Implication-Based and Cardinality-Based Inclusions in Information Retrieval

Patrick Bosc, Laurent Ughetto, Olivier Pivert, and Vincent Claveau

***Abstract*— This paper investigates the use of fuzzy logic mechanisms coming from the database community, namely graded inclusions, to model the information retrieval process. Two kinds of graded inclusions are considered. In this framework, documents and queries are represented by fuzzy sets, which are paired with operations like fuzzy implications and T-norms. Through different experiments, it is shown that only some among the wide range of fuzzy operations are relevant for information retrieval. When appropriate settings are chosen, it is possible to mimic classical systems, thus yielding results rivaling those of state-of-the-art systems. These positive results validate the proposed approach, while negative ones give some insights on the properties needed by such a model. Moreover, this paper shows the added-value of this graded inclusion-based model, which gives new and theoretically grounded ways for a user to easily weight his query terms, to include negative information in his queries, or to expand them with related terms.**

I. Introduction

Information Retrieval (IR) and Databases (DB) communities share the goal of providing the users with the information they ask for. But, it is well-known that classical DB querying mechanisms are not adapted to IR needs. First, they lack the required flexibility for the approximate matching between queries and documents. Secondly, they seldom provide a means to rank the retrieved documents. However, recent studies in the field of fuzzy database querying have provided new mechanisms that may be adapted to the IR framework. Following the recent work of [1], [2], this paper investigates the use of one of these mechanisms, the graded inclusion, in textual IR.

The first goal of this study is to provide some insights on the practical use of the graded inclusions in IR, considering that: "the more the set of words of the query is included into the set of words of a document, the better this document". In this model, documents and queries are represented by fuzzy sets of words. Then, documents and queries are matched using graded inclusion and, for each query, documents are ranked according to their inclusion degree.

Fuzzy extensions of the two classical views of the set inclusion (implication-based and cardinality-based), are considered and compared in this paper. We have shown in [1] that the implication-based approach corresponds to the division of fuzzy relations. Thus, it is a fuzzy extension of the Boolean model for Information Retrieval, which corresponds to the division of classical relations. We also have experimentally validated this approach [3], through experiments recalled and discussed in Section V, where numerous settings were explored, using numerous pairs of implication and T-norm.

Among other comparisons, it has been shown that, to compute the matching degree, our implication-based model focuses on the words from the query absent from the documents, while classical IR models focus on the query words present in the documents. This led us to consider the other view of inclusion, the cardinality-based one. This new approach is studied here, both theoretically and experimentally. It is shown that it corresponds to the Vector Space Model (VSM), which is a standard IR method.

The positive obtained results show that, with appropriate settings, it is possible for our two fuzzy-based models to mimic classical systems and thus to yield results rivaling state-of-the-art ones. It is also possible to determine which of the settings are suited or not to build a retrieval system and, from negative obtained results, some insights are given about the properties needed by such a model.

The paper is organized as follows: the next section reviews some of the existing studies using fuzzy logic in information retrieval and see how they are related to our approach. The theoretical background of the graded-inclusion approach is presented in Section III. The practical implementation and the experimental results of this approach are detailed and discussed in Sections IV and V, and several theoretical extensions allowed by this framework are proposed in Section VI.

II. Related work

Parts of Fuzzy Logic (FL) have been introduced into IR models, since the early 80s (e.g. see [4]). It is natural since the Boolean IR model has been extended using degrees, and FL generalizes Boolean logic with degrees. FL have different uses in IR, as for instance, managing uncertainty in the term representation [5], improving the ranking of the documents [6], enhancing the expressiveness of the querying language... Others have extended the classical IR model to take into account particular situations, for instance to use ordinal terms weights [7], or to use both possibility and necessity measures for terms weights [8]. Most of these papers use parts of FL to deal with particular problems, while our approach proposes a comprehensive theoretical framework.

Among the studies using fuzzy logic in the matching mechanism between a query and a document, one can notice the recent papers of Herrera-Vielma et al. [9] or Oussalah et al. [10]. The latter work is close to ours: it also proposes

P. Bosc and O. Pivert are with ENSSAT - IRISA, Lannion, France (email: {bosc, pivert}@enssat.fr). V. Claveau is with CNRS - IRISA, Rennes, France (email: vincent.claveau@irisa.fr). L. Ughetto is with Université Rennes 2 - IRISA, Rennes, France (email: laurent.ughetto@irisa.fr).

the use of fuzzy implications to compute a similarity measure between a document and a query. However, in their approach, $D \rightarrow Q$ is computed, as usual in IR logical approaches [11], while here the implication is used the other way round, for reasons detailed in the next section.

III. IR AND THE DIVISION OF RELATIONS

Information Retrieval Systems (IRSs) are based on models characterized by three main components: the representation of documents, the query language, and the matching mechanism. This section shows how our graded-inclusion approach generalizes the IR Boolean model on these three components. The next subsections show the link between the Boolean model and the division of relations, how the extension to fuzzy relations is linked to a graded IR approach, and the theoretical basis of our IRS.

A. Division of relations and the Boolean IR model

In the relational model of data, a universe is modeled as a set of relations, manipulated with operations known as the relational algebra. Among these operations, the division of the relation $C(D,T)$ by $Q(T)$ denoted by $C[T \div T]Q$, where T is a set of attributes common to C and Q, aims at determining the D-values connected in C with all the T-values appearing in Q. This operation can be defined by:

$$d \in C[T \div T]Q \quad \Leftrightarrow \quad \forall d \in Q,\ (d,t) \in C\ , \qquad (1)$$

or equivalently, with $\Omega^{-1}(d) = \{t | (d,t) \in C\}$:

$$d \in C[T \div T]Q \quad \Leftrightarrow \quad Q \subseteq \Omega^{-1}(d)\ . \qquad (2)$$

Consider the Boolean IR model in which each document d is described as a set of terms $d = \{t_1, \ldots, t_m\}$, with $t_i \in T$, the set of the index terms, and a query q is a set of expected terms $Q = \{t'_1, \ldots, t'_n\}$. If the collection of documents is represented by a normalized (in a DB sense) relation (C) where a document d of m terms corresponds to the m tuples: $\langle d, t_1\rangle, \ldots, \langle d, t_m\rangle$, and a query is a unary relation (Q), then the query may be answered by the division of C by Q.

This Boolean approach, clearly related to DB querying mechanisms, was at the origin of IR systems. However, it has rapidly shown its limitations and is no more used in IR. Among the reasons, the Boolean approach do not allow to represent and use the relative importance of terms indexing the documents or representing the queries.

B. A fuzzy extension

Most of the extensions of the Boolean IR model take into account the relative importance of terms, through weighting mechanisms. In FL, it naturally consists in representing documents by a fuzzy set of terms [4]. Each term t belongs to a document d of the collection C with a degree $\mu_C(d,t)$ assessing its significance [12], [13]. A query q can also be represented by a fuzzy set of terms, or a more general expression structured with fuzzy operators (AND, OR, NOT) [14]. Using weights (degrees) for terms in both documents and queries raised the problem of their interpretation. This problem is discussed further on.

There are two other classical steps in IRSs. The first one consists in matching documents and queries, computing *individual scores* $S_q(d,t)$ for each term t in the query q, and each document $d \in C$. The second one consists in aggregating the scores $S_q(d,t)$, $\forall t \in q$, to obtain a global score $S_q(d)$ for each document, assessing the satisfaction degree of d w.r.t. q. This degrees allow to rank the documents according to their relevance. Fuzzy IRSs can use matching and aggregation functions defined on the unit interval.

This gradual extension of the Boolean IR model may appear *ad hoc*. However, generalizing the Boolean case described above, the answer to a query q can be obtained by the division of two fuzzy relations, whose tuples are weighted in the unit interval. Then, replacing the set inclusion with a graded one g, expression (2) becomes:

$$C[T \div T]Q(d) = g(Q \subseteq \Omega^{-1}(d))\ , \qquad (3)$$

where $\Omega^{-1}(d) = \{\mu/t | \mu/(d,t) \in C\}$ is a fuzzy set of terms. The semantics of this division depends on both the inclusion operator and the meaning of the weights associated with the tuples in relations C and Q [15].

The two classical views of the set inclusion are the implication-based one:

$$A \subseteq B \quad \Leftrightarrow \quad \forall x,\ (x \in A \Rightarrow x \in B)\ , \qquad (4)$$

and the cardinality-based one:

$$A \subseteq B \quad \Leftrightarrow \quad \text{card}(A \cap B) = \text{card}(A)\ . \qquad (5)$$

Fuzzy extensions of these two approaches are considered for the inclusion operator g. If it is represented by a fuzzy implication $\rightarrow$, (3) becomes:

$$C[T \div T]Q(d) = \min_{t \in Q}\ (\mu_Q(t) \rightarrow \mu_C(d,t))\ , \qquad (6)$$

In this formula, the matching ($\rightarrow$) and aggregation functions ($\min$) appear clearly. Thus, the extension of the Boolean IR model is not *ad hoc* in our framework.

