# Different interpretations of fuzzy gradual-inclusion-based IR models

**L. Ughetto[1] V. Claveau[2]**

[1]IRISA - Université Rennes 2 - Campus de Beaulieu, F-35042 Rennes cedex, France
[2]IRISA - CNRS - Campus de Beaulieu, F-35042 Rennes cedex, France
laurent.ughetto@irisa.fr vincent.claveau@irisa.fr

**Abstract**

Recently, a theoretical fuzzy IR system, based on gradual inclusion measures, has been proposed [1]. In this model, derived from the division of fuzzy relations, the gradual inclusion of a query in a document is modeled by a fuzzy implication. In previous papers, we have shown that, under some assumptions, this model can be seen as a Vector Space Model. This paper also studies other interpretations of our fuzzy IR models based on gradual inclusions. It is shown that the fuzzy models can be interpreted as language models for IR. The links with logical models to IR are also recalled. More generally, this paper discusses the links between these models, shown from the point of view of our fuzzy models.
**Keywords**: IR, IR models, language models

## 1. Introduction

The Information Retrieval (IR) and Databases (DB) communities share the same goal: allowing users to obtain the information they need. However, it is well-known that classical querying methods from DB cannot be used in IR, as they lack the required flexibility to perform an approximate matching between documents and queries, and they seldom offer a mean to order the results. However, recent studies in the field of flexible DB querying lead to new querying mechanisms which are more suited to IR. Moreover, from the work by Bosc et al. [1, 2] on the division of fuzzy relations, new fuzzy IR models based on gradual inclusion have been proposed and experimentally validated [3, 4]. The considered gradual inclusions are founded either on an implication or on a cardinality measure.

Several similarities have been noticed between our fuzzy models and some classical IR models, either by construction, or from their score formula. For instance, the fuzzy model based on an implication has been obtained from the relational DB division operator in [1], but can be obtained by a straight extension of the Boolean IR model [5]. Some links between fuzzy and classical IR models have already been mentioned, as for instance the link with Vector Space Models [1] or with Logical IR Models [6]; they are recalled and detailed in this paper. It also shows that our fuzzy IR models can be re-interpreted as language models. Then, considering the problem from the other side, these links show that several classical IR models, while derived from various paradigms, may be re-interpreted as models based on gradual inclusion measures.

First, principles of our fuzzy IR models are recalled in Section 2. Then, Section 3 details the above-mentioned similarities, and shows that these fuzzy IR models can be re-interpreted by several classical models, which is the main goal of this theoretical paper.

## 2. Gradual inclusion-based fuzzy IR model

If documents and queries are considered as sets of terms, inclusion can be seen as a simple IR model: a document is relevant if and only if it contains all the query terms.

One of the first IR models, namely the Boolean model, is founded on this inclusion model, while making use of Boolean logic (to allow for more general queries). In this model, a document is a set of terms. A query is a logical formula composed of terms linked with AND, OR, NOT operators (written in conjunctive normal form). Then, a document is relevant if and only if, for at least one clause in the query, non-negated terms are present in the document and negated ones are absent.

However, queries are often considered as "bag-of-words", where each term is requested, which corresponds to a simple formula where terms are linked using the AND operator. These queries contain neither negative terms (no NOT) nor alternative (no OR). In this case, the Boolean model boils down to the simple inclusion model.

For the sake of simplicity, only bag-of-words queries are considered is this paper, without loss of generality. Indeed, fuzzy OR and NOT operators can be easily added into our fuzzy models.

This section shows how the inclusion IR model, once extended to a gradual inclusion measure, can lead to an efficient IR model.

### 2.1. Fuzzy IR models

Most extensions of the Boolean IR model try to overcome some of its well-known limitations:

- lack of terms weighting: relative importance of terms in the document, or user's preference in the query, cannot be taken into account;
- binary relevance: a document is relevant or not and, as a consequence, relevant document cannot be ordered;
- no flexibility: a document is not relevant as soon as one query term is missing (or one negated term is present), even if it contains all the others.

To achieve this goal, these extensions modify two parameters of the model: a notion of relative importance of terms is added, usually using term weights, and the binary inclusion measure is replaced by a graded, more flexible one, e.g. a similarity measure. Our fuzzy models also address these parameters.

#### *2.1.1. Terms weighting*

First of all, weighting mechanisms are natural in fuzzy logic. Here, the terms weighting consists in representing a document as a fuzzy subset of the set of indexation terms $T$ [7]. Each term $t_j \in T$ belongs to a document $d_i$ from collection $C$ to a degree $\mu_C(d_i, t_j) \in [0, 1]$ which assesses the representativity degree of the term w.r.t. the document [8, 9]. One can note:

$$d_i = \{\alpha_1/t_1, \ldots, \alpha_m/t_m\} \ , \qquad (1)$$

where $\{t_1, \ldots, t_m\}$ are the terms from document $d_i$ and $\alpha_j = \mu_C(d_i, t_j)$ is the membership degree of term $t_j$ to document $d_i$.

Similarly, a query $q$ can be represented either as a fuzzy subset of $T$, or as a more complex and structured query, using fuzzy logic operators (AND, OR, NOT) [10]. Weighting query terms, raises the problem of the semantics of the given weights $\mu_q(t)$. Most of the time, they encode a user's preference, but they could also represent a discrimination capability.

#### *2.1.2. Inclusion measure*

In classical set-based approaches, relevant documents are the ones containing all the query terms. Then, the relevance of a document $d_i$ is given by the following set-inclusion:

$$q \subseteq d_i \ . \qquad (2)$$

From an axiomatic point of view, this inclusion is represented either by a logical formula:

$$q \subseteq d_i \quad \Leftrightarrow \quad \forall t \in T, \ (t \in q \Rightarrow t \in d_i) \ , \qquad (3)$$

or using a constraint on the sets cardinality:

$$q \subseteq d_i \quad \Leftrightarrow \quad \text{card}(q \cap d_i) = \text{card}(q) \ . \qquad (4)$$

In the following sections, IR fuzzy models based on these two representations (which are no more equivalent in the fuzzy framework) are briefly presented. In these fuzzy models, inclusion becomes gradual. Documents and queries are matched using this gradual inclusion, and the obtained inclusion degree (understood as a relevance degree) allows to rank the documents by relevance.

As shown later, two classical steps from IR systems can be found in our fuzzy systems. First, a matching function computes individual terms scores $S_q(d_i, t_j)$ for each term $t_j$ from a query $q$ and each document $d_i$. Then, an aggregation function is used to compute a global score $S_q(d_i)$ for each document $d_i \in C$ (aggregating the individual terms scores for this document), assessing the relevance degree of each document for the query. In our fuzzy IR system, these matching and aggregation functions are fuzzy ones, taking values in the unit interval.

### 2.2. Implication-based inclusion

#### *2.2.1. Implication-based IR in the literature*

The fuzzy extension of formula (3) in an IR model has been initially proposed in [5], replacing the material implication by a fuzzy one. In this approach, documents and queries are matched at the term level, and the implication degree $\mu_q(t) \rightarrow \mu_C(t, d_i)$ is computed for each term. Then, these individual scores, aggregated by the universal quantifier in (3), are aggregated using the greatest t-norm min (according to the minimum specificity principle), leading to the inclusion degree:

$$\text{Inc}_q(d_i) = \min_{t \in q} \ \big(\mu_q(t) \rightarrow \mu_C(d_i, t)\big) \ , \qquad (5)$$

This degree $\text{Inc}_q(d_i)$ corresponds to the notion of relevance of a document $d_i$ for a query $q_i$, as expressed in the Boolean model.

Independently, this approach has been proposed again in 2008 [1]. Working on the division of fuzzy relations in the fuzzy DB framework (e.g. see [11]), the authors noticed the link between the Boolean IR model and the division of relations, and they envisioned that the division of *fuzzy* relation could be an interesting extension of the Boolean IR model. The proposed model exactly correspond to formula (5).

Note also that the link between the fuzzy extension of formula (3) and the division of fuzzy relations has been also briefly mentioned in [12].

However, these works were theoretical studies only, and this kind of IR fuzzy model has never been experimentally tested and validated until the work reported in [3].

#### *2.2.2. Operators in the implication-based model*

Once the classical model is extended using formula (5), several open problems remain. First, the implication operator has to be chosen among several families of operators (e.g. R-implication and S-implications), while the conjunction min seems to be set by theory. However, it has been shown

in [3] that the efficiency of the fuzzy model closely depends on the chosen operators. Moreover, the semantics of the query terms weights $\mu_q(t)$ depends on the chosen family of implications (R- or S-implication). Some important properties of the operators, mentioned by the different authors having considered this model, are recalled below.

**Absorption property of the min operator.** The min conjunction in formula (5) is required by the minimum specificity principle in the approach by [5]. It is also required in [1] for the result of the division have the properties of a quotient.

However, this operator has a bad property for IR: it is absorbent, which means that only one of its operands makes the result (e.g. $\min(0.3, 0.5) = 0.5$, and 0.3 is *absorbed*). As it aggregates the individual terms scores in (5), the global score of a document is given by the lowest individual term score. The other scores are not taken into account. By contrast, it has been shown that efficient IR systems are the ones taking into account each term, by some balancing formulas or operators. And experiments reported in [1] have also shown that operators with an absorbent part as min or the bounded sum $(\max(0, a+b-1))$ lead to poor results, while other "product-like" t-norms as the product, or Einstein t-norm $(a.b/(2-a-b+a.b))$ lead to the best results.

Thus, to obtain a working IR system, the min has to be replaced by another t-norm $\top$, and formula (5) becomes:

$$\mathrm{Inc}_q(d_i) = \top_{t \in q} \left(\mu_q(t) \rightarrow \mu_C(d_i, t)\right) \ . \qquad (6)$$

This replacement could be seen as an infringement of the theory. However, formula (6) remains a gradual inclusion measure (which is not maximal), and it is sufficient to insure the validity of our IR model. Indeed, the goal of an IR system is not to obtain an absolute, maximal inclusion degree for each document, but to *rank* the documents according to an inclusion degree.

**Threshold and R-implications.** When using a R-implication [13], denoted $\rightarrow_R$, the term weight $\mu_q(t)$ is seen as a requirement threshold. Total satisfaction is obtained as soon as $\mu_C(d_i, t)$ reaches this threshold for all terms $t$ from $q$. When this threshold is not reached, a penalty is applied.

A R-implication can be written the following way:

$$a \rightarrow_R b = 1 \text{ if } a \leq b, \quad f(a, b) \text{ otherwise}, \qquad (7)$$

where $f(a, b)$ expresses a partial satisfaction (less than 1) when antecedent $a$ is not reached by conclusion $b$. The interpretation of $\mu_q(t)$ as a threshold is clear in formula (7), where the implication degree is 1 as soon as term weight in the document $b = \mu_C(d_i, t)$ reaches the requirement $a = \mu_q(t)$.

However, once again, this threshold effect leads to bad results in an IR system which is supposed to rank documents, and not only to determine if they are relevant or not. Using a R-implication, the system cannot rank two documents containing the query terms to the requested degrees, event if the terms have stronger weights in one document, meaning that this document is more relevant than the other.

This point has been experimentally verified: experiments in [3] have shown that R-implications can give good results only when weights are chosen such that the thresholds are never reached.

**Importance and S-implications.** The other widely used family of implications is the one of S-implications. In the second interpretation, $\mu_q(t)$ defines the importance of term $t$ (and the degree $\mu_C(d_i, t)$ is modulated by this importance). In the logical framework imposed by an implication, the underlying notion is that of a guaranteed satisfaction (to a degree $> 0$) when this importance is not total: when $\mu_q(t) < 1$ the term $t$ is not completely important, and it can be forgotten to some extent.

A document $d_i$ is totally satisfactory when $\mu_C(d_i, t) = 1$ for each term $t$ of $q$ whatever its importance. And it is totally unsatisfactory (the global score is 0) only if for at least one term $t$ in $q$, both $\mu_q(t) = 1$ (the requirement has the maximum level of importance) and $\mu_C(d_i, t) = 0$ (the tuple does not fulfil the requirement at all). This behavior is modeled by using an S-implication [13] denoted by $\rightarrow_S$, which writes:

$$p \rightarrow_S q \ = \ \bot(1-p, q) = 1 - \top(p, 1-q) \qquad (8)$$

where $\bot$ stands for a triangular conorm.

#### *2.2.3. Experimental results*

An IR system, founded on this fuzzy model has been implemented and tested on different standard collections of documents [3, 14]. It was parametrized using a terms weighting scheme adapted from the one in BM25 (which is one of the best), normalized to fit properties of membership degrees. Numerous fuzzy operators have been tested. It has been shown that, with an appropriate choice of parameters and operators, the fuzzy systems is rivalling with Okapi (which is the best-scoring, state-of-the-art Vector Space Model). Necessary properties the fuzzy operators must have in order to perform well in an IR context have also been identified.

### 2.3. Cardinality-based inclusion

The other axiomatic approach to inclusion, presented in Section 2.1.2 may also be extended using fuzzy logic. To our knowledge, this extension of formula (4) in an IR model has only been studied in our previous works [4, 15].

#### *2.3.1. From implication to cardinality*

Most often in IR, a document may be relevant even if it does not contain all the query terms. This is why, in Vector Space Models (VSM), the absence

of a term has no effect on the document's score (computed from the other terms). This is achieved using the identity element of the aggregation function (which aggregates the individual terms scores into a global score) as the individual score for a missing term (e.g., 0 in VSM where the aggregation function is the sum). Moreover, a very representative term (rare in the collection and frequent in a document) greatly increases the score. One can conclude that, in VSM, the terms with a large individual score have a more important contribution to the final score than terms with a low individual score.

By contrast, the fuzzy model based on an implication gives more importance to terms with a low score. This is due to the conjoint use of an implication to compute individual score, and of a t-norm to aggregate the individual scores, as the maximal individual score 1 is the identity element of any t-norm. This behavior is closer to the DB world than to the IR world. Indeed, in DB, it is normal for retrieved tuples to be totally satisfying. When exceptions are allowed (as missing or approximate values), these exceptions are given a penalty, and these penalties make the score. Then, in flexible DB systems, low score terms are more important in the final score than (normal) high score terms.

This lead us to consider another approach, more focused on the query terms present in the document: the cardinality-based approach. It consists in computing the fuzzy cardinality of the intersection between $q$ and $d_i$, normalized by the fuzzy cardinality of $q$. Thus, by construction, the score computation is closer to the ones of classical IR systems, which only depends on terms shared by both the document and the query.

#### *2.3.2. Features of the cardinality-based approach*

The inclusion measure, extended to fuzzy sets from formula (4), is given by:

$$Inc_q(d_i) = \frac{|q \cap d_i|}{|q|} \text{ if } |q| \neq 0,\ 1 \text{ otherwise,} \qquad (9)$$

where $|E|$ is the fuzzy cardinality of $E$.

The notion of a fuzzy subsethood measure generalizing $Inc$ and based on the concept of fuzzy entropy has been axiomatized in [16]. Using the definition of the scalar cardinality of a fuzzy set introduced in [17] and often called Zadeh's cardinality:

$$|E| = \sum_{x \in U} \mu_E(x) \ , \qquad (10)$$

where $U$ is the universe of $E$, and using a triangular norm $\top$ for the intersection, formula (9) becomes:

$$Inc(A,B) = \frac{\sum_{x \in U} \top(\mu_A(x),\ \mu_B(x))}{\sum_{x \in U} \mu_A(x)}$$
$$\text{if} \quad \Sigma_{x \in U} \mu_A(x) \neq 0, \quad 1 \text{ otherwise.} \qquad (11)$$

When the query is not empty (which should be always the case), $1/\Sigma_{x \in U} \mu_q(x)$ is a strictly positive constant, denoted $k$ below, whose role is to normalize the inclusion measure in the unit interval $[0, 1]$. Then, the score function writes:

$$Inc_q(d_i) = k. \sum_{x \in U} \top(\mu_q(x),\ \mu_{d_i}(x)) \ , \qquad (12)$$

with $k = 1/\Sigma_{x \in U} \mu_q(x) > 0$.

Let us mention that a *division* interpreted by means of a cardinality-based inclusion cannot be called a division *stricto sensu* since its result is not a quotient in general [18]. Anyway, in the framework considered here, this aspect is not crucial; as already mentioned, an inclusion measure is sufficient for an IR system.

#### *2.3.3. Experimental results*

To allow a fair comparison with the other models, this approach has been tested with the same parameters than the implication-based approach. It has been compared to Okapi, on the same collections of documents and the same weighting scheme have been used.

The only remaining free parameter of this model is then the t-norm $\top$ in formula (12). Numerous operators have been tested and, even if the t-norm plays a different function in this model, we got the same kinds of results: min-like operators lead to poor results while product-like ones lead to the best results.

It can be noticed that, when the chosen t-norm is min, and using BM25 weightings (up to a normalization), formula (12) corresponds to the score formula of Okapi. Thus, it is not surprising to obtain results very close to the Okapi ones in this case. As this paper is not devoted to experiments, details are not given here, and can be found in [4] or [15].

## 3. Links with standard models

The previous section has presented fuzzy IR models based on gradual inclusions and derived from the Boolean IR model. This section shows that these models can also be interpreted as extensions of other classical IR models.

### 3.1. Boolean IR model

As shown in Section 2.1, the link between our fuzzy models and the Boolean IR model is immediate, as the fuzzy models are extensions of the Boolean model by construction.

In this paper, only bag-of-words queries are considered or, if the logical counterpart is considered, queries represented by a conjunction of terms. In the implication-based model, the conjunction corresponds to the t-norm $\top$ in formula (6). This formula may be easily extended to take into account

a disjunction (e.g. the dual t-conorm of $\top$) and a negation $(1-x)$.

In the cardinality-based model, it can be done by set combination; cardinality ratios being computed on several intersections, and combined by fuzzy AND, OR and NOT operators.

### 3.2. Implication-based fuzzy model and logical IR models

#### *3.2.1. Logical IR models*

IR logical models have been studied by many authors during the 90s. Keith van Rijsbergen was the first to propose a logical interpretation of information retrieval, using the concept of implication of a query by a document $d \rightarrow q$, where $\rightarrow$ is an implication operator from the considered logic [19]. From this first work, several authors have studied the role logic may play in IR models. An overview, and a fine analysis of the different approaches in the literature have been proposed by Sebastiani [20] and Lalmas [21].

In this approach, documents and query are represented by logical formula (hence the name), and most of the time a conjunction of the index terms they contain. For instance, a document $d_i$ defined by the set of terms $\{t_1, \ldots, t_n\}$ is represented by the formula $d_i = t_1 \wedge \ldots \wedge t_n$. Although it can be represented by a more general formula, a query $q$ is often a conjunction of terms as in "bag-of-words"-like models.

In order to determine if $d_i$ is relevant to $q$, a logical IR model checks the status of the formula $d_i \rightarrow q$. When the formula is valid, the document is relevant. Checking the validity of this formula can be done in four ways, which are equivalent in propositional logic. The first ones come from model theory:

- $\models d_i \rightarrow q$: formula $d_i \rightarrow q$ is valid (i.e., true whatever the truth of terms $t_j$),
- $d_i \models q$: formula $q$ is a logical consequence of $d_i$ (i.e., valuations satisfying $d_i$ also satisfy $q$).

The two others come from proof theory:

- $\vdash d_i \rightarrow q$: formula $d_i \rightarrow q$ is a theorem,
- $d_i \vdash q$, formula $q$ may be derived from formula $d_i$ (using a proof method).

With other logics, more accurate in IR, these methods may not be equivalent (or even feasible). See [21] for examples.

#### *3.2.2. Two notations for a single model*

At first glance, set-based models (as the Boolean model) and logical models lead to opposite formalizations: $q \rightarrow d_i$ and $d_i \rightarrow q$. However, it can be shown that it is just a matter of notation, and that the logical IR model based on propositional logic corresponds to the Boolean IR model. The difference is due to the formalization process, in which $d_i$ and $q$ do not represent the same thing in the two approaches.

In the case of conjunctive queries, the equivalence is immediate. In the Boolean model, a document is relevant if it contains all the query terms. In the logical model a document is relevant if for each valuation satisfying $d_i$ (i.e., when all the document terms are true) formula $q$ is also true. When $q$ is a conjunction of the terms it contains, it is true only if all its terms are true, and for that they have to be in the document. This means, as in the Boolean model, that the query terms must be included in the set of document's terms. Formulas differ, but the condition is the same. In the case of more general queries, the equivalence can be formally proved. Then, the Boolean IR model and the logical IR models (with propositional logic) are two formalizations of the same paradigm.

Their representations, $q \rightarrow d_i$ and $d_i \rightarrow q$, seem opposed but have the same meaning, under different formalisms. If $d_i \rightarrow q$ is a good notation for logical models, where $d_i$ and $q$ represent the entire document and query, $q(t) \rightarrow d_i(t)$ (and maybe $\forall t,\ q(t) \rightarrow d_i(t)$) would be better for set-based models, where implication is at the terms level.

By construction, our fuzzy model based on an implication is a set-based model. If its logical counterpart is considered, and given the above equivalence, one can conclude that it is also a logical IR model, using fuzzy logic.

### 3.3. Vector Space Models

#### *3.3.1. Score formula in Vector Space Models*

In vector space models (VSM), each document $d_i$ is represented by a vector, each dimension being a term $t \in T$. The values of the vector components $w_{t,d_i}$ depend on the chosen weighting scheme. The weight of a missing term is 0 in general. Queries are also represented by vectors. The weighting scheme can be the one of the documents, or a specific one.

A document score is then given using a similarity measure (most often the cosine, which is equivalent to a $L_2$ distance for normalized vectors) between the query vector and the document vector. Once normalized by the length of both vectors, the score is given by:

$$\mathrm{sim}(d_i, q) = \frac{\sum_{t \in q} w_{t,d_i} \cdot w_{t,q}}{\sqrt{\sum_{t \in q} w_{t,d_i}^2} \cdot \sqrt{\sum_{t \in q} w_{t,q}^2}} \ , \qquad (13)$$

where $w_{t,d_i}$ is the weight of term $t$ in document $d_i$, and $w_{t,q}$ the weight of $t$ in query $q$. Denoting $1/k_{d_i} = \sqrt{\sum_{t \in q} w_{t,d_i}^2}$ the document vector length and $1/k_q = \sqrt{\sum_{t \in q} w_{t,q}^2}$ the query vector length, formula (13) becomes:

$$\mathrm{sim}(d_i, q) = \sum_{t \in q} k_{d_i}.w_{t,d_i} \cdot k_q.w_{t,q} \ , \qquad (14)$$

which is the general form of scores in vector space models.

#### 3.3.2. Fuzzy generalization

Theses score formulas are generally viewed as first the computation of individual terms scores for each term $t \in T$, by means of a matching function (e.g. the product) between the term weights in the document $w_{t,d_i}$ and in the query $w_{t,q}$, followed by an aggregation of these scores (e.g. by the sum).

As already mentioned, these two steps can be found in the fuzzy models. In formula (6), the query and document weights are matched by the implication, then aggregated by the t-norm. In the fuzzy models, the weights can be chosen as in a VSM, up to a normalization, as we did in experiments.

This link is more evident with the cardinality-based model corresponding to formula (9). Indeed, the aggregation function is also the sum, and the matching function is a t-norm, which may be the product.

Thus, our fuzzy models can be seen as generalizations of VSM, with the good consequence that they may benefit from the various improvements of tf-idf weighting schemes.

### 3.4. Language Models for IR

#### 3.4.1. From language models to IR

A language model is a function which gives a probability to a term, or a sequence of terms of the language from a given corpus. The more popular is the $n$-gramme model. It assumes that the appearance probability of a term only depends on the $n-1$ previous terms. In IR, this $n$-gramme model is often used in its simpler form: unigrammes. Thus, it does not take into account the terms positions in the documents, which are considered once again as bags-of-words. Hence, the probability for a term $t$ to be generated by a document $d_i$ is estimated by its frequency in the document, normalized by the length of the document:

$$P(t|d_i) = \frac{\mathrm{tf}_{t,d_i}}{\sum_{u \in d_i} \mathrm{tf}_{u,d_i}} \ . \tag{15}$$

The score of a document $d_i$ for a query $q$ is the probability that the document generates the query; it is given by the product of individual probabilities of terms from the query:

$$\mathrm{score}(d_i, q) = \prod_{t \in q} P(t|d_i) \ . \tag{16}$$

However, when a query term is absent from the document, formula (15) gives a null score, meaning a null score for the whole document (due to the product in formula (16)). In order to obtain a score tolerant to missing query terms, and better estimation of probabilities, numerous *smoothing functions* have been proposed. Their principle is to give a non-null score to every term of the collection in formula (15). For instance, in the model by Hiemstra and Kraaij [22], the smoothed probability is obtained by an interpolation formula between the probability that the term is generated by the document, and the probability that it is generated by the corpus:

$$P_l(t|d_i) = \lambda.P(t|d_i) + (1-\lambda)P(t|C) \qquad \lambda \in ]0,1[ \ . \tag{17}$$

Another method, named *absolute discounting* [23] consist in substracting a small, constant value from the probability of each term, and in redistributing it equitably on all the corpus terms.

#### 3.4.2. Language models and VSM

In [22], Hiemstra and Kraaij show that the score formula of their IR model (based on a language model), may be rewritten in an equivalent form as a VSM. Indeed, with a smoothing, the score formula (16) becomes:

$$\begin{aligned} &\mathrm{score}(d_i, q) \\ &= \prod_{t \in q} \left(\lambda.\frac{\mathrm{tf}_{t,d_i}}{\sum_{u \in d_i} \mathrm{tf}_{u,d_i}} + (1-\lambda).\frac{\mathrm{df}_t}{\sum_{u \in C} \mathrm{df}_u}\right) \\ &\propto \sum_{t \in q} \mathrm{tf}_{t,d_i}.\log\left(1 + \frac{\mathrm{tf}_{t,d_i}}{\mathrm{df}_t.\sum_{u \in d_i} \mathrm{tf}_{u,d_i}} + \right. \\ &\qquad \left. \frac{\lambda.\sum_{u \in C} \mathrm{df}_u}{(1-\lambda)}\right). \end{aligned}$$

Up to a constant, it corresponds to the score formula of a VSM (where the equivalent of a TF and an IDF can be found):

$$\mathrm{sim}(d_i, q) = \sum_{t \in q} w_{t,d_i} \cdot w_{t,q} \ , \tag{18}$$

where $w_{t,d_i} = \mathrm{tf}_{t,d_i}$ and where

$$w_{t,d_i} = \log\left(1 + \frac{\mathrm{tf}_{t,d_i}}{\mathrm{df}_t.\sum_{u \in d_i} \mathrm{tf}_{u,d_i}} + \frac{\lambda.\sum_{u \in C} \mathrm{df}_u}{(1-\lambda)}\right).$$

#### 3.4.3. Fuzzy models and language models

As our fuzzy models may be interpreted as VSMs (cf. Section 3.3.2) and as some language models can be interpreted as VSMs also (cf. Section 3.4.2), it is natural to wonder if fuzzy models can be interpreted as Language models.

In the fuzzy implication-based model, the main operator of formula (6) is a conjunction as in formula (16), the one of language models. Moreover term weights in both models, probabilities in language models, and membership degrees in fuzzy models, take values in the unit interval. The membership degrees just need a normalization to sum up to one, as probabilities do.

By the way, these properties raise similar problems in the models. Indeed, the problem of null probabilities for absent terms also occurs in the fuzzy implication-based model. It occurs when a query term with maximal importance 1 in the query

is absent from the document. To avoid this situation, in the experiments reported in [3] the documents terms weights where bounded to received at least a low, but non-null value $\epsilon$. This mechanism can be considered as an *absolute discounting* smoothing [23].

The fuzzy cardinality-based model can also be rewritten as a language model, applying the opposite transformation than the one proposed by Hiemstra, and presented in the previous section. Thus, our fuzzy models can also be seen as language models.

### 3.5. Straight fuzzy extension of language models

Recently, in [24], a straight extension of Hiemstra IR language model has been proposed using fuzzy logic. In Hiemstra score formula:

$$\text{score}(d_i, q) = \prod_{t \in q} \Big(\lambda.P(t|d_i) + (1-\lambda)P(t|C)\Big) \quad (19)$$

where $\lambda \in ]0,1[$, the product has been replaced with a t-norm, leading to:

$$\text{score}(d_i, q) = \top_{t \in q} \Big(\lambda.P(t|d_i) + (1-\lambda)P(t|C)\Big) \quad (20)$$

where $\lambda \in ]0,1[$. Several operators have been experimentally tested for $\top$, in the same context as experiments mentioned Section 2.2.

It has been shown that, with a good choice of operators, this fuzzy extension gives results rivaling with Hiemstra model, which validates this approach. These experiments confirmed the previously mentioned properties for t-norms in the context of IR.

Then, other aggregation functions has been tested to replace $\top$ in formula (20) as mean operators, and the best results where approximately the same than Hiemstra model's results (sometimes slightly better, sometimes worse).

## 4. Concluding remarks

The principle of fuzzy IR models founded on gradual inclusions, either implication-based or cardinality-based, has been recalled. It has been shown that these families of IR models (or sometimes just one of them) may be seen as several classical IR models: extended Boolean IR model, logical IR models, vector space models and Language models.

These fuzzy models were proposed as generalizations of classical models. If the first goal was to imitate classical systems to validate the approach (which has been done), it is now time to study other operators, and try to find new, original approaches. For instance, in the fuzzy extension of Hiemstra model, several extensions remain to be tested, in particular about the smoothing mechanism. And if we get rid of probabilities, and replace them by possibilities, the weighting mechanism between $P(t|d_i)$ and $P(t|C)$ may be reconsidered.

From an opposite point of view, it is interesting to notice that score formulas of several classical IR models, while having various theoretical foundings, boil down to gradual inclusion measures. These links will be studied in future works, trying to find how our fuzzy models could benefit from the various qualities of the different classical systems they generalize.

## References


[1] P. Bosc and O. Pivert. On the use of tolerant graded inclusions in information retrieval. In *Actes de la 5$^e$ Conférence en Recherche d'Information et Applications, CORIA'08*, pages 321–336, Trégastel, France, 2008.

[2] P. Bosc and O. Pivert. On a parameterized antidivision operator for database flexible querying. In *Proceedings of the 19th International Conference on Database and Expert Systems Applications, DEXA'08*, pages 652–659, Turin, Italy, 2008.

[3] P. Bosc, V. Claveau, O. Pivert, and L. Ughetto. Graded-inclusion-based information retrieval systems. In *Proceedings of the European Conference on Information Retrieval, ECIR'09*, pages 321–336, Toulouse, France, 2009.

[4] P. Bosc, L. Ughetto, O. Pivert, and V. Claveau. Implication-based and cardinality-based inclusions in information retrieval. In *Proceedings of the IEEE International Conference on Fuzzy Systems (Fuzz-IEEE'09)*, pages 2088–2093, Jeju Island, South Korea, 2009.

[5] G. Pasi. A logical formulation of the Boolean model and of weighted Boolean models. In *LUMIS workshop at ECSQARU'99*, London, 1999.

[6] L. Ughetto, G. Pasi, V. Claveau, O. Pivert, and P. Bosc. Implication in information retrieval systems. In G. Pasi, editor, *e-Proceedings of the 9th international conference on Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO'09)*, Paris, France, 2010.

[7] D.A. Buell. An analysis of some fuzzy subset applications to information retrieval systems. *Fuzzy Sets & Systems*, 7:35–42, 1982.

[8] W.G. Waller and D.H. Kraft. A mathematical model of a weighted Boolean retrieval system. *Information Processing & Management*, 15:235–245, 1979.

[9] D.A. Buell and D.H. Kraft. Threshold values and Boolean retrieval systems. *Information Processing & Management*, 17:127–136, 1981.

[10] A. Bookstein. Fuzzy requests: an approach to weighted Boolean searches. *Journal of the American Society for Information Science*, 31:240–247, 1980.

[11] P. Bosc, D. Dubois, O. Pivert, and H. Prade. Flexible queries in relational databases – the

example of the division operator. *Theoretical Computer Science*, 171:281–302, 1997.

[12] M. Baziz, M. Boughanem, Y. Loiseau, and H. Prade. Fuzzy logic and ontology-based information retrieval. In P. Wang, D. Ruan, and E. Kerre, editors, *Studies in Fuzziness and Soft Computing*, volume 215/2007, pages 193–218. Springer, 2007.

[13] J. Fodor and R.R. Yager. *Fundamentals of Fuzzy Sets — The Handbook of Fuzzy Sets Series (D. Dubois and H. Prade eds.)*, chapter Fuzzy Set-theoretic Operators and Quantifiers. Chap. 1.2, pages 125–193. Kluwer Academic Publishers, 1999.

[14] L. Ughetto, O. Pivert, V. Claveau, and P. Bosc. Sri à base d'inclusion graduelle. In *Actes de la COnférence en Recherche d'Infomations et Applications (CORIA'09)*, pages 235–250, Presqu'île de Giens, France, 2009.

[15] L. Ughetto, O. Pivert, V. Claveau, and P. Bosc. Recherche d'information et inclusions graduelles. In *Actes des Journées Francophones sur la Logique Floue et ses Applications (LFA'09)*, pages 125–132, Annecy, France, 2009.

[16] V. Young. Fuzzy subsethood. *Fuzzy Sets & Systems*, 77:371–384, 1996.

[17] A. De Luca and S. Termini. A definition of non-probabilistic entropy in the setting of fuzzy sets theory. *Information and Control*, 17:301–312, 1972.

[18] P. Bosc, D. Rocacher, and O. Pivert. Characterizing the result of the division of fuzzy relations. *International Journal of Approximate Reasoning*, 45:511–530, 2007.

[19] C. J. van Rijsbergen. A non-classical logic for information retrieval. *The Computer Journal*, 29(6):481–485, 1986.

[20] F. Sebastiani. On the role of logic in information retrieval. *Information Processing and Management*, 34(1):1–18, 1998.

[21] M. Lalmas. Logical models in information retrieval: Introduction and overview. *Information Processing & Management*, 34(1):19–33, 1998.

[22] D. Hiemstra and W. Kraaij. Twenty-one at trec-7: ad-hoc and cross-language track. In *Proceedings of the 7th Text Retrieval Conference TREC-7, NIST Special Publication 500-242*, pages 227–238, 1999.

[23] H. Ney, U. Essen, and R. Kneser. On structuring probabilistic dependencies in stochastic language modelling. *Computer Speech and Language*, 8:1–38, 1994.

[24] R. Harastani. Information retrieval: From language models to fuzzy logic. Master's thesis, Université de Rennes 1 - IRISA, Rennes, France, 2010.