not need too many properties, we shall in general resort to the second method.

THEOREM 3.2. *For both the weak and the strong quantifiers it is*

$$\forall x.\forall y.p \equiv \forall y.\forall x.p \tag{18}$$

(The proof is a simple application of the definition and associativity.)

The strong conjunction and the strong quantifier have a problem, which is particularly pernicious for our application. In most of the logic systems, the formula

$$\phi \to (\phi \sqcap \phi) \tag{19}$$

is *not* a theorem of Fuzzy logic[1]. The reason is that, for any

[1] An exception is Gödel logic, in which this is taken as an axiom. Gödel logic, however, entails that

$$e(\phi \sqcap \psi) = \min\{e(\phi), e(\psi)\}$$

that is, the t-norm $*$ is "min".

t-norm that is not *min*, we have $x * x < x$ so

$$e(\phi \to (\phi \sqcap \phi)) = e(\phi) \Rightarrow (e(\phi) * e(\phi)) \qquad (20)$$

and, setting $x = e(\phi)$ and $y = e(\phi) * e(\phi) < x$, we have

$$e(\phi \to (\phi \sqcap \phi)) = \sup\{z | x * z < y\} < 1 \qquad (21)$$

The fact that $e(\phi \sqcap \phi) < e(\phi)$ means that, if we take a series of predicates $p(x_i)$ such that $e(p(x_i)) = x$, the value

$$e(\forall x.p) = \overbrace{x * x * \cdots * x}^{n} \qquad (22)$$

will become, for n large enough, equal to zero: the quantification of a large enough number of predicates that are not entirely true will yeld false. For example, consider the case of the Lukasiewicz norm. Here $x * x = \max\{0, 2x - 1)\}$ and

$$\overbrace{x * x * \cdots * x}^{n} = \max\{0, (n+1)x - n\} \qquad (23)$$

so that for $n > \frac{x}{1-x}$ the quantifier will be false. As we shall see in the following, if we look for a set $\mathcal{R}$ with n results, we shall have to do several universal quantifications on universes with n members and, unless n is very small or the relevance of the documents is very close to 1, we shall get a score of 0 for all sets.

4. DIVERSITY AND NOVELTY

We now have the tools to express the diversity and novelty of a set of result under the fuzzy interpretation of relevance. For the sake of clarity, we shall derive two separate predicates, one for diversity and one for novelty that we shall then join in a conjunction to derive the statement *set $\mathcal{R}$ is novel (non-redundant) and diverse*. Here we assume that in all quantifications, the variables d and d' will range over $\mathcal{R}$, while the variable τ will range over the set $\mathcal{T}$ of topics. That is, we shall use the following short forms:

$$\begin{aligned} \forall d.p &\equiv \forall d.(d \in \mathcal{R} \to p) \\ \exists d.p &\equiv \exists d.(d \in \mathcal{R} \wedge p) \\ \forall \tau.p &\equiv \forall \tau.(\tau \in \mathcal{T} \to p) \\ \exists \tau.p &\equiv \exists \tau.(\tau \in \mathcal{T} \wedge p) \end{aligned} \qquad (24)$$

A result set $\mathcal{R}$ is *diverse* if for every topic there is a document in the set that is relevant for it. That is, the statement $\mathfrak{D}(\mathcal{R})$ can be expressed simply as

$$\mathfrak{D}(\mathcal{R}) \equiv \forall \tau . \exists d . r(d, \tau) \qquad (25)$$

A document is *novel* (or *non-redundant*) if there is at least one topic for which only that document is relevant, and a set is *novel* if all its documents are novel. That is:

$$\mathfrak{N}(\mathcal{R}) \equiv \forall d.\exists \tau.(r(d,\tau) \wedge \forall d'.(r(d',\tau) \to d = d')) \qquad (26)$$

We shall call this the *weak* novelty. There is another possibility of defining novelty, which we shall call *strong.* We can require that there be no overlapping between the topics covered by the documents, that is, whenever a document d is relevant for a topic, no other document is relevant for that topic. That is:

$$\mathfrak{N}'(\mathcal{R}) \equiv \forall d.\forall \tau.(r(d,\tau) \to \forall d'.(r(d',\tau) \to d = d')) \qquad (27)$$

We leave as an exercise to the reader to prove, using the definition of the quantifiers, the axioms and modus ponens, that, for an arbitrary $\mathcal{R}$,

$$\mathfrak{N}'(\mathcal{R}) \to \mathfrak{N}(\mathcal{R}) \qquad (28)$$

A set $\mathcal{R}$ is *qualified* if it is diverse and novel. Since we have two versions of novelty, we have correspondingly two definitions of qualification. The strong qualification is defined as

$$\begin{aligned} &\mathfrak{S}(\mathcal{R}) = \mathfrak{D}(\mathcal{R}) \wedge \mathfrak{N}'(\mathcal{R}) \\ &= \forall \tau.\exists d.(r(d,\tau)) \wedge \forall d.\forall \tau.(r(d,\tau) \to \forall d'.(r(d',\tau) \to d = d')) \\ &= \forall \tau. \left(\exists d.r(d,\tau) \wedge \forall d.(r(d,\tau) \to \forall d'.(r(d',\tau) \to d = d'))\right) \end{aligned} \qquad (29)$$

while the weak qualification is defined as

$$\begin{aligned} &\mathfrak{S}(\mathcal{R}) = \mathfrak{D}(\mathcal{R}) \wedge \mathfrak{N}(\mathcal{R}) \\ &= \forall \tau.\exists d.(r(d,\tau)) \wedge \forall d.\exists \tau.(r(d,\tau) \wedge \forall d'.(r(d',\tau) \to d = d')) \end{aligned} \qquad (30)$$

Before we write down the evaluation functions for these formulas, we consider the translation of the logical function (of d and τ)

$$r(d,\tau) \to \forall d'.(r(d',\tau) \to d = d') \qquad (31)$$

The statement $d' = d$ is crisp, so it evaluates to 0 or to 1. If $d' \neq d$, then

$$e(r(d',\tau) \to d = d') = e(r(d',\tau) \to \bar{0}) = e(\neg r(d',\tau)) \qquad (32)$$

while if $d = d'$

$$e(r(d',\tau) \to d = d') = e(r(d',\tau) \to \bar{1}) = 1 \qquad (33)$$

The quantification, whichever form it takes, is a conjuntion (either strong or weak), and 1 is its unit, so, in the case of the quantification we have

$$\forall d'.(r(d',\tau) \to d = d') = \bigwedge_{d' \neq d} \neg r(d',\tau) \qquad (34)$$

and, in the case of strong quantification we have

$$\forall d'.(r(d',\tau) \to d = d') = \sqcap_{d' \neq d} \neg r(d',\tau). \qquad (35)$$

As we have seen, in addition to the difference in the formula, we have different ways of implementing the quantifiers. Using the strong quantifiers on the strong formula leads to the ${}^{S}\mathbf{S}$ (strong-strong) evaluation function

$$\begin{aligned} {}^{S}\mathbf{S}(\mathcal{R}) = \sqcap_{\tau=1}^{T} \Big[&\neg \sqcap_{d=1}^{D} \neg r(d,\tau) \sqcap \\ &\sqcap_{i=1}^{D} (r(d,\tau) \to \sqcap_{d' \neq d} \neg r(d \tau)) \Big] \end{aligned} \qquad (36)$$

while if we use the weak quantifiers, we get the $\mathcal{SW}$ evaluation function

$${}^{W}\mathbf{S}(\mathcal{R}) = \bigwedge_{\tau=1}^{T} \left[\bigvee_{d=1}^{D} r(d,\tau) \;\wedge\; \bigwedge_{i=1}^{D} (r(d,\tau) \to \bigwedge_{d' \neq d} \neg r(d',\tau)) \right] \qquad (37)$$

Similarly, the two versions of the weak formula are

$$\begin{aligned} {}^{S}\mathbf{W}(\mathcal{R}) = &\sqcap_{\tau=1}^{T} (\neg \sqcap_{d=1}^{D} \neg r(d,\tau)) \;\sqcap \\ &\sqcap_{d=1}^{D} \left[\neg \sqcap_{\tau=1}^{T} \neg (r(d,\tau) \sqcap \sqcap_{d' \neq d} \neg r(d',\tau)) \right] \end{aligned} \qquad (38)$$

and

$${}^{W}\mathbf{W}(\mathcal{R}) = \bigwedge_{\tau=1}^{T} \bigvee_{d=1}^{D} r(d,\tau) \;\wedge\; \bigwedge_{d=1}^{D} \bigvee_{\tau=1}^{T} \left[r(d,\tau) \wedge \bigwedge_{d' \neq d} \neg r(d',\tau) \right] \qquad (39)$$

The observations of the previous section, in particular eq. (23) advise against the use of the strong quantifiers in large scale problems, so in the following we shall in general limit our considerations to the evaluation functions (37) and (39).

With these functions, we can formulate our two versions of the diversity and novelty optimization problem.

> Strong Fuzzy Diversity(n): Given a data base of documents $\mathcal{D}$, a set of T categories $\mathcal{T}$, and the relevance measures $r(d,\tau)$ with $d \in \mathcal{D}$ and $\tau \in \mathcal{T}$, find the subset $\mathcal{R} \subseteq \mathcal{D}$ with $|\mathcal{R}| = n$ such that ${}^{W}\mathbf{S}(\mathcal{R})$ is maximum.

The problem Weak Fuzzy Diversity(n) is analogous but, in this case, the function that is maximized is ${}^{W}\mathbf{W}(\mathcal{R})$.

5. COMPLEXITY

Information retrieval with novelty and diversity often generates intractable problems [10] and our formulation is not, unfortunately, an exception, as we following theorems show. In order to show NP-completeness we have to transform the optimization problems into equivalent decision problems. The decision problem corresponding to Strong Fuzzy Diversity(n) is the following:

> Strong Fuzzy Decision(n): Given a data base of documents $\mathcal{D}$, a set of T categories $\mathcal{T}$, the relevance measures $r(d,\tau)$ (with $d \in \mathcal{D}$ and $\tau \in \mathcal{T}$), and a number $\rho \in [0,1]$ does there exist a subset $\mathcal{R} \subseteq \mathcal{D}$ with $|\mathcal{R}| = n$ such that ${}^{W}\mathbf{S}(\mathcal{R}) \geq \rho$?

The problem Weak Fuzzy Decision(n) is defined analogously.

Theorem 5.1. Weak Fuzzy Decision*(n) is NP-complete.*

Proof. We shall prove the theorem with a reduction from **X3C** (Exact cover by 3-sets). The statement of the problem is as follows: given a set X with $|X| = 3q$ and a collection C of 3-element subsets of X, does C contain a subset $C' \subseteq C$ such that every element of X occurs exactly in an element of C'?

Note that, although the number of sets in C' is not explicitly stated in the theorem, the constraints of the problem entail that C' contains q sets.

We reduce the problem to Weak Fuzzy Decision as follows. The set $\mathcal{T}$ of categories will have one category for each element of X. There will be a document for each subset $c \in C$, and we shall set $r(d,\tau) = 1$ if c contains the element of X represented by τ, and 0 otherwise.

We claim that Weak Fuzzy Decision(q) has a solution with $\rho = 1$ if and only if **X3C** has a solution.

Rmember that we can write

$$ {}^{W}\mathbf{W}(\mathcal{R}) = \mathfrak{D}(\mathcal{R}) \wedge \mathfrak{N}(\mathcal{R}) \tag{40}$$

where the logic quantifiers in $\mathfrak{D}$ and $\mathfrak{N}$ are interpreted in the weak sense. Consider the term $\mathfrak{D}$. We have

$$ \mathfrak{D}(\mathcal{R}) = \min_{\tau \in \mathcal{T}} \max_{d \in \mathcal{D}} d(d,\tau) \tag{41}$$

$\mathfrak{D}(\mathcal{R}) = 1$ if and only if all the "max" that appear in the equation have a value of 1, that is, if and only if for each category (viz. element of X) there is a document (viz. subset in $\mathcal{R}$) that contains it. In other words, $\mathfrak{D}(\mathcal{R}) = 1$ iff

$$ X \subseteq \bigcup_{c \in C'} c \tag{42}$$

Note however, that X is the universe of discourse, and that no subset c can contain any element not in X. So $\mathfrak{D}(\mathcal{R}) = 1$ iff

$$ X = \bigcup_{c \in C'} c \tag{43}$$

Suppose now that there is a solution to **X3C**. In this case, (43) holds, so $\mathfrak{D}(\mathcal{R}) = 1$. What about $\mathfrak{N}(\mathcal{R})$? Suppose, by contradiction, that $\mathfrak{N}(\mathcal{R}) < 1$. Then there has to be at least one pair (d,τ) such that

$$ e(r(d,\tau) \wedge \bigwedge_{d' \neq d} \neg r(d',\tau)) < 1 \tag{44}$$

(The actual condition is stronger: there must be one such d for *every* τ, but the weaker condition will do here.) So, there has to be d' such that $e(r(d,\tau) \wedge \neg r(d',\tau)) < 1$, that is, $e(r(d,\tau) \wedge r(d',\tau)) > 0$. Since the values of relevance are always 0 or 1, this means $e(r(d,\tau) \wedge r(d',\tau)) = 1$, so the element of X represented by τ belongs to both d and d', i.e. the set represented by d and d' are not disjoint, contradicting the fact that a solution was found.

Suppose now that there is a $\mathcal{R}$ such that ${}^{W}\mathbf{W}(\mathcal{R}) = 1$. In this case, by (43), $X = \bigcup c$, that is, the douments in $\mathcal{R}$ cover all categories. Since there are q documents, $3q$ categories, and each document covers only 3 categories, if there were a category represented by more than a document there would also be a category not represented by any dcument. Since this is not the case, there are no overlaps between the documents, that is, the sets of C' are disjoint.

Note that in this case we didn't need the condition $\mathfrak{N}$: the constraints on the problem guarantee that even without this condition we would have solved **X3C**. □

Theorem 5.2. Strong Fuzzy Decision*(n) is NP-complete.*

The proof is based on the same reduction as that of the previous theorem.

6. THE BEHAVIOR OF THE FUNCTIONS

In this section we shall carry out a preliminary study of the two fuzzy evaluation functions that we are considering: ${}^{W}\mathbf{S}$ and ${}^{W}\mathbf{W}$. Before this, we should make a few methodological considerations. There are, roughly speaking, three categories of methods that we can use to study these functions. We can study them analytically, expressing them in closed form; we can generate data using a known statistical distribution and determine the functions' behavior *vis à vis* certain controlled variables; or we can resort to user data collected from an existing system.

It should be evident that the latter solution, despite its widespread use, is inadequate in this case, since it doesn't allow a fine control over the independent variables and the controlled parameters of the evaluation. Tests on "real" are good for obtaining a qualitative impression of how a whole system works, but would make little sense in our predicament.

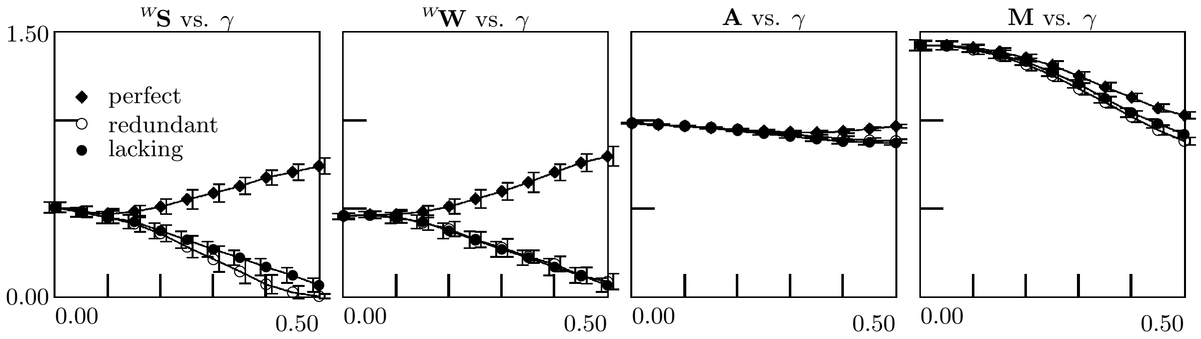


Figure 1: Redundant sets have $r = 2$, lacking sets have $r = -2$; all the results have $c = 24$ and $r = 6$.

Closed form solutions are clearly the best way to study a function, but they may be difficult to obtain under very general hypotheses. Here, we study analytically the behavior of our evaluation functions under a simple but telling special case: that of two topics. As we shall see later on, this setting is fairly representative of more general situations. For a more general setting, we recur to numerical calculations with controlled data sets. In this case, we not only calculate our two evaluation functions ${}^{W}\mathbf{S}$ and ${}^{W}\mathbf{W}$, but compare them with two examples of the state of the art appeared in the literature: the probabilistic measure presented in [1] (and indicated in the following as $\mathbf{A}$), and the *undirected compensatory* measure of [12] (indicated with $\mathbf{M}$).

6.1 Closed-form model

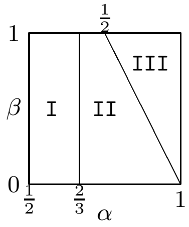
We consider a system with two categories, and result sets of two documents. We shall consider three sets, R_1, R_2, and R_3. The documents of each set are represented as vectors, where the value $\alpha > \frac{1}{2}$ represents relevance while the value $\beta < \frac{1}{2}$ represents irrelevance for a particular topic. The three sets of two documents are as follows:

$$R_1 : \begin{cases} d_1 = [\alpha, \beta] \\ d_2 = [\beta, \alpha] \end{cases}$$
$$R_2 : \begin{cases} d_1 = [\alpha, \alpha] \\ d_2 = [\beta, \alpha] \end{cases} \tag{45}$$
$$R_3 : \begin{cases} d_1 = [\alpha, \beta] \\ d_2 = [\beta, \beta] \end{cases}$$

R_1 is the "perfect" set: document d_1 is relevant for category τ_1, and document d_2 is relevant for τ_2. The two documents cover the category range completely and without redundancy. In R_2 the second document is redundant, as d_1 already covers all categories, while in R_3 no document covers category τ_2. We shall say that R_2 is *redundant* (viz. has positive redundancy) and that R_3 is *lacking* (viz. has negative redundancy).

Consider first the function ${}^{W}\mathbf{S}(\mathcal{R})$ which is, for each of the three result sets, a function of α and β defined in the square $\alpha \in [\frac{1}{2}, 1]$, $\beta \in [0, \frac{1}{2}]$.

In order to determine the behavior of the function, we shall need to divide the square in three regions as illustrated below together with the values of the function in the three regions.

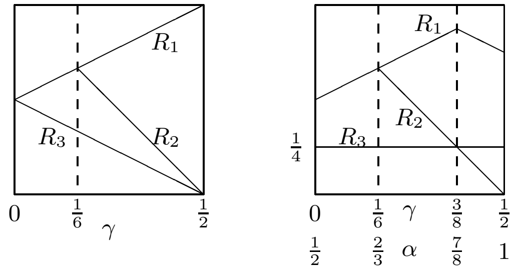


	I	II	III
R_1	α	α	$2-(\alpha+\beta)$
R_2	α	$2(1-\alpha)$	$2(1-\alpha)$
R_3	β	β	β

It must be noted that, for $\alpha < \frac{2}{3}$, this function doesn't discriminate between the "perfect" set and the redundant one. The interpretation of this phenomenon hinges on the definition of redundance. For low values of α, it not so obvious that having two documents about the same topic constitutes a true redundancy, since the relevance of a document is low enough that a second document does indeed add relevance. To have a better idea of this phenomenon, consider two different parametrizations of α and β. First, we consider a path in which α and β start from a situation of complete confusion and diverge to a situation of crisp (binary) relevance. In particular, we shall consider the parametrization

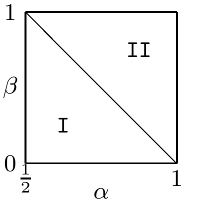
$$\alpha = \frac{1}{2} + \gamma \qquad \beta = \frac{1}{2} - \gamma \tag{46}$$

with $\gamma \in [0, \frac{1}{2}]$. Then we shall consider the same parametrization of α, but keeping $\beta = \frac{1}{4}$. The value of the function ${}^{W}\mathbf{S}$ for the three result sets, as a function of γ with the two parametrizations is the following



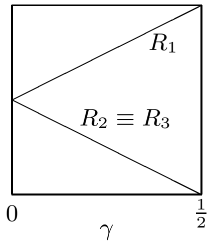
The behavior of the second curve for high values of γ (and therefore of α) is due to the presence of region III. In this case, each one of the two documents of the "perfect" set has a certain relevance not only for the category for which it is nominally relevant, but for the other as well ($\beta = \frac{1}{4}$). When α is high, the fact that, say, d_1 is extremely relevant for τ_1 while d_2 is also somewhat relevant for the same category creates some redundancy. It is therefore not surprising that in this region the value of the evaluation function begins to decrease, behaving exactly as it does in the case of the redundant set R_2.

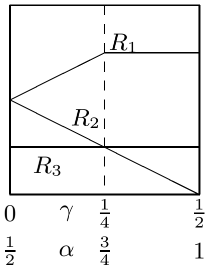
In the case of the ${}^W\mathbf{W}$ evaluation function, we only have to distinguish two regions, represented here with the corresponding function expressions.



	I	II
R_1	α	$1-\beta$
R_2	$1-\alpha$	$1-\alpha$
R_3	β	β

Considering again the parametrization γ and the two previous examples ($\alpha = \frac{1}{2}+\gamma$, $\beta = \frac{1}{2}-\gamma$ and $\alpha = \frac{1}{2}+\gamma$, $\beta = \frac{1}{4}$, respectively, we obtain the following behaviors (behaviors that, in this case, reserve no surprises).





6.2 Numerical tests

In order to extend the range of configurations in which we evaluate the functions, and in order to compare them with other functions appeared in the literature, we resorted to numerical evaluation in statistically controlled conditions. We consider a situation with c topics, in which we seek a result set of s documents. These values are always chosen in such a way that $p = c/s$ is a natural number (this assumption doesn't restrict the scenario appreciably, and simplifies data generation). The "perfect" result set contains s documents, each one of which is relevant to p topics, without overlaps. This entails that this set is optimally diverse and novel. Imperfect sets are created using a redundancy parameter r, and having each one of the documents in the result set be relevant for $p + r$ topics. If $r < 0$ the set will be lacking (some topics will not be covered), while if $r > 0$ the set will be redundant. Note that it must be $1 - p \leq r \leq c - p$. Relevance and irrelevance scores are modeled as two equally distributed random variables obtained starting with a normal distribution and clipping them to $[0, 1]$. That is, if

$$x'_r = N(\alpha, \sigma) \quad x'_{\bar{r}} = N(\beta, \sigma) \tag{47}$$

with $\alpha \geq 1/2$ and $\beta \leq 1/2$, then the scores for relevance and non-relevance are

$$\begin{aligned} x_r &= \text{if}(x'_r < 0, 0, \text{if}(x'_r > 1, 1, x'_r)) \\ x_{\bar{r}} &= \text{if}(x'_{\bar{r}} < 0, 0, \text{if}(x'_{\bar{r}} > 1, 1, x'_{\bar{r}})) \end{aligned} \tag{48}$$

The distribution of the normal, for reasonable values of α and β, if $\sigma < 0.2$; for $\sigma > 0.2$ the distortion due to clipping becomes preponderant and the results become hard to interpret. We chose to do all the measures with $\sigma = 0.1$.

The first diagram is a replica, in the new situation, of the analytical results, using the parametrization (46). The behavior, for the four functions under test, is shown in figure 5.

For $\gamma = 0$ all documents are statistically the same, so none of the methods distinguish between them. As γ increases, and the average difference between relevant and irrelevant documents becomes significant, all four methods separate the perfect set from the redundant and lacking ones (the t-test shows that with $\gamma = 0.05$ the separation is already significant for all methods; this result applies to all other measurements so, from now on, in order to simplify the graphs, we will omit the indication of the variance). Qualitatively, we can observe that the two fuzzy measures appear to give a sharper separation between the perfect set and the other, as reflected by the separation of the curves.

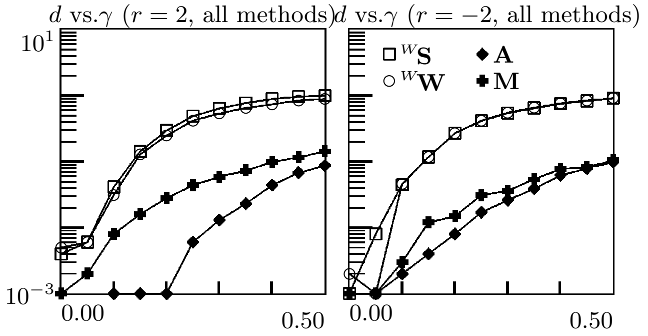


Figure 2: Discrimination results for the four measures under test. Redundant sets have $r = 2$ (graph on the left), lacking sets have $r = -2$ (graph on the right). All the results have $c = 24$ and $r = 6$.

In order to verify this effect, we have performed a series of *discrimination* measures. The idea is that, in order to separate the good results from the bad, we are often more interested in the relative difference between the scores than in the absolute values. For this reason, if u is the score given to a perfect set, and v is the score of a redundant or lacking set, we define the *discrimination coefficient* between the two as

$$d = \frac{|u - v|}{u}. \tag{49}$$

This coefficient is independent of the scale of the measure, and it gives us the degree of separation between the perfect and redundant results as a fracction of the perfect score. The two graphs in figure 6.2 show the discrimination coefficients for the four measures under test[2].

Here too we observe that the discrimination coefficient grows in a much sharper way in the case of the fuzzy measures than it does in the case of the other two.

As a final measure, we analyze the discrimination as a funcion of the redundancy (figure 6.2). We fix the averages of the relevance values to $\alpha = 0.75$ and $\beta = 0.75$ We still have $c = 24$ and $r = 6$, which leads to $p = 4$, so that the redundancy musy be in the range $-3 \leq r \leq 20$. In order to make the graph clearer, we plot $1 - d$ in lieu of d, so that the plot attains its maximum of 1 for $r = 0$, and decreases as r assumes positive or negative values. The graph confirms the main difference that we had already observed between the logic measure an the others that we are analyzing: in the case of the logic measures, the relative difference in score between the "perfect" score and the others is much more

[2] Note that the organization here is different from that of fig. 5: here each graph is relative to a single redundancy, and contains curves for all four measures. This solution would have been too confused for figure 5 due to the presence of the variance.