When g is a ratio of cardinalities, (3) is not a division *stricto sensu* since its result is not a quotient in general [16]. However, the formula becomes:

$$\sum\nolimits_{t \in Q} \top(\mu_Q(t),\ \mu_C(d,t)) \quad / \quad \sum\nolimits_{t \in Q} \mu_Q(t)\ , \qquad (7)$$

assuming that $\Sigma_{t \in Q} \mu_Q(t) \neq 0$. Here again, the matching ($\top$) and aggregation functions (normalized sum) appear clearly.

C. Graded inclusion based on a fuzzy implication

In (6), the matching function is a fuzzy implication. The semantics of inclusion degrees and queries terms weights depends on the chosen implication, which may be either a R- or a S-implication. Numerical examples are given in [1], which illustrate the differences.

1) Threshold and R-implications: In the first case, the degree $\mu_Q(t)$ is seen as a threshold and the complete satisfaction requires that this threshold is attained by $\mu_C(d,t))$ for each term t of Q. When this threshold is not reached, a penalty is applied.

This behavior is obtained using a residuated implication (or R-implication) [17], denoted by $\rightarrow_R$. Any R-implication may be rewritten:

$$p \rightarrow_R q = 1 \text{ if } p \leq q, \;\; f(p,q) \text{ otherwise,} \tag{8}$$

where $f(p,q)$ expresses a partial satisfaction (a value less than 1) when the antecedent p is not reached by the conclusion q. The threshold interpretation is clear from formula (8), where the satisfaction degree is 1 as soon as $\mu_C(d,t)$ reaches $\mu_Q(t)$.

2) Importance and S-implications: In the second interpretation, $\mu_Q(t)$ defines the importance of term t (and then the degree $\mu_C(d,t)$ is modulated). In the logical framework imposed by an implication, the underlying notion is that of a guaranteed satisfaction when this importance is under 1: when $\mu_Q(t) < 1$ the requirement is not completely important, and it can be forgotten to some extent. The complete satisfaction requires that $\mu_C(d,t)$ equals 1 for each value t of Q whatever its importance. And a document is totally unsatisfactory ($\mu_{C[T \div T]Q}(d) = 0$) only if for at least one t in Q, both $\mu_Q(t) = 1$ (the requirement has the maximum level of importance) and $\mu_C(d,t) = 0$ (the tuple does not fulfill the requirement at all). This behavior is modeled by using an S-implication [17] denoted by $\rightarrow_S$, as follows ($\perp$ stands for a triangular conorm):

$$p \rightarrow_S q \;\; = \;\; \perp(1-p,q) = 1 - \top(p, 1-q) \tag{9}$$

One can notice that the regular division is recovered from formulas (3) and (6) in the presence of regular relations.

3) Absorption effect: This approach is logical and conjunctive but has an "absorption effect". Indeed, the division operator only retains the smallest degree of implication between Q and C, due to the $\min$ aggregation in (6). This is why (6) is to be relaxed using another T-norm in our IR model.

D. Graded inclusion based on a ratio of cardinalities

One can also consider a set-based view for the graded inclusion used as a matching function in (3). The classical inclusion (5) can be extended to fuzzy sets as follows:

$$Inc(A,B) = \frac{|A \cap B|}{|A|} \quad \text{if } |A| \neq 0, \quad 1 \text{ otherwise.} \tag{10}$$

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 [18]. Using the definition of the scalar cardinality of a fuzzy set introduced in [19] and often called Zadeh's cardinality: $|E| = \sum_{x \in U} \mu_E(x)$, where U is the universe of E, and using a triangular norm $\top$ for the intersection, formula (10) becomes:

$$Inc(A,B) = \frac{\sum_{x \in U} \top(\mu_A(x),\; \mu_B(x))}{\sum_{x \in U} \mu_A(x)}$$

$$\text{if} \quad \Sigma_{x \in U} \mu_A(x) \neq 0, \quad 1 \text{ otherwise.} \tag{11}$$

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 [16]. Anyway, in the framework considered here, this aspect is not crucial and it is worth comparing the behavior of an IRS based on such a *division-like* operator with that based on a *logical division*.

IV. Implementation and experiments

A. Document collections

The two proposed IRSs have been tested using 2 collections of documents from IR evaluation campaigns. The first one is the INIST collection: it contains 163,308 documents (paper abstracts from various scientific disciplines) and a set of 30 queries. The second is a part of the TREC-3 TIPSTER collection, containing 173.252 documents from the Wall Street Journal and 50 queries. For both collections, documents and queries have been lemmatized. The queries are composed of several files: a title, a subject, a description and a set of associated concepts. In the experiments reported below, only title and associated concepts fields have been used as actual queries in INIST, and title filed only in TIPSTER (associated concepts were not available).

B. General IRS features

Our IRS implements the two fuzzy approaches described in Section III. Thus, the score of a document d in front of a query q is computed either as:

$$S(d,\; q) = \top_{t \in q}(w_q(t) \rightarrow w_d(t)) \;\; , \tag{12}$$

or as:

$$S(d,\; q) = \sum\nolimits_{t \in q} \top(w_q(t),\; w_d(t)) \;/\; \sum\nolimits_{t \in q} w_q(t) \;\; . \tag{13}$$

In formula (12), several parameters must be set: the weights of the terms in both the queries and the documents, and the aggregation and matching operators. It is similar in (13), except for the aggregation operator (sum).

C. Features of the implication-based approach

1) Aggregation operator: When $\top$ is the $\min$ operation, $S(d,q)$ equals the inclusion degree of the term $t \in q$ which is the least included in d (the degree of the weakest term reflects the acceptability of the document). As explained below, this approach fails in IR. This is why a large range of T-norms has been tested in (12).

2) Graded inclusion operator: As mentioned in Section III, two classes of operators have been used: R-implications and S-implications. The most representative (and widely used) of each class have been chosen for this first series of tests. Among the R-implications: Gödel, Goguen, Łukasiewicz. Among the S-implications: Kleene-Dienes, Łukasiewicz, Reichenbach, Willmott. See for instance [17] for the definition of these operators.

3) Weights of the terms in the documents: In the context of the division of fuzzy relations, the weights have to carry a clear semantics (importance, threshold...). The OKAPI-BM25 weighting scheme has been used, as it accurately carries the notion of relative importance of terms:

$$w_{BM25}(t,d) = \frac{log(\frac{N-n(t)+0.5}{n(t)+0.5}) \cdot (k_1+1) \cdot \text{tf}_d(t)}{k_1 \cdot ((1-b)+b \cdot \frac{L_d}{L_{\text{avg}}}) + \text{tf}_d(t)} \quad , \quad (14)$$

where $k_1 = 2$ and $b = 0.75$ are constants, L_d the document length, L_{avg} the average document length in the collection, N the total number of documents and $n(t)$ the number of documents containing t. However, in the context of fuzzy computations, the weights must belong to the $[0,1]$ interval. Thus, the OKAPI-BM25 weights $w_{BM25}(t,d)$ have been normalized and bounded.

4) Weights of the terms in the queries: As for the term-weights of documents ($w_d(t)$), the term-weights of the queries ($w_q(t)$) have to carry a clear semantics. Specifically, this is the case when one is dealing with R-implications, in which $w_q(t)$ is a threshold to be reached by $w_d(t)$. For now, this could only have been achieved by a manual terms weighting of the queries. It would have been subjective, and above all would not have allowed a fair comparison with other IRSs. This is why an automatic (and classic) weighting mechanism has been used in this first work, at the expense of the semantics. The term weights rely on the frequency of the terms in the queries, and are normalized and bounded.

D. From an implication- to a cardinality-based approach

Often in IR, a relevant document does not contain all the terms of the query. In most vector-space models, the absence of a query term in a document does not decrease the score of this document; it is just neutral. This is why such models use addition to aggregate the scores of individual terms. By contrast, a very representative term (rare in the collection, and frequent in the document) greatly increases the score. Thus, from an IR point of view, the "best" terms are more important than the "weakest". Unfortunately, the implication-based approach focuses on the terms in q absent from the document, or on the least representative ones. This behavior is due to the fuzzy implication, and the T-norm-based aggregation. The implication gives a *non-one* degree only for unsatisfactory terms; the min keeps the weakest degree; totally satisfactory terms obtain 1 which is absorbent with any t-norm.

Moreover, in order to rank the documents, the document-score should take into account each individual term-score, while the $\min$ operator only considers one. This is why (6) has been relaxed into (12). Then, each individual term score takes part in the final score, leading to an accurate IR model.

However, this consideration led us to consider another approach, more focused on the terms from q present in the documents: the cardinality-based approach. It consists in computing the (fuzzy) cardinality of the intersection between q and d, normalized by the (fuzzy) cardinality of q. Thus, by construction, this ratio focuses on the elements of q present in d, as in classical IR models.

E. Features of the cardinality-based approach

The remarks made above about the weights of the terms in the queries and the documents also apply in this case. As to the T-norm involved in formula (13), the same range as in the previous case has been tested, i.e., min, drastic, Einstein, Łukasiewicz, and product.

It is worth noticing that, when the product is chosen for $\top$ in (13), the formula corresponds to the one of most Vector Space Models used in IR:

$$S(d,q) = \sum_{t \in q} w_q(t) \cdot w_d(t) \ . \quad (15)$$

In particular, using (14) for $w_d(t)$ and $w_q(t) = (k_3+1) \cdot \text{tf}_q(t)/(k_3 + \text{tf}_q(t))$ (with $k_3 = 1000$), (15) correspond to the best-scoring, state-of-the-art VSM known as OKAPI-BM25 [20]. In our model, this score has to be normalized.

This interesting result shows that our cardinality-based approach can be viewed as a generalization of the IR Vector-Space Models.

V. Experimental results

Experiments have been carried out varying the different parameters. Then, in each case, the results have been compared to those of an OKAPI-like IRS, which is considered the best model for now.

A. Implication-based approach

This section shows both positive results, which validate the proposed model, and negative ones, along with explanations of the causes, for the implication-based model. Part of these results have been given in [3]. Absorption, zero and threshold properties are responsible for most of the poor results obtained and are discussed hereafter.

1) Zero property of T-norms: Conjunctive aggregation suffers from the zero property: if one term is scored 0, the document gets 0, whatever the score of the other terms.

A term score is given by: $w_q(t) \rightarrow w_d(t)$. With most R-implications, this score is 0 as soon as $w_d(t)$ is 0, i.e. the term is absent from the document. The situation is better with S-implications, as the score is $1 - w_q(t)$ in this case. To deal with this problem, the adopted strategy is the same than in language modeling approaches with smoothing techniques: if a word does not occur in a document, his weight is a small predefined and strictly positive value. It means that a term, even absent from a document *may be* representative of this document (as it is the case for synonyms).

2) Threshold property of R-implications: With R-implication, $w_q(t)$ is the required degree for $\mu_d(t)$ in totally satisfactory documents. As a consequence, as soon as $w_d(t) > w_q(t)$, the score for t, namely $w_q(t) \rightarrow w_d(t)$ is 1.

And as a bad consequence, two documents with different weights $w_{d_1}(t) \neq w_{d_2}(t)$ both above the threshold $w_q(t)$ get the same score 1 and cannot be ranked. This is why R-implications lead to poor results, and should not be used

INIST	OKAPI	IRS using an implication-based graded inclusion			
implic.		Reichenbach		Łukasiewicz	
t-norm		Einstein %	Product %	Łukasiewicz %	Product %
NIAP	21.75	**23.22 (+6.79%)**	**23.13 (+6.37%)**	**0.03 (-99.85%)**	**23.03 (+5.90%)**
IAP	24.13	**25.60 (+6.10%)**	**25.50 (+5.70%)**	**0.20 (-99.17%)**	25.38 (+5.17%)
Rprec	25.85	28.20 (+9.09%)	27.94 (+8.08%)	**0.03 (-99.90%)**	**28.09 (+8.69%)**
P5	50.00	45.33 (-9.33%)	49.33 (-1.33%)	**0.00 (-100.00%)**	48.00 (-4.00%)
P10	42.67	42.67 (0.00%)	42.00 (-1.56%)	**0.00 (-100.00%)**	43.67 (+2.34%)
P100	17.03	**18.27 (+7.24%)**	**18.20 (+6.85%)**	**0.03 (-99.80%)**	**18.23 (+7.05%)**
P500	5.39	**5.64 (+4.70%)**	**5.61 (+4.08%)**	**0.03 (-99.38%)**	**5.63 (+4.58%)**

