A NEW SEMANTICS FOR THE DIVISION OF FUZZY RELATIONS IN RELATIONAL DATABASES

Patrick Bosc
IRISA/ENSSAT
Technopole Anticipa BP 447
22305 Lannion Cedex France
bosc@enssat.fr

Olivier Pivert
IRISA/ENSSAT
Technopole Anticipa BP 447
22305 Lannion Cedex France
pivert@enssat.fr

Summary

Two principal semantics have been assigned to the division of fuzzy relations. In the first one, the weights of the divisor are seen as thresholds and in the other, they act as grades of importance. This paper discusses a third approach in which the weights play the role of ideal values. In this context, the notion of a graded inclusion between (fuzzy) sets moves to that of a graded equality. It is shown that the desired behavior can be reached using an appropriate fuzzy implication.

Keywords: Database systems, Fuzzy relations, Division, Graded set inclusion, Graded set equality, Fuzzy implications.

1 INTRODUCTION

In several retrieval applications, the user need is to search for elements of the system which are connected (in a certain sense) with a given set of items. In a relational setting, this need can be expressed thanks to a particular operator, known as the division of relations. In the context of a database that contains the relations PARTS(#part, part-name, price) and SALES(#store, #part, quantity), a typical query calling on a division is: "retrieve the stores that have sold (at least) all the parts priced under $40 in a quantity over 50". The result is the set of stores (#store) that are connected (in SALES) with at least all parts whose price is less than $40 with a quantity greater than 50.

It has been shown that the search activity undertaken in information retrieval where users are looking for those documents indexed by a given set of keywords pertains to the same category. This is also the case in the engineering system described in [4] where the user (a software developer) is looking for software components having a given set of characteristics. In all these cases, the set inclusion is the key tool to model the problem in which a set of desired features plays the role of a model. Then, it seems relevant to consider situations where weights are introduced in order to refine the notions of presence/absence of features. For instance, in the information retrieval framