On the use of tolerant graded inclusions in information retrieval

Patrick Bosc, Olivier Pivert

IRISA-ENSSAT

Technopole Anticipa, BP 80518

F-22305 Lannion cedex

bosc@enssat.fr, pivert@enssat.fr

RÉSUMÉ. *Dans cet article, un modèle de recherche d'information fondé sur la théorie des ensembles flous est considéré. Tout d'abord, nous montrons que le mécanisme de recherche dans un tel modèle peut être défini en termes d'inclusion graduelle. Cette approche est fortement liée à la notion de division dans un contexte de bases de données relationnelles. Dans un deuxième temps, nous mettons en évidence plusieurs axes d'extension de l'inclusion graduelle, l'objectif étant de rendre l'indicateur d'inclusion (et donc le mécanisme de matching document-requête) plus tolérant, aux exceptions notamment. Il est montré que l'utilisation de tels indicateurs d'inclusion tolérante permet de réduire le risque d'obtention de réponses vides.*

ABSTRACT. *In this contribution, a fuzzy-set-based information retrieval model is considered. First, we show that the retrieval mechanism of such a retrieval model can be defined in terms of graded inclusion. This approach derives from the notion of division of fuzzy relations in the framework of Database Management Systems. Then, we point out different lines of extension of the graded inclusion aimed at making it more tolerant (to exceptions, in particular). It is shown that the use of such tolerant inclusion indicators reduces the risk of obtaining empty answers.*

MOTS-CLÉS : *Modèle flou de recherche d'information, inclusion graduelle, inclusion tolérante, relations floues, termes d'index pondérés, requêtes pondérées.*

KEYWORDS: *Fuzzy Information Retrieval, graded inclusion, tolerant inclusion, fuzzy relations, index term weights, query weights.*

1. Introduction

In the literature, several fuzzy approaches to extend Boolean information retrieval systems have been defined (see, e.g., (Radecki, 1979), (Bookstein, 1980), (Buell, 1982), (Yager, 1987), (Bordogna et al., 1995)). All these approaches introduce weights to extend both the documents' representation (based on index terms) and the queries.

The objective of this contribution is to show how the notion of a tolerant graded inclusion can be of use in the query-document matching process, so as to make it more flexible. The idea is to consider that a set *E* (of keywords) can sometimes be viewed as (approximately) included in a set *F* even when some elements of *E* are not in *F*. In this paper, we point out different rationales that appear relevant to found such a notion of tolerant graded inclusion.

The paper will be organized as follows: in section 2, it will be shown that the inclusion operation plays a key role in Boolean IR, and the connections between Boolean information retrieval and the division of relations is established. In particular, in front of a set of expected keywords stated in a query, it is shown that the retrieval mechanism may be seen as the division of a binary relation describing the associations between documents and keywords by a relation containing the keywords of the query. In section 3, this view is generalized to the case of fuzzy relations, i.e., whose tuples are weighted, which gives birth to several types of semantics, depending on the meanings of the weights and the nature of their interaction. It will be seen that the query-document matching process is strongly connected with different possible interpretations of the inclusion degree of (fuzzy) sets. Section 4 is devoted to different lines of extension of the graded inclusion indicator so as to make it more tolerant in order to limit the risk of obtaining empty answers. The conclusion summarizes the main points of the paper and proposes some perspectives for future work.

2. Boolean information retrieval and the division of relations

IR systems are aimed at the handling of large sets of documents and retrieving those documents which correspond to a user need. Documents generally consist of texts which are indexed to represent their contents (keywords or index terms) and queries are based on the specification of terms used to identify topics of interest (Salton *et al.*, 1984). In most of the commercial IR systems based on the Boolean IR model, a query is used to find the documents which are related to a given topic, i.e., which contain (and/or do not) a given set of keywords. For example, one may look for documents talking about "fuzzy sets, fuzzy inclusion, fuzzy relations" excluding "possibility theory, measures". By describing a document through a set of keywords *d* and the query through the set of expected keywords *P* on the one hand and the set of excluded keywords *N* on the other hand, two operations are required to decide whether the document is relevant or not. In fact, *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 information retrieval.

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 R_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 $R(A, X)$ by $S(A)$ denoted by $R[A \div A]S$, where A is a set of attributes common to R and S, aims at determining the X-values connected in R with all the A-values appearing in S. This operation can be defined equivalently in the following ways:

- $x \in R[A \div A]S \Leftrightarrow \forall a \in S,\ (x, a) \in R$ (1)
- $x \in R[A \div A]S \Leftrightarrow S \subseteq \Omega^{-1}(x)$ where $\Omega^{-1}(x) = \{a \mid (x, a) \in R\}$. (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 (UR) where a tuple has the form: $<d, t_1, \ldots, t_m>$ or as a normalized relation (NR) where the information stored in the previous tuple is split through m tuples: $<d, t_1>, \ldots, <d, t_m>$. The keywords appearing in the query may be seen as a unary relation (P) and the query may be answered as the division of NR by P.

Example 1. Let us consider the relations in Table 1:

ARCHIVE

doc.	keyword
d_1	k_1
d_1	k_3
d_1	k_4
d_2	k_1
d_2	k_2
d_2	k_3
d_3	k_2
d_3	k_3

EXPECTED TERMS

keyword
k_1
k_2
k_3

Table 1: relations representing an archive and a query

The result of the division ARCHIVE [keyword÷keyword] EXPECTED-TERMS returns the document d_2, which corresponds to the only document containing at least the three desired keywords $\{k_1, k_2, k_3\}$.♦

3. Fuzzy information retrieval and the division of fuzzy relations

IR systems are based on models characterized by three main components: the representation of documents, the query language, and the matching mechanism.

The documents' representation is generated by the indexing process which produces surrogates of the document information content, generally consisting of index terms, manually associated with the documents or automatically extracted from them (Salton *et al.*, 1984).

Users submit their information needs to the system through queries expressed in the system query language. Generally, queries consist of atomic selection criteria (which are basically single terms, sometimes weighted), and aggregation operators (which can be either implicit in the query, as in the Vector Space model, or explicit).

The matching mechanism evaluates a user's query against the representations of documents and retrieves those documents which are considered to be relevant. The relevance assessments are determined by a retrieval function which is peculiar to each model. In the following subsections, the two main aspects of a fuzzy model of information retrieval are briefly described. 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.

3.1. *Fuzzy document representation*

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 (Waller *et al.*, 1979) (Buell *et al.*, 1981). 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 possible to represent a document as a fuzzy set of terms (Buell, 1982): $R(d) = \{\mu_d(t)/t, t \in T\}$ in which $\mu_d(t) = F(d, t)$. Based on this fuzzy documents' representation 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 $F(d, t)$:

$$E(d, q) = F(d, t) \quad \forall\ q = t \in T \cup Q'.$$

3.2. *Fuzzy queries*

To make the Boolean query language less limited in its expressiveness, a fuzzy IR model such as that described in (Bookstein, 1980) extends atomic selection criteria by introducing query term weights. An example of Boolean weighted (or fuzzy) query is: $<t_1, w_1>$ AND ($<t_2, w_2>$ OR $<t_3, w_3>$) 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 weight semantics 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 $Q' = T \times [0, 1]$. 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 2, namely the division of two fuzzy relations R and S. In this case, the result of the division is defined as a fuzzy set, i.e. a fuzzy relation $R[T \div T]S$, and a natural extension stems from expression 2 where the usual set inclusion operator is changed into a grade of inclusion g:

$$\mu_{R[T \div T]S}(d) = g(S \subseteq \Omega^{-1}(d)) \qquad (3)$$

$\Omega^{-1}(d)$ being a fuzzy set of keywords defined as:

$$\Omega^{-1}(d) = \{\mu/t \mid \mu/(d, t) \in R \text{ and } d \in D\}.$$

The notation μ/t expresses that the membership degree of t to to $\Omega^{-1}(d)$ equals μ.

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 R and S (Bosc *et al.*, 1997).

A view of the inclusion consists in defining the grade of inclusion $g(S \subseteq \Omega^{-1}(d))$ using a fuzzy implication (denoted by → in the following), and then we obtain the indice:

$$g(S \subseteq \Omega^{-1}(d)) = \min_{t \in S} (\mu_S(t) \rightarrow \mu_R(d, t)) \qquad (4)$$

Two different interpretations may be distinguished depending on the nature of the interaction of the degrees in the two relations. In the first case, the degree $\mu_S(t)$ is seen as a threshold and the complete satisfaction requires that this threshold is attained by $\mu_R(d, t)$ for each value t of S. When the threshold is not reached, a penalty is applied. This behavior is obtained using a residuated implication (or R-implication) (Fodor *et al.*, 1999), denoted by $\rightarrow_{\text{R-i}}$, and defined as:

$$p \rightarrow_{\text{R-i}} q = \sup \{u \in [0, 1] \mid \top(p, u) \leq q\} \qquad (5)$$

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

$p \rightarrow_{\text{R-i}} q = 1$ if $p \leq q, f(p, q)$ otherwise

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's implication:

$p \rightarrow_{\text{Gö}} q = 1$ if $p \leq q$, q otherwise

which is obtained by choosing $\top(a, b) = \min(a, b)$ in formula (5). Other representatives of R-implications are Goguen's (respectively Lukasiewicz') implication obtained with $\top(a, b) = a \times b$ (respectively $\max(a + b - 1, 0)$):

$p \rightarrow_{\text{Gg}} q = 1$ if $p \leq q$, q/p otherwise

$p \rightarrow_{\text{Lu}} q = 1$ if $p \leq q$, $1 - p + q$ otherwise.

In the second interpretation, $\mu_S(t)$ defines the importance of value t (and then the degree $\mu_R(d, t)$ 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 $\mu_S(t) < 1$ the requirement is not completely important, and it can be forgotten to some extent. The complete satisfaction requires that $\mu_R(d, t)$ equals 1 for each value t of S whatever its importance and $\mu_{R[T \div T]S}(d) = 0$ only if for at least one t in S, both $\mu_S(t) = 1$ (the requirement has the maximum level of importance) and $\mu_R(d, t) = 0$ (the tuple does not fulfill the requirement at all). This behavior is modeled by using an S-implication (Fodor *et al.*, 1999) denoted by $\rightarrow_{\text{S-i}}$, as follows:

$$p \rightarrow_{\text{S-i}} q = \perp(1 - p, q) = 1 - \top(p, 1 - q) \qquad (6)$$

As it is the case for R-implications, there exists an infinity of such implications and their most commonly used representative, Kleene-Dienes' implication, defined as:

$p \Rightarrow_{\text{KD}} q = \max(1 - p, q)$

is the minimal element obtained from expression 6 with the smallest co-norm, i.e., the maximum. Another S-implication is obtained from expression 6 using the probabilistic sum, namely Reichenbach's implication defined as:

$p \Rightarrow_{\text{Rb}} q = 1 - p + pq$.

It turns out that Lukasiewicz' implication is also an S-implication obtained from expression 6 with $\perp(a, b) = \min(a + b, 1)$.

One can notice that the regular division is recovered from formulas 3-4 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 $1 \rightarrow 1 = 1$).

3.4. *Comments and an example*

This approach is logical and conjunctive and an "absorption effect" occurs: the division operator only retains the smallest degree of implication between S and R. The S-grades, i.e. the query term weights w, can express either a threshold or an importance, which makes sense in the context of document retrieval. If we assume that the degree μ attached to a term t in a document d refers to the relevance of d with respect to t, the weight w tied to the expected term e stands for a minimal relevance with the threshold interpretation while it represents the importance of e with the second interpretation. Consequently, the solutions suggested before for the division of fuzzy relations may be an interesting basis for plausible interpretations of document retrieval.

Example 2. Let us consider the archive represented by the fuzzy relation in Table 2 and the queries q and q' represented by the fuzzy relations in Table 3.

	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

Table 2. Relation representing an archive of documents as a fuzzy relation R

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

Table 3. Each row is a fuzzy relation S representing a query

	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 4. Result of the queries of Table 3 referred to the archive of Table 2.

Depending on the semantics chosen, the result of q and q' are given in Table 4.♦

4. Fuzzy information retrieval and tolerant graded inclusions

4.1. *Some limitations of the classical inclusion*

It may happen that a query calling on a division – thus on an inclusion –, even of fuzzy relations, leads to an empty answer, while some elements would have been almost satisfactory if some tolerance could take place. This line is the basis for the design of tolerant inclusion operators, but it is important to notice that such operators can be used directly in user queries and not only to repair "failing" queries. We now review four types of situation where different ideas of tolerance make sense.

Let us consider that the query (relation *S*) contains 20 keywords ($t_1, \ldots, t_{20}$) and that the associations document-keyword are represented by the binary relation *R* as follows:

$\{<d_1, t_1>, <d_1, t_6>, <d_1, t_{20}>,$
$<d_2, t_1>, <d_2, t_2>, \ldots, <d_2, t_{19}>,$
$<d_3, t_1>, <d_3, t_2>, \ldots, <d_3, t_{18}>\}.$

Neither d_1, nor d_2, nor d_3 is satisfactory as to the division of *R* by *S*. Nevertheless, if it seems legitimate to think that d_1 is definitely inadequate, d_2 and d_3 are almost satisfactory since they are associated with respectively 19 and 18 terms of the query. Thus, one may be interested in distinguishing between these quite different situations through a tolerance to exceptions. An "all or nothing" approach will accept d_2 and d_3 provided that a 10% ratio of exceptions is allowed. It is also possible to adopt a graded view according to which exceptions are a matter of preferences and then their ratio (or number) is a matter of degree. For instance, satisfaction decreases from full acceptance (if exceptions are under 8%) to total rejection (above 12%).

In the previous situation, exceptions are treated on a quantitative basis, i.e., according to their number. So, it is impossible to compensate a large number of small exceptions, i.e., to account for the notion of low-intensity exceptions which may only occur in the context of fuzzy relations. For instance, let us consider the fuzzy relations:

$R = \{1/<d_1, t_1>, 1/<d_1, t_2>, 1/<d_2, t_1>, 0.8/<d_2, t_2>, 1/<d_3, t_1>, 0.9/<d_3, t_2>\},$

$S = \{1/t_1, 0.8/t_2, 0.1/t_3, \ldots, 0.1/t_{10}\}.$

If Gödel's or Goguen's implication is used, the result of the division of *R* by *S* is empty while *S* is "almost" included (in Zadeh's sense, i.e. $E \subseteq F \Leftrightarrow \forall x \in U, \mu_E(x) \leq \mu_F(x)$, where *E* and *F* are two fuzzy sets defined on the universe *U*) in the set of keywords associated with d_1, d_2 and d_3 in the dividend *R*. Of course, the notion of qualitative exceptions may be dealt with in a gradual way, i.e., an exception is more or less a low-intensity one, so as to prevent from a sharp behavior of the tolerance mechanism.

A third kind of approach to query relaxation is well-known in the field of information retrieval. It corresponds to taking into account the notion of synonymy between the keywords, using a thesaurus (Salton *et al.*, 1984). Then, it is possible to extend the query in the following way: a keyword *t* from the query is replaced by (*t* OR t'_1 OR ... OR t'_n) where the t'_i's are the synonyms of *t* present in the thesaurus. Here again, in general, the notion of synonymy may be a graded one in order to express nuances, and we describe in the following (Subsection 4.3.1) a tolerant graded inclusion that takes into account this notion of approximate synonymy.

In the preceding case, the tolerance conveyed by the query relaxation process may also be viewed as enlarging the dividend relation. One may think of a dual mechanism whose effect is to restrict the divisor. From a semantic point of view, such a modification can be based on the notion of "significance" of the elements of the divisor. Several approaches to significance can be envisaged, for instance, in the presence of fuzzy relations, the grades assigned to the index terms of the divisor could be diminished according to the attainment of a threshold.

These various types of tolerance are studied and formalized in the next two sections where tolerant inclusion operators are proposed.

4.2. *Dealing with exceptions*

As mentioned previously, an approach to softening the inclusion can be based on some tolerance to exceptions. This can be understood in two ways depending on the nature of the exceptions which can be either quantitative or qualitative.

4.2.1. *Quantitative exception-tolerant inclusion*

In the continuation of the first case evoked in Section 3, some tolerance may be introduced on the basis of a number of exceptions with respect to the universal quantifier present in formula 4. The definition of a quantitative exception-tolerant inclusion $E \subseteq_{\text{tol}} F$ is based on the "ignoration" of the fact that some elements of E are not sufficiently (or even not at all) in F. In other words, a certain number of keywords of the query can be more or less ignored depending on the desired level of relaxation. The principle adopted is to weaken the universal quantifier into a relaxed quantifier "almost all" (see (Kerre *et al.*, 1998), (Zadeh, 1983) for fuzzy relative quantifiers). In the following, the quantifier "almost all" induces grades defined as follows:

$\mu_{almost\ all}(0) = w_n = 0$, $\mu_{almost\ all}(1) = w_0 = 1$,

$\mu_{almost\ all}(1 - i/n) = w_i$ expresses the degree of satisfaction when i out of the n elements of the query are ignored.

By definition: $1 = w_0 \geq w_1 \geq ... \geq w_n = 0$

and if we denote: $k_1 = \max \{j \mid w_j = 1\}$, $k_2 = \max \{j \mid w_j > 0\}$,

the quantifier allows for the total "ignoration" of k_1 exceptions and the partial ignoration of up to k_2 exceptions. Basically, the idea is to search for the best compromise (value k) such that k elements of the F are in E and k is compatible with almost all, which leads to the following definition of the quantitative exception-tolerant inclusion: (Bosc *et al.*, 2006b)

$$\forall d \in D, g(E \subseteq_{quant\text{-}exc} F) = \max_{k \in [1, n]} \min(\alpha_k, w_{k-1}) \qquad (7)$$

where α_k is the k^{th} smallest implication degree ($\mu_E(x) \rightarrow \mu_F(x)$) and w_k is the degree of ignoration $w_k = \mu_{almost\ all}(1 - k/n)$, issued from the quantifier "almost all" for set E of cardinality equal to n, i.e., n implication values intervene in formula 10. It is worth noticing that: i) formula 4 is recovered by formula 10 when the universal quantifier is used ($w_0 = 1$, $w_1 = \ldots = w_n = 0$), i.e., when no exception is admitted, ii) as expected the result obtained is a superset of that returned by the non-tolerant division, and iii) formula (10) rewrites:

$$\forall d \in D, g(E \subseteq_{quant\text{-}exc} F) = \min_{i \in [1, n]} \max(\alpha_i, w_i) \qquad (8)$$

which establishes a clear connection with formula 4. Indeed, any implication value insufficiently satisfactory (from the smallest one to the largest one) is possibly replaced by the degree of satisfaction corresponding to the number of values ignored so far (according to "almost all"). When w_i is 1, total ignoration takes place, whereas if w_i is 0, the associated element α_i is completely taken into account as such. It appears that grades of ignoration define degrees of guaranteed satisfaction, i.e., if p implication values are ignored, the satisfaction level is at least w_p.

A quantitative exception tolerant division is obtained by replacing in formula 3 $g(S \subseteq \Omega^{-1}(d))$ by $g(S \subseteq_{quant\text{-}exc} \Omega^{-1}(d))$. Obviously, the user can choose the fuzzy implication to be used in formulas 10 and 11 so as to specify the role played by the degrees of the divisor (threshold or importance). Furthermore, if the quantifier Q_1 is included in Q_2 (in Zadeh's sense), the result of the tolerant division founded on Q_1 is included in that of the tolerant division division based on Q_2 (Bosc *et al.*, 2006a).

Example 3. One considers the quantifier "almost all" defined as:

$\mu_{almost\ all}(f) = 0$ if $f \in [0, 0.75]$,
$\mu_{almost\ all}(f) = 1$ if $f \in [0.95, 1]$,
$\mu_{almost\ all}(f)$ linearly increasing if $f \in [0.75, 0.95]$.

In this perspective, the degrees issued from the quantifier are $w_0 = 1$, $w_1 = 0.75$, $w_2 = 0.25$, $w_3 = \ldots = w_{10} = 0$ if s contains 10 elements. Using Gödel's implication and the following extensions of R and S:

$R = \{0.1/<d_1, t_2>, 0.2/<d_1, t_3>, 0.5/<d_1, t_4>, 0.7/<d_1, t_5>, 0.9/<d_1, t_6>, 1/<d_1, t_7>, 1/<d_1, t_8>, 0.2/<d_1, t_9>, 0.5/<d_1, t_{10}>, 0.8/<d_2, t_1>, 1/<d_2, t_3>, \ldots, 1/<d_2, t_{10}>\}$,

$S = \{1/t_1, 0.9/t_2, 0.9/t_3, 0.9/t_4, 0.9/t_5, 0.8/t_6, 0.7/t_7, 0.4/t_8, 0.2/t_9, 0.1/t_{10}\}$,

the result of the non-tolerant division is empty since $<d_1, t_1>$ and $<d_2, t_2>$ are missing in R. On the other hand, when the tolerant division is performed, the grade obtained by d_1 is:

min(max(0, 0.75), max(0.1, 0.25), max(0.2, 0), max(0.5, 0), max(0.7, 0), max(1, 0), …, max(1, 0)) =
max(min(0, 1), min(0.1, 0.75), min(0.2, 0.25), min(0.5, 0), min(0.7, 0), min(1, 0), …, min(1, 0)) = 0.2

and for d_2 :

min(max(0, 0.75), max(0.8, 0.25), max(1, 0), … , max(1, 0)) =
max(min(0, 1), min(0.8, 0.75), min(1, 0.25), min(1, 0), … , min(1, 0)) = 0.75.

For d_1, the two implication values 0 and 0.1 are ignored thanks to w_1 and w_2, while for d_2, only the first implication value 0 is ignored (thanks to w_1). ♦

4.2.2. *Qualitative exception-tolerant inclusion*

In Subsection 4.2.1, exceptions have been dealt with in a quantitative way. In this context, the quantitative inclusion of E in F expresses that "*almost all* elements of E are included in F according to the chosen implication". An alternative approach is to take a qualitative view and to consider a qualitative inclusion operator expressing "all elements of E are *almost* included in F according to the chosen implication". Then, exceptions are also taken into account according to the idea of "almost inclusion", which leads to a qualitative view. In other words, the idea is to (more or less) compensate the initial value of the implication when it expresses a sufficiently "low-intensity" exception (Bosc *et al.*, 2007).

Intuitively, the idea is to consider exceptions with respect to the inclusion in the following sense: if one looks for the inclusion of E in F, compensation takes place for an element x such that $\mu_E(x)$ and $\mu_F(x)$ are sufficiently close to the situation of (full) inclusion. It seems reasonable to consider that the closeness in this situation is a matter of degree rather than based on a crisp boundary. For instance one may think that in reference to a, $a \pm 0.1$ is totally acceptable, a shift beyond 0.3 cannot be tolerated and the satisfaction is linear in-between. Of course, this does not make sense for regular relations, for which exceptions correspond to the case where $\mu_E(x)$ equals 1 and $\mu_F(x)$ is zero, thus to a "full" exception.

If an S-implication is considered, full satisfaction, i.e. degree 1 is obtained when the antecedent is close to 0 and/or the conclusion is close to 1. As a consequence, "low-intensity" exceptions occur only for (then the compensation mechanism applies to) fairly high values of the initial implication degrees. From this, it appears that such a qualitative tolerant mechanism is not useful for reducing the risk of obtaining empty answers (even if it is of interest for the design of a tolerant division as such) and we move to the case of R-implications.

Let us first recall that any R-implication is completely satisfied if the conclusion (let us denote it by $\mu_F(x)$) attains the antecedent (let us denote it by $\mu_E(x)$). Consequently, the intensity of an exception depends on the difference between $\mu_E(x)$ and $\mu_F(x)$. Intuitively, if this difference is positive but small enough, we are in the presence of a "low-intensity" exception, which is somewhat tolerable. In this context, the compensation mechanism may affect any implication value (a large one, i.e., close to 1, as well as a small one), which makes it convenient for dealing with empty answers. The definition of the qualitative exception-tolerant inclusion is:

$$\forall d \in D,\ \mathrm{g}(E \subseteq_{qual\text{-}exc} F) = \min_{x \in E} \mu_E(x) \rightarrow_{\text{R-i}} (\mu_F(x) + \delta) \qquad (9)$$

where δ = 0 if $\mu_E(x) - \mu_F(x) \geq \beta$,
$\mu_E(x) - \mu_F(x)$ if $\mu_E(x) - \mu_F(x) \leq \alpha$,
linear in-between.

A qualitative exception tolerant division is obtained by replacing $g(S \subseteq \Omega^{-1}(d))$ by $g(S \subseteq_{qual\text{-}exc} \Omega^{-1}(d))$ in formula 3.

Example 4. Let us consider the document collection D represented by the binary relation $R = \{0.7/<d_1, t_1>, 0.4/<d_1, t_3>\}$ and the query $S = \{1/t_1, 0.1/t_2, 0.6/t_3\}$.

The usual division based on Gödel's implication delivers an empty answer due to the absence of t_2 in d_1. If the qualitative exception-tolerant division is performed with α = 0.1, β = 0.3, the result returned is:

$res = \{\min(1 \rightarrow_{\text{Gö}} 0.7, 0.1 \rightarrow_{\text{Gö}} 0 + 0.1, 0.6 \rightarrow_{\text{Gö}} 0.4 + 0.05)/<d_1>\} = \{0.45/<d_1>\}$

which is no longer empty.♦

4.3. *Modifiying the operands of the inclusion*

In this section, tolerant inclusions are designed on the basis of some semantic transformation of the operands. In order to get a relaxation of the original inclusion, the tolerant one may either enlarge the right-hand side argument:

$$tol\text{-}inc_1(E, F) \equiv E \subseteq F' \qquad (10)$$

where F' is a superset of F, or diminish the left-hand side argument:

$$tol\text{-}div_2(E, F) \equiv E' \subseteq F \qquad (11)$$

where E' is a subset of E, or both. In the next two subsections, the principle guiding these two types of modifications of the operands of an inclusion is discussed.

4.3.1. *Synonymy-based tolerant inclusion*

The rationale to enlarge the right-hand side argument is to compose it with a synonymy relation expressing that some keywords may be somewhat close semantically speaking. In the Boolean context, this is usually done thanks to an *equivalence* relation (reflexive, symmetric and transitive) – represented by a thesaurus (Salton *et al.* 1984) –, which can be extended to a *similarity* relation (where transitivity is replaced by: $\forall x, y, z\ sim(x, y) \geq \sup_z \min(sim(x, z), sim(z, y))$) in the fuzzy framework. Other variants can be used where the transitivity is skipped (*proximity* relation). Whatever the type of *resemblance* relation used (denoted later by *rsb*), the right-hand argument (let us denote it by F) is *dilated* in the sense of the adjunction of any element of the referential which is close to an element initially present in it. The idea is to consider that an element x missing in F can be replaced by another x' which is present and such that x' is a synonym (to some extent) of x.

The resemblance-based tolerant inclusion is defined as follows:

$$g(E \subseteq_{rsb\text{-}tol} \mathrm{F}) = \min_{x \in S} \mu_E(x) \rightarrow \mu_{dil(F)}(x) \qquad (12)$$

which is a refinement of expression 10 and *dil(F)* is a dilated variant of *F* obtained using the resemblance relation *rsb*, i.e.:

$$\mu_{dil(F)}(x) = \sup_{x' \in U} \top(\ \mu_F(x'), \mu_{rsb}(x, x')) \qquad (13)$$

where $\top$ is a triangular norm. It is worth noticing that if a value initially absent can be added to *F* (with a given degree of membership), the degree of a value initially present can also be increased. The less demanding the norm used in formula 13, the larger the final dilated set obtained and then the greater the result of the approximate inclusion. A resemblance-tolerant division is obtained by replacing $g(S \subseteq \Omega^{-1}(d))$ by $g(S \subseteq_{rsb\text{-}tol} \Omega^{-1}(d))$ in formula 3. Let us mention that the approach proposed here can be seen as a generalization of that described in (Miyamoto *et al.*, 1986) in the sense that this author only considers the t-norm minimum for the dilation process (formula 13), but it also differs from that approach in the sense that we use a matching process based on a graded inclusion, while in (Miyamoto *et al.*, 1986) the interaction between the query weights and the document weights rests on a minimum.

Example 5. Let us consider the document collection *D* represented by the following binary relation:

R = {0.3/<d_1, *grand prix*>, 0.6/<d_1, *speedcar*>, 0.4/<d_1, *automobile*>,
1/<d_2, *race*>, 0.7/<d_2, *formula 1*>},

the query *S* = {1/<*grand prix*>, 0.5/<*formula 1*>}

and the following graded synonymy relation (where the pairs 1/(x, x) are omitted and α/(y, x) is implicit when α/(x, y) is present):

rsb = {0.7/(*grand prix*, *race*), 0.6/(*automobile*, *speedcar*),
0.5/(*automobile*, *formula 1*), 0.9/(*speedcar*, *formula 1*)}

The regular division of *R* by *S* using Goguen's implication delivers an empty result. The dilation of *R* with the norm minimum (the largest one) yields:

dil(R) = {0.3/<d_1, *grand prix*>, 0.3/<d_1, *race*>, 0.6/<d_1, *speedcar*>,
0.6/<d_1, *automobile*>, 0.6/<d_1, *formula 1*>, 1/<d_2, *race*>,
0.7/<d_2, *grand prix*>, 0.7/<d_2, *formula 1*>, 0.7/<d_2, *speedcar*>,
0.5/<d_2, *automobile*>},

and the resemblance-tolerant division of *R* by *S* with Goguen's implication in formula 12 returns the relation *res* = {0.3/<d_1>, 0.7/<d_2>}).♦

4.3.2. *Significance-based tolerant inclusion*

In the preceding section, one of the arguments of the inclusion, namely the one on the right-hand side, is extended in order to have a better chance to get a non-empty answer to the division. A dual point of view consists in reducing (eroding) the left-hand side argument and the key question is about the rationale of the *erosion* mechanism. In the information retrieval context, the general idea is to reduce the query considering that weakly significant keywords can more or less be removed. It turns out that different examples of the notion of significance can be pointed out depending on the nature (fuzzy or not) of the query into play.

If S is a fuzzy query (i.e., a fuzzy set of keywords), it can be eroded by removing the elements with a small grade (recall that such elements may cause an empty answer with implications like Gödel's or Goguen's).

Example 6. Let us consider the document collection D represented by the binary relation $R = \{0.7/\langle d_1, t_1\rangle, 0.4/\langle d_1, t_3\rangle\}$ and the query $S = \{1/t_1, 0.35/t_2, 0.6/t_3\}$.

The usual division of R by S based on Gödel's (or Goguen's) implication in formula (4) delivers an empty answer due to the absence of t_2 in d_1. If elements of the query with a degree less than 0.4 are considered non-significant and removed according to the erosion mechanism used in the significance-based tolerant division, the result obtained is: $res = \{\min(1 \rightarrow_{\text{Gö}} 0.7, 0.6 \rightarrow_{\text{Gö}} 0.4)/\langle d_1\rangle\} = \{0.4/\langle d_1\rangle\}$.♦

On the other hand, with a non-weighted query, the notion of significance is not as straightforward. A simple strategy could be based on the fact that, most often, the order of the keywords in the query reflects the relative importance that the user attaches to these keywords. Then, one could eliminate the last one, or the last two ones if necessary, and so on. Another idea could be to use an ontology in order to construct some clusters of keywords from the initial query. The keywords inside a cluster would satisfy a given relationship with respect to the ontology, and the keywords which do not belong to any cluster would be discarded from the query. This issue is left for future works since it is not obvious what type of relationship could be used as a good basis for such a clustering process.

Since regular sets are just a special case of fuzzy ones, all these cases can be modelled in the fuzzy framework as follows. The significance-based tolerant inclusion of E by F is defined as follows:

$$g(E \subseteq_{sgf\text{-}tol} F) = \min_{x \in U} \mu_{ero(E)}(x) \rightarrow \mu_F(x) \quad (14)$$

which is a refinement of expression 11 and *ero(E)* is an eroded variant of E obtained using one of the previously evoked mechanisms. A significance-based tolerant division is obtained by replacing $g(S \subseteq \Omega^{-1}(d))$ by $g(S \subseteq_{sgf\text{-}tol} \Omega^{-1}(d))$ in formula 3.

5. Conclusion

In this paper, a fuzzy model of information retrieval has been considered and an interpretation of the document-query matching process in terms of a division of fuzzy relations (which relies on the notion of inclusion) has been presented. This works stems from the fact that in the Boolean IR model, the concept of inclusion plays a central role in retrieval. We have pointed out different ways of extending the inclusion in order to make it more tolerant, which leads to different tolerant matching mechanisms and reduces the risk of obtaining empty answers to an IR query. This paper focused on semantic aspects and, as a perspective for future work, it would now be worthy to implement a search engine based on the principles described here in order to assess its performances in terms of precision and recall (for example by means of experiments on the TREC or INEX collections) and compare them with those obtained by classical systems such as Smart, Okapi or Mercure.

6. Bibliography

Bookstein A., « Fuzzy requests: an approach to weighted Boolean searches », *J. of the American Society for Information Science*, vol. 31, 1980, p. 240-247.

Bordogna G., Carrara P., Pasi G., « Query term weights as constraints in fuzzy information retrieval », *Information Processing and Management*, vol. 27, 1991, p. 15-26.

Bosc P., Dubois D., Pivert O., Prade H., « Flexible queries in relational databases – The example of the division operator », *Theoretical Computer Science*, vol. 171, 1997, p. 281-302.

Bosc P., Hadjali A., Pivert O., « Preference-based divisions to overcome empty answers », *Proc. of the 3rd Multidisciplinary Workshop on Advances on Preference Handling (M-PREF'07)*, in conjunction with *VLDB'07*, Vienna, Austria, September 24, 2007.

Bosc P., Pivert O., « About approximate inclusion and its axiomatization », *Fuzzy Sets and Systems*, vol. 157, 2006a, p. 1438-1454.

Bosc P., Pivert O., Rocacher R., « A propos de division usuelle et approchée de relations floues », *Technique et Science Informatiques*, vol. 25, 2006b, p. 631-660.

Buell D.A., « An analysis of some fuzzy subset applications to information retrieval systems », *Fuzzy Sets and Systems*, vol. 7, 1982, p. 35-42.

Buell D.A., Kraft D.H., « Threshold values and Boolean retrieval systems », *Information Processing & Management*, vol. 17, 1981, p. 127-136.

Fodor J.,Yager R.R., « Fuzzy-set theoretic operators and quantifiers », in: *Fundamentals of Fuzzy Sets – The Handbook of Fuzzy Sets Series*, D. Dubois and H. Prade (Eds.), Kluwer Academic Publishers, 1999, p. 125-193.

Kerre E.E., Liu Y., « An overview of fuzzy quantifiers – Interpretations », *Fuzzy Sets and Systems,* vol. 95, 1998, p.1-22.

Miyamoto S., Nakayama K., « Fuzzy information retrieval based on a fuzzy pseudo-thesaurus », *IEEE Transactions on Systems, Man and Cybernetics*, vol. 16, 1986, p. 278-282.

Radecki T., « Fuzzy set theoretical approach to document retrieval », *Information Processing and Management*, vol. 15, 1979, p. 247-260.

Salton G., McGill M.J., *Introduction to modern information retrieval*, McGraw-Hill Int. Book Co., 1984.

Waller W.G., Kraft D.H., « A mathematical model of a weighted Boolean retrieval system », *Information Processing & Management*, vol. 15, 1979, p. 235-245.

Yager R.R., « A note on weighted queries in information retrieval systems », *J. of the American Society for Information Science*, vol. 38, 1987, p. 23-24.

Zadeh L.A., « A computational approach to fuzzy quantifiers in natural languages », *Computer Mathematics with Applications*, vol. 9, 1983, p. 149-183.