TIPSTER	OKAPI	IRS using an implication-based graded inclusion			
implic.		Reichenbach		Łukasiewicz	
t-norm		Einstein %	Product %	Łukasiewicz %	Product %
MAP	18.14	18.61 (2.61%)	18.66 (2.87%)	**2.53 (-86.08%)**	18.66 (2.87%)
IAP	20.09	**20.83 (3.69%)**	20.90 (4.06%)	**2.70 (-86.55%)**	20.90 (4.02%)
Rprec	22.42	22.85 (1.91%)	23.31 (4.00%)	**3.47 (-84.54%)**	23.32 (4.02%)
P5	31.60	32.40 (2.53%)	32.80 (3.80%)	**5.60 (-82.28%)**	32.80 (3.80%)
P10	30.40	32.00 (5.26%)	31.80 (4.61%)	**6.00 (-80.26%)**	32.00 (5.26%)
P100	17.14	17.14 (0.00%)	17.08 (-0.35%)	**3.64 (-78.76%)**	17.06 (-0.47%)
P500	7.33	7.37 (0.49%)	7.34 (0.11%)	**0.85 (-88.43%)**	7.35 (0.27%)

TABLE I

RESULTS WITH THE IMPLICATION-BASED IRS

INIST	OKAPI	IRS using a cardinality-based graded inclusion			
t-norm		Einstein %	Łukasiewicz %	Min %	Product %
NIAP	21.75	22.82 (+4.91%)	**12.05 (-44.42%)**	20.80 (-4.34%)	**23.17 (+6.55%)**
IAP	24.13	25.33 (+4.99%)	**14.85 (-38.29%)**	23.19 (-3.91%)	**25.65 (+6.32%)**
Rprec	25.85	**27.39 (+5.95%)**	**15.47 (-40.16%)**	25.43 (-1.62%)	**28.65 (+10.84%)**
P5	50.00	46.67 (-6.67%)	**36.00 (-28.00%)**	45.33 (-9.33%)	**45.33 (-9.33%)**
P10	42.67	43.67 (+2.34%)	**28.67 (-32.81%)**	39.33 (-7.81%)	42.67 (0.00%)
P100	17.03	**18.00 (+5.68%)**	**10.13 (-40.51%)**	17.27 (+1.37%)	**18.33 (+7.63%)**
P500	5.39	**5.62 (+4.33%)**	**3.78 (-29.83%)**	**5.61 (+4.08%)**	**5.69 (+5.69%)**

TIPSTER	OKAPI	IRS using a cardinality-based graded inclusion			
t-norm		Einstein %	Łukasiewicz %	Min %	Product %
MAP	18.14	**18.01 (-0.71%)**	**16.10 (-11.23%)**	18.10 (-0.22%)	18.14 (+0.04%)
IAP	20.09	**20.01 (-0.39%)**	**17.93 (-10.72%)**	20.03 (-0.28%)	20.09 (-0.01%)
Rprec	22.42	22.28 (-0.62%)	**20.08 (-10.43%)**	22.39 (-0.14%)	22.45 (+0.14%)
P5	31.60	**30.00 (-5.06%)**	**28.80 (-8.86%)**	31.60 (0.00%)	31.60 (0.00%)
P10	30.40	30.20 (-0.66%)	**28.60 (-5.92%)**	30.40 (0.00%)	30.40 (0.00%)
P100	17.14	17.06 (-0.47%)	**15.50 (-9.57%)**	17.06 (-0.47%)	17.14 (0.00%)
P500	7.33	7.32 (-0.22%)	**6.92 (-5.67%)**	7.32 (-0.11%)	7.33 (0.00%)

TABLE II

RESULTS WITH THE CARDINALITY-BASED IRS

in the general case. However, if the weights in the queries $w_q(t)$ are chosen higher than the weights in the documents, the threshold is never reached, and results obtained with R- and S-implication are comparable.

3) Absorption property of min-like operators: Some aggregation operators have an absorption effect, as $\min$, $\max$... With this class of operators, only one term (or few terms) is taken into account to compute the score of a document. Here again, the consequence is that documents cannot be accurately ranked and thus lead to poor results.

4) Results: Among the many possible combinations of implications, aggregations, etc., only some (positive and negative) representative figures are given here. Table I presents the results for Reichenbach implication associated with Product or Einstein T-norm, and for Łukasiewicz implication, associated with Product or Łukasiewicz T-norm. The bold values are those considered as statistically significant according to a t-test.

Unsurprisingly, when the different parameters are chosen to avoid the above-mentioned unwanted properties, our IRS obtains positive results, comparable —and in some cases slightly better— than those of OKAPI.

For both collections, the best results are obtained using Reichenbach implication and Einstein or Product T-norm. In some cases, Łukasiewicz implication, and Larsen pseudo-implication (product) have also given good results. Some parameterized implications gave good results, but mainly when their behavior was close to the product.

B. Cardinality-based approach

Unsurprisingly, when the parameters are set to match the OKAPI-BM25 formula, the results are identical.

To allow for a comparison between our two models, the results presented here, in Table II, have been obtained with the weighting scheme used for the implication-based IRS. The weights for the terms in the documents $w_d(t)$ are the ones from OKAPI-BM25, but the weights of the terms in the queries $w_q(t)$ are based on the term frequency, linearly normalized and bounded (between 0.5 and 0.9). Thus, only the T-norm remains to be set.

On the INIST collection, Einstein and product t-norms give results slightly better than the ones of OKAPI. Łukasiewicz t-norm shows bad results. The min will be discussed later.

The results on the TIPSTER collection are less representative here. Indeed, there are no *associated concepts* in the queries, and only the subject is used. Then, queries are expressed with few keywords, and the term frequency is most often 1. This explains why the min (wrongly) seems to be a good operator here. As most of the weights in the documents are low, $min(0.9, w_d(t)) = w_d(t)$, while in OKAPI $1 \cdot w_d(t)) = w_d(t)$. It means that, most of the time, only the $w_d(t)$ are taken into account in (13), when using the min t-norm. When the $w_q(t)$ are different, which occurs with long queries, where representative terms are given several times, the term frequency is not taken into account with min, yielding poor results. With the TIPSTER collection, only one of the 50 queries gave different results than OKAPI. There were more numerous in the INIST collection, leading to worse results. For the same reasons, the product gave results almost identical to the OKAPI ones.

To conclude, Einstein and product t-norms in (13) give results rivaling with the ones of OKAPI (as expected, since the formulas are like-looking), as for the implication-based approach. Other t-norms give worse results in the general case.

VI. EXPRESSIVENESS OF THE MODEL

The proposed model has been tested using classical weighting mechanisms in order to validate our approach, but has not been entirely exploited yet. Better results are expected using user-defined weights for queries terms. Indeed, the frequency of query terms does not accurately represent the user's need; yet, asking for a user to weight his terms with real numbers is not generally feasible. The proposed graded approach makes it possible to simplify the manual weighting: for example, the user can just rank the query terms by importance, using an ordinal scale, or he can classify them into a few importance categories (e.g. filling 3 or 5 fields

in a formular). In both cases, numerical weights may be automatically given, representing their relative importance, according to the user simplified representation.

Queries can also be expanded using related terms (synonyms, hypernyms…). This kind of expansion is often done (e.g. [21]), but remains the problem of accurately weighting the added terms, or finishing a chain (the hypernyms of the hypernyms of the…). In our framework, new terms could be weighted relatively to the original terms using a notion of semantic proximity. First experiments, based on ideas from [2], consisting in enlarging the dividend (the documents) in the fuzzy division, have shown good results.

They could also be linked using fuzzy operators, like disjunctions (e.g. meaning that a term OR one of its synonyms is required). For instance, "*air pollution, greenhouse effect*" should give better results represented by (*air* AND *pollution*) OR (*greenhouse* AND *effect*) than using the 4 terms independently. The rich set of operators in FL allows here to modulate the meaning of the conjunctions and disjunctions. For instance, min/max carry the notion of independency. Other operators, like product/probabilistic sum, carry the notion of reinforcement. Using the probabilist sum (instead of max), the more associated concepts in a document, the better its score.

If most of these proposed extensions are not really original ones, they will rely on the well founded proposed approach. Then, operators, weights, and obtained results will benefit from a clear semantics. This could help improve the results.

Besides, some theoretical results, which would enrich the model, remain to be experimentally validated. For instance, negative terms can also be added to refine the query, and processed by an operation of antidivision [2]. The antidivision of $C(A, X)$ by $Q(A)$, dual from the division, retrieves the elements x in C such that $\forall a \in Q$, $(a, x) \notin C$ thus, in our model, the documents which do not contain the negative terms.

