Graded-Inclusion-Based Information Retrieval Systems

Patrick Bosc[1], Vincent Claveau[2], Olivier Pivert[1], and Laurent Ughetto[3]

[1] IRISA - ENSSAT, BP 80518, F-22305 Lannion, France
{bosc,pivert}@enssat.fr
[2] IRISA - CNRS - Campus de Beaulieu, F-35042 Rennes cedex, France
[3] IRISA - Université de Rennes 2 - Campus de Beaulieu, Rennes, France
{vincent.claveau,laurent.ughetto}@irisa.fr

Abstract. This paper investigates the use of fuzzy logic mechanisms coming from the database community, namely graded inclusions, to model the information retrieval process. 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.

Key words: IRS models, fuzzy logic, graded inclusion, fuzzy implication, query expressiveness

1 Introduction

Information retrieval and database querying share the same goal of providing a user with information he is asking for. Yet, it is well known that classical approaches used in a database context are not suited to the specificities of information retrieval: first, they do not provide the needed flexibility to the approximate matching between the textual queries and the documents, and secondly, they rarely offer a way to rank the returned results as it is usual in IR. However, recent studies in the field of database interrogation and fuzzy logic have provided new querying mechanisms that may be adapted to retrieve documents. Following the recent work of [1], this paper investigates the use of one of these mechanisms —the graded inclusion, which is presented in greater depth in the next section— in textual information retrieval. In this model, documents and queries are represented by fuzzy sets, which are matched using a graded inclusion, that is, using fuzzy operations like fuzzy implications and T-norms.

The first goal of this study is to provide some insights on the practical use of the graded inclusion in IR. Thus, through the experiments reported in Section 5, numerous settings were explored, using numerous pairs of implication and T-norm. In this fuzzy-logic framework, documents are represented as fuzzy sets of words. Then, well-known weighting schemes have been adapted to automatically assign the degrees of membership of these words. The positive results obtained show that, with appropriate settings (fuzzy operators and weighting schemes), it is possible for our fuzzy-based model to mimic classical systems and thus to yield results rivaling state-of-the-art ones. It is also possible to determine which of the different possible settings are actually suited or not to build a retrieval system, and from negative obtained results some insights can be given about the properties that are needed by such a model.

Apart from these experimental results, this study shows that this fuzzy-based IRS model makes it possible to foresee a better interaction with the user, with more expressive queries. Particularly, it is shown how this model can provide the user with a theoretically grounded way to easily weight query terms and to include negative information.

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 then presented in Section 3. The practical implementation and experimental results of this approach are detailed and discussed in Sections 4 and 5, and several theoretical extensions allowed by this framework are proposed in Section 6.

2 Related work

Parts of Fuzzy Logic (FL) theory have been used in IR models since the early 80s [2, inter alia]. This is rather natural since the Boolean IR model has been extended to graded ones, as FL is an extension of Boolean logic using grades of membership. Many studies have introduced FL in IR models, pursuing different goals. For instance, it has been used to managed uncertainty in the terms representation [3], to improve the ranking of the documents [4], to enhance 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 [5], or to use both possibility and necessity measures for terms weights [6]. Most of these papers are not really related to our approach.

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. [7] or Oussalah et al. [8]. The latter work is close to ours: it also proposes the use of fuzzy implications to compute a similarity measure between a document and a query. In their approach, $D \rightarrow Q$ is computed, as it is common in logical approaches to IR (see [9]), while here the implication is used the other way round, computing an inclusion degree of query words in the document (as explained in the next section). Our approach is also related to [10], as it extends the Boolean model and is in between the Boolean and Vector-Space models.

3 Information retrieval 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 graded inclusion, which is at the heart of our approach, can be derived as a generalization of the simple yet well-known Boolean model. The next subsections successively present the Boolean approach and its link with the division of relations, and how the extension to graded (or fuzzy) relations is linked to a graded IR approach. At last, the theoretical basis of our IRS is given.

3.1 A Boolean approach

In the Boolean model, a document can be considered as a set d of terms, and similarly a query can be represented by the set of its expected keywords P and the set of its excluded keywords N. In this framework, two operations are required to decide whether the document is relevant or not: P must be contained in d ($P \subseteq d$) and no element must be a member of both d and N ($d \cap N = \emptyset$). This shows the central role played by set operations (inclusion and intersection) in such IR systems.

In the framework of the relational model of data, a universe is modeled as a set of relations (in a mathematical sense, i.e., a relation C_i is a subset of the Cartesian product of some domains) which can be manipulated with the help of specific operators known as the relational algebra (set operations, selection, projection...). Among these operations, the division of the relation $C(A, X)$ by $Q(A)$ denoted by $C[A \div A]Q$, where A is a set of attributes common to C and Q, aims at determining the X-values connected in C with all the A-values appearing in Q. This operation can be defined equivalently in the following ways:

$$x \in C[A \div A]Q \quad \Leftrightarrow \quad \forall a \in Q,\ (x, a) \in C \tag{1}$$

$$x \in C[A \div A]Q \quad \Leftrightarrow \quad Q \subseteq \Omega^{-1}(x) \quad \text{where} \quad \Omega^{-1}(x) = \{a | (x, a) \in C\} \tag{2}$$

Let us 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. Moreover, let us restrict to the case in which a query q looks for those documents indexed by a set of expected terms $P = \{t'_1, \ldots, t'_n\}$. The set of documents of the archive may be represented as an unnormalized relation (C_U) where a tuple has the form: $\langle d, t_1, \ldots, t_m \rangle$ or as a normalized relation (C_N) where the information stored in the previous tuple is split through m tuples: $\langle d, t_1 \rangle, \ldots, \langle d, t_m \rangle$. The keywords appearing in the query may be seen as a unary relation (P) and the query may be answered as the division of C_N by P.

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.

3.2 A graded approach

Let us now describe two main aspects of a fuzzy model of information retrieval. In such a model, the retrieval function can be formalized in two steps. In the first step the function E evaluating queries constituted by a single (weighted) term is defined: $E: \; D \times Q' \rightarrow [0,1]$ in which Q' is the set of queries with a single (weighted) term. Function E computes the Retrieval Status Value (RSV) constituting the degree to which a document d matches a query $q \in Q'$. In the second step, a function $E*$ is defined as: $E*: \; D \times Q \rightarrow [0,1]$ (where Q is the set of all the legitimate queries) which evaluates the final RSV of a document, reflecting the satisfaction of the whole query; by interpreting the operators AND, OR and NOT, as fuzzy intersection, union and complement respectively.

The first step towards a fuzzy IR model was to extend the representation within fuzzy set theory by associating with each document-term pair a weight $F(d,t) \in [0,1]$, named index term weight, indicating the degree of aboutness or significance $F(d,t)$ of document d with respect to term t [11, 12]. The computation of $F(d,t)$ is generally based on the number of occurrences of t in the document d and in the whole archive D. The introduction of the index term weight made it possible to represent a document as a fuzzy set of terms [2]: $C(d) = \{\mu_d(t)/t, t \in T\}$ in which $\mu_d(t) = F(d,t)$. The notation α/t is the classical one for fuzzy sets defined on a discrete universe and means: value t with membership degree α (here, $\alpha = \mu_d(t)$). Based on this fuzzy representation of documents, the retrieval mechanism has been extended with the ability to rank the retrieved documents in decreasing order of their significance with respect to the user query. In fact, in this case the retrieval function evaluating an atomic query consisting of a single term t yields $F(d,t)$: $E(d,q) = F(d,t) \quad \forall q = t, \; t \in T \cup Q'$.

To make the Boolean query language less limited in its expressiveness, a fuzzy IR model such as that described in [13] extends atomic selection criteria by introducing query term weights. An example of Boolean weighted (or fuzzy) query is: $\langle t_1, w_1 \rangle$ AND ($\langle t_2, w_2 \rangle$ OR $\langle t_3, w_3 \rangle$) in which t_1, t_2, t_3, are search terms and w_1, w_2, $w_3 \in [0,1]$ are numeric weights. The concept of query weights has raised the problem of their interpretation: several authors have realized that the semantics of query weights should be related to the concept of "importance" of the terms. The semantics of the weights determines the definition of function E; as weights are introduced at the level of single query terms, function E is defined on the sets D and Q' in which $Q' = T \in [0,1]$. Function E is then evaluated for a document $d \in D$, a term $t \in T$ and its query weight $w \in [0,1]$.

3.3 Division, graded inclusions and fuzzy implications

The answer to a query q may be devised as the generalization of the Boolean case described in Section 3.1, namely the division of two fuzzy relations (i.e. relations with weighted tuples) C and Q. In this case, the result of the division is defined as a fuzzy set, i.e. a fuzzy relation $C[T \div T]Q$, and a natural extension stems from expression (2) where the usual set inclusion operator is changed into a grade of inclusion g: $C[T \div T]Q(d) = g(Q \subseteq \Omega^{-1}(d))$, $\Omega^{-1}(d)$ being a fuzzy set of keywords defined as: $\Omega^{-1}(d) = \{\mu/t | \mu/(d,t) \in C\}$. Then the semantics of the division depends on both the choice of the inclusion grade and the intended meaning of the weights associated with the tuples in relations C and Q [14].

A view of the inclusion consists in defining the grade of inclusion $g(Q \subseteq \Omega^{-1}(d))$ using a fuzzy implication (denoted by → in the following), which leads to the indice:

$$g(Q \subseteq \Omega^{-1}(d)) = \min_{t \in Q} \ (\mu_Q(t) \rightarrow \mu_C(d,t)) \ . \tag{3}$$

Two different interpretations may be distinguished depending on the nature of the interaction of the degrees in the two relations.

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) [15], denoted by $\rightarrow_R$, and defined as:

$$p \rightarrow_R q \ = \ \sup \ \{u \in [0,1] | \top(p,u) \leq q\} \ , \tag{4}$$

where $\top$ stands for a triangular norm. Any R-implication may be rewritten:

$$p \rightarrow_R q = 1 \text{ if } p \leq q, \ f(p,q) \text{ otherwise,} \tag{5}$$

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 minimal element of this class of implications is Gödel implication: $p \rightarrow_{Gd} q = 1$ if $p \leq q$, q otherwise, which is obtained by choosing $\top(a,b) = min(a,b)$ in formula (4). Other representatives of R-implications are Goguen (respectively Lukasiewicz) implication obtained with $\top(a,b) = a.b$ (respectively $\max(a+b-1,0)$): $p \rightarrow_{Gg} q = 1$ if $p \leq q$, q/p otherwise, $p \rightarrow_{Lu} q = 1$ if $p \leq q$, $1-p+q$ otherwise. The threshold interpretation of $\mu_Q(t)$ with R-implications is clear from formula (5), where the satisfaction degree is 1 as soon as $\mu_C(d,t)$ reaches $\mu_Q(t)$.

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 [15] denoted by $\rightarrow_S$, as follows ($\perp$ stands for a triangular conorm):

$$p \rightarrow_S q \ = \ \perp(1-p,q) = 1 - \top(p, 1-q) \tag{6}$$

As it is the case for R-implications, there exists an infinity of such implications and their most commonly used representative, Kleene-Dienes implication, is defined as: $p \rightarrow_{KD} q = \max(1 - p, q)$. It is the minimal element obtained from (6) with the smallest T-conorm $\perp$, i.e., the maximum. Using the probabilistic sum, one gets Reichenbach implication: $p \rightarrow_{Rb} q = 1 - p + p \cdot q$. It turns out that Lukasiewicz implication is also an S-implication obtained with $\perp(a, b) = \min(a+b, 1)$. One can notice that the regular division is recovered from formula (3) in the presence of regular relations due to the fact that any fuzzy implication coincides with the usual one in that case (in particular $1 \rightarrow 0 = 0$ and $1 \rightarrow 1 = 1$).

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

Example 1. Table 1 shows the fuzzy relations representing a collection of 2 documents d_1 and d_2, 2 queries q and q', and the results depending on the chosen semantics. The threshold effect clearly appears with q' and d_2.

	t_1	t_2	t_3	t_4
d_1	1	0.9	1	0.2
d_2	0.7	0.6	0.3	0.8

	t_1	t_2	t_3	t_4
q	1	0.4	0	0.6
q'	0.6	0.6	0.3	0.5

	query weight	semantics	d_1	d_2
q	importance	Kleene-Dienes	0.4	0.6
		Reichenbach	0.52	0.76
q'	threshold	Gödel	0.2	1
		Goguen	0.4	1
		Lukasiewicz	0.7	1

Table 1. top left: fuzzy relation C representing a collection — top right: each row is a fuzzy relation Q representing a query — bottom: results

4 Implementation and experiments

4.1 Document collections

The proposed IRS has been tested on 3 collections. The first one, hereafter called ELDA, is a small collection in French containing 3,499 documents (question/answer of the European commission) and a set of 29 queries. The second one is the INIST collection which contains 163,308 documents (paper abstracts from various scientific disciplines) and a set of 30 queries. Both collections come from the IR evaluation campaign Amaryllis. The third one is the Wall Street Journal subcollection from the TREC-3 TIPSTER collection, which contains 173,252 documents and a set of 50 queries. For all collections, documents and queries have been lemmatized. The queries are composed of several fields: 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 (except in TIPSTER where concepts are not available).

4.2 IRS features

Our IRS implements the fuzzy approach described in Section 3. Thus, the score of a document d in front of a query q is computed as follows:

$$S(d,q) = \top_{t \in q}(w_q(t) \rightarrow w_d(t)) \ , \tag{7}$$

where t is a term (in the query), $w_q(t)$ is weight in the query, $w_d(t)$ (which can also be denoted $w_C(d,t)$) its weight in the document, → the fuzzy implication underlying the chosen graded inclusion, and $\top$ the aggregating T-norm. From formula (7), one can see that several parameters can be tuned: the weights of the terms in both the queries and the documents, and the aggregation and implication operators.

Aggregation operator. When $\top$ is the min operation, the obtained degree $S(d,q)$ is the degree to which d belongs to the quotient of the fuzzy division of the set of documents by the query s. As min is the largest T-norm, it gives the maximal degree $S(d,q)$. This degree then corresponds to the inclusion degree of the term $t \in q$ which is the least included in d. Thus, it corresponds to a classical DB point of view, where each term of the query is expected to be in the retrieved document. The degree of the weakest term reflects the acceptability of the document. As explained below, this approach fails in IR.

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". 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 term. This is why equation (3) has been relaxed into (7), which remains an inclusion measure, and a large range of T-norms has been tested, as for instance, min, Drastic, Einstein, Lukasiewicz, Product, as well as some parameterized T-norms: Dubois and Prade, Hamacher, Yager...

Graded inclusion operator. As mentioned in Section 3, 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, Lukasiewicz. Among the S-implications: Kleene-Dienes, Lukasiewicz, Reichenbach, Willmott. See for instance [15] for the definition of these operators.

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. However, in the context of fuzzy computations, and as it is explained above, the weights must belong to the $[0,1]$ interval. Thus, the OKAPI-BM25 weights $w_{BM25}(t,d)$ have been normalized and bounded.

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 satisfaction 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 relies on the frequency of the terms in the queries, and are normalized and bounded.

5 Experimental results

Experiments have been carried out varying the different parameters. In each case, the results have been compared with OKAPI. This section shows both positive results, which validate the proposed model, and negative ones, along with explanations of the causes.

5.1 Properties involving poor results

Absorption, zero and threshold properties are responsible for most of the poor results obtained. They can occur at different levels, and involve aggregation operators, implications and weights.

Zero property of T-norms. In our model, the individual scores given to terms are aggregated conjunctively. This aggregation suffers from the zero property: if one term is scored 0, the document is scored 0, whatever the score of the other terms.

A term score is given by: $w_q(t) \to w_d(t)$. 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. Thus, one gets a 0 score for the term (and hence the document) only if $w_q(t) = 1$, meaning that the t has a maximal importance in the query. 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).

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 term t, namely $w_q(t) \to w_d(t)$ 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. Here again, this is more a classical DB approach, where the system has to retrieve the relevant tuples only. In IR, the documents also have to be ranked. (Note that is is also the case in flexible querying of DB.) This is why R-implications lead to poor results, and should not be used 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.

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.

5.2 Results

Among the many possible combinations of implications, aggregations, etc., only some (positive and negative) representative figures are given here. Table 2 presents the results for Reichenbach implication associated with Product or Einstein T-norm ($a \top_E b = (a.b)/(2-a+b-a.b)$), and for Lukasiewicz implication, associated with Product or Lukasiewicz T-norm ($a \top_L b = \max(0, a+b-1)$).

The results are evaluated in terms of mean average precision (MAP), interpolated average precision (IAP), R-precision (Rprec) and precision on top-k lists (Pk); the bold values are those considered as statistically significant according to a t-test.

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

Operators. For both collections, the best results are obtained using Reichenbach implication and Einstein or Product T-norm. In some cases, Lukasiewicz 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.

Interestingly enough, T-conorms (disjunctions), used to aggregate the scores of the terms have given results similar (while worse in general) than the associated T-norm. At first, it may seem surprising since their semantics is very different (all vs. at least one). However, it underlines that the important point is the way each individual term score is taken into account to make the document score. This way is often similar in the T-norm and its associated T-conorm.

6 Expressiveness of the graded-inclusion 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

INIST	OKAPI	Graded inclusion-based IRS							
implic.		Reichenbach				Lukasiewicz			
t-norm		Einstein	%	Product	%	Lukasiewicz	%	Product	%
MAP	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%)**

ELDA	OKAPI	Graded inclusion-based IRS							
implic.		Reichenbach				Lukasiewicz			
t-norm		Einstein	%	Product	%	Lukasiewicz	%	Product	%
MAP	57.14	56.86	(-0.49%)	56.91	(-0.42%)	**1.11**	**(-98.06%)**	56.29	(-1.50%)
IAP	58.09	57.89	(-0.36%)	57.88	(-0.37%)	**1.98**	**(-96.59%)**	57.38	(-1.23%)
Rprec	55.33	53.82	(-2.73%)	54.64	(-1.26%)	**0.67**	**(-98.78%)**	**53.03**	**(-4.16%)**
P5	77.24	76.55	(-0.89%)	74.48	(-3.57%)	**1.38**	**(-98.21%)**	75.17	(-2.68%)
P10	68.28	68.62	(+0.51%)	68.97	(+1.01%)	**0.69**	**(-98.99%)**	67.93	(-0.51%)
P100	27.00	26.86	(-0.51%)	26.83	(-0.64%)	**1.00**	**(-96.30%)**	26.83	(-0.64%)
P500	6.67	6.66	(-0.10%)	6.67	(+0.00%)	**0.87**	**(-86.97%)**	6.66	(-0.10%)

TIPSTER	OKAPI	Graded inclusion-based IRS							
implic.		Reichenbach				Lukasiewicz			
t-norm		Einstein	%	Product	%	Lukasiewicz	%	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 2. Results with the ELDA, INIST and TIPSTER collections

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. [16]), 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 for instance a notion of distance. They could also be linked using fuzzy operators, like disjunctions (e.g. meaning that a term OR one of its synonyms is required). This would lead to more complex queries. And our theoretically grounded framework also allows for complex, while semantically sound, queries, using fuzzy conjunction or disjunction operators [7]. It is of interest to accurately take into account the associated concepts of the queries. 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. In: $\max(\min(\mu_d(air), \mu_d(pollution)), \min(\mu_d(greenhouse), \mu_d(effect)))$ the disjunction max means that both associated concepts are not required, only the "best" one makes the score; the max means that both terms in a concept are required, as the "worst" one makes the score, e.g. "*effect*" without "*greenhouse*" leads to a low score. 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 increase the results.

Besides, some theoretical results, which would enrich the model, remain to be experimentally validated. For instance, negative terms can also been added to refine the query, and processed by an operation of antidivision [17]. 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 does not contain the negative terms.

7 Concluding remarks

The graded-inclusion IR model proposed in this paper seems promising. It has been shown that, with adequate settings, the model is able to mimic state-of-the-art systems, yet keeping its strong theoretical background. 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 key components (fuzzy operators) of the model 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. Particularly, 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 suited IR collections.

Apart from the perspectives already mentioned in the previous section, several other issues concerning our model should be investigated. For instance, the use of qualitative or quantitative exception tolerant inclusions [1] to deal with the zero property of T-norms should be explored. Last, another interesting issue is the study of the inclusion model itself. From the experiments, it is clear that current IR mechanisms focus mainly on the intersection between documents and queries, while DB ones usually focus on the inclusion, i.e. the terms from the query outside the document. Cardinality-based inclusions could bridge the gap since they are more focused on the common elements between two fuzzy sets. Studying fuzzy intersection-based models also seems of interest.

References

1. Bosc, P., Pivert, O.: On the use of tolerant graded inclusions in information retrieval. In: Proceedings of CORIA'2008. (2008) 321–336
2. Buell, D.: An analysis of some fuzzy subset applications to information retrieval systems. Fuzzy Sets & Systems **7** (1982) 35–42
3. Kraft, D.H., Pasi, G., Bordogna, G.: Vagueness and uncertainty in information retrieval: how can fuzzy sets help? In: Proceedings of IWRIDL'2006. (2006) 1–10
4. Boughanem, M., Loiseau, Y., Prade, H.: Improving document ranking in information retrieval using ordered weighted aggregation and leximin refinement. In: Proceedings of EUSFLAT'2005. (2005) 1269–1274
5. Herrera-Viedma, E.: 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 **52** (2001) 460–475
6. Brini, A., Boughanem, M., Dubois, D.: A model for information retrieval based on possibilistic networks. In: Proceedings of SPIRE'2005. (2005) 271–282
7. Herrera-Viedma, E., López-Herrera, A., Luque, M., Porcel, C.: A fuzzy linguistic IRS model based on a 2-tuple fuzzy linguistic approach. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems **15**(2) (2007) 225–250
8. Oussalah, M., Khan, S., Nefti, S.: Personalized information retrieval system in the framework of fuzzy logic. Expert Systems with Applications **35** (2008) 423–433
9. Lalmas, M.: Logical models in information retrieval: Introduction and overview. Information Processing & Management **34**(1) (1998) 19–33
10. Salton, G., Fox, E., Wu, H.: Extended boolean information retrieval. Communications of the ACM **26**(12) (1983) 1022–1036
11. Waller, W., Kraft, D.: A mathematical model of a weighted Boolean retrieval system. Information Processing & Management **15** (1979) 235–245
12. Buell, D., Kraft, D.: Threshold values and Boolean retrieval systems. Information Processing & Management **17** (1981) 127–136
13. Bookstein, A.: Fuzzy requests: an approach to weighted Boolean searches. J. of the American Society for Information Science **31** (1980) 240–247
14. Bosc, P., Dubois, D., Pivert, O., Prade, H.: Flexible queries in relational databases – the example of the division operator. Theoretical Comp. Sc. **171** (1997) 281–302
15. Fodor, J., Yager, R.: Fuzzy Set-theoretic Operators and Quantifiers. Chap. 1.2. In: Fundamentals of Fuzzy Sets – The Handbook of Fuzzy Sets Series (D. Dubois and H. Prade eds.). Kluwer Academic Publishers (1999) 125–193
16. Voorhees, E.: Using WORDNET for Text Retrieval. In: C. Fellbaum (ed.), WORDNET: An Electronic Lexical Database. The MIT Press (1998) 285–303
17. Bosc, P., Pivert, O.: On a parameterized antidivision operator for database flexible querying. In: Proceedings of DEXA'2008. (2008) 652–659