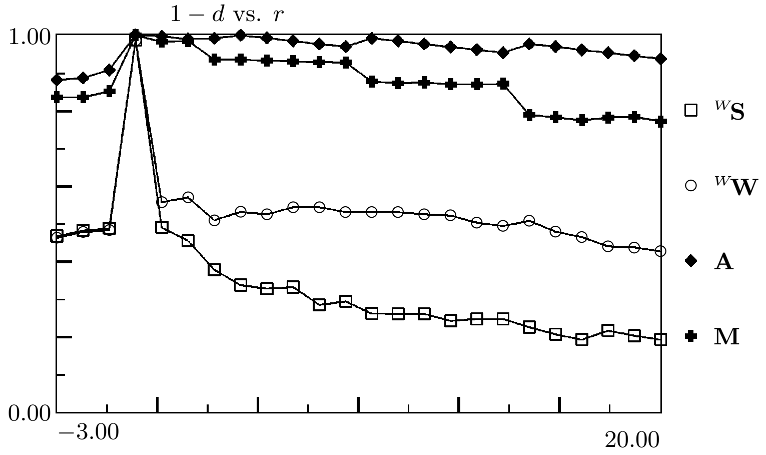


Figure 3: Discrimination $(1 - d)$ results for versus redundancy for the four measures under test. All the results have $c = 24$ and $r = 6$, which leads to $p = 4$ and a range for the redundancy of $[-3, 20]$. Here we set $\alpha = 0.75$, $\beta = 0.25$.

pronounced; even relatively minor defects in the result will result in a considerable drop in the score.

7. CONCLUSIONS

We have presented a model of novelty and diversity consistent with the idea that relevance measures can be interpreted as fuzzy truth values, overcoming the binary relevance simplification. We have derived two different evaluation functions, depending on the specific form of the quantifier used, and we have compared them with two examples of the state of the art.

With respect to other functions, the main characteristics of the logic ones is the sharp decrease in the relative score difference between "perfect" sets and sets with even limited redundancy or lack. Whether this sharpness is an asset or a liability depends, of course, on the specifics of the system that one is designing. At the very least, however, the availability of the logic model provides additional tools to the designer of information retrieval and recommender systems.

A possible way to reduce this discrimination, that we shall study in the future, is to make use of other quantifiers. For example, instead of expressing logically the statement *for each* document d there is a category τ that only d has, we could use a different type of fuzzy quantifier to express the statement *for most* documents d there is a category τ that only d has.

8. REFERENCES

[1] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Leong. Diversifying search results. In *Proceedings of WDSM '09*. ACM, 2009.

[2] J. C. Bezdek and S. Pal. *Fuzzy models for pattern recognition*. New York:IEEE Press, 1996.

[3] Charles Clarke, Maheedhar Kolla, Gordon Cormack, Olga Vechtomova, Azon Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the International ACM SIGIR Conference in Research and Developmens in Information Retrieval*. ACM, 2008.

[4] D. Dubois and H. Prade. A review of fuzzy set aggregation connectives. *Information Sciences*, 36:85–121, 1985.

[5] Didier Dubois and Henri Prade. Possibility theory, probability theory and multiple-valued logics: A clarification. *Annals of Mathematics and Artificial Intelligence*, 32(1-4):35–66, 2001.

[6] Francesc Esteva and LLuís Godo. Monoidal t-norm based logic: towards a logic for left-continuous t-norms. *Fuzzy sets and systems*, 124:271–88, 2001.

[7] Petr Hájek. Basic fuzzy logic and BL-algebras. Technical report V736, Institute of Computer Science, Academy of Science of the Czech Republic, December 1996.

[8] Michal Holčapek. **L**-fuzzy quantifiers of the type $\langle 1^n, 1 \rangle$. *Fuzzy sets and systems*, 159:1811–35, 2008.

[9] S. Robertson. The probability ranking principle in IR. *Journal of documentation*, 33:294–304, 1977.

[10] Simone Santini and Pablo Castells. Intractable problems in novelty and diversity. In *Actas de las XVI Jornadas de Ingeniería del Software y bases de datos*, 2011.

[11] J. Wang and J. Zhu. Portfolio theory of information retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference in Research and Developmens in Information Retrieval*. ACM, 2009.

[12] Yunjie Xu and Hainan Yin. Novelty and topicality in interactive information retrrieval. *Journal of the American Society for Information Science and Technology*, 59(2):201–15, 2008.

[13] Lofti A. Zadeh. A computational approach to fuzzy quantifiers in natural languages. *Computers and Mathematics with Applications*, 9:149–84, 1983.