VII. Concluding remarks

The graded-inclusion IR model proposed in this paper seems promising. It has been shown that, with adequate settings, the implication-based model is able to mimic state-of-the-art systems, yet keeping its strong theoretical background. The cardinality-based model, while it does not really correspond to the division of fuzzy relations, is an extension of the Vector Space Model and, with appropriate settings, corresponds to the very efficient OKAPI-BM25 scheme. Note that language-modeling systems could be mimicked as well using probabilities (smoothed maximum-likelihood estimates) as degrees of membership and a product T-norm. Necessary properties the fuzzy operators must have in order to perform well in an IR context have also been identified.

Maybe more interesting, this fuzzy approach also provides new ways to build and handle expressive queries. In particular, easy and intuitive weighting procedures can be applied with our graded-inclusion model. Unfortunately, large-scale experimental validation of such techniques is hard to obtain due to the lack of suitable IR collections.

Apart from the perspectives already mentioned in the previous section, several other issues still have to be investigated. For instance, the use of qualitative or quantitative exception tolerant inclusions [1] to obviate the zero property of T-norms should be explored.

References

[1] P. Bosc and O. Pivert, "On the use of tolerant graded inclusions in information retrieval," in *Proceedings of CORIA'08*, 2008, pp. 321–336.

[2] ——, "On a parameterized antidivision operator for database flexible querying," in *Proceedings of DEXA'08*, 2008, pp. 652–659.

[3] P. Bosc, V. Claveau, O. Pivert, and L. Ughetto, "Graded-inclusion-based information retrieval systems," in *Proceedings of ECIR'09*, to appear, pp. 321–336.

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

[5] D. H. Kraft, G. Pasi, and G. Bordogna, "Vagueness and uncertainty in information retrieval: how can fuzzy sets help?" in *Proceedings of IWRIDL'06*, 2006, pp. 1–10.

[6] M. Boughanem, Y. Loiseau, and H. Prade, "Improving document ranking in information retrieval using ordered weighted aggregation and leximin refinement," in *Proceedings of EUSFLAT'05*, 2005, pp. 1269–1274.

[7] E. Herrera-Viedma, "Modeling the retrieval process for an information retrieval system using an ordinal fuzzy linguistic approach," *Journal of the American Society for Information Science and Technology*, vol. 52, pp. 460–475, 2001.

[8] A. Brini, M. Boughanem, and D. Dubois, "A model for information retrieval based on possibilistic networks," in *Proceedings of SPIRE'05*, 2005, pp. 271–282.

[9] E. Herrera-Viedma, A. Lopez-Herrera, M. Luque, and C. Porcel, "A fuzzy linguistic IRS model based on a 2-tuple fuzzy linguistic approach," *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 15, no. 2, pp. 225–250, 2007.

[10] M. Oussalah, S. Khan, and S. Nefti, "Personalized information retrieval system in the framework of fuzzy logic," *Expert Systems with Applications*, vol. 35, pp. 423–433, 2008.

[11] M. Lalmas, "Logical models in information retrieval: Introduction and overview," *Information Processing & Management*, vol. 34, no. 1, pp. 19–33, 1998.

[12] W. Waller and D. Kraft, "A mathematical model of a weighted Boolean retrieval system," *Information Processing & Management*, vol. 15, pp. 235–245, 1979.

[13] D. Buell and D. Kraft, "Threshold values and Boolean retrieval systems," *Information Processing & Management*, vol. 17, pp. 127–136, 1981.

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

[15] P. Bosc, D. Dubois, O. Pivert, and H. Prade, "Flexible queries in relational databases – the example of the division operator," *Theoretical Comp. Sc.*, vol. 171, pp. 281–302, 1997.

[16] P. Bosc, D. Rocacher, and O. Pivert, "Characterizing the result of the division of fuzzy relations," *Int. J. of Approximate Reasoning*, vol. 45, pp. 511–530, 2007.

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

[18] V. Young, "Fuzzy subsethood," *Fuzzy Sets & Systems*, vol. 77, pp. 371–384, 1996.

[19] A. D. Luca and S. Termini, "A definition of non-probabilistic entropy in the setting of fuzzy sets theory," *Inform. Control*, vol. 17, pp. 301–312, 1972.

[20] K. S. Jones, S. Walker, and S. Robertson, "A probabilistic model of information retrieval: Development and comparative experiments," *Information Processing & Management*, vol. 36, pp. 779–808, 809–840, 2000.

[21] E. Voorhees, *C. Fellbaum (ed.), WORDNET: An Electronic Lexical Database*. The MIT Press, 1998, ch. Using WORDNET for Text Retrieval, pp. 285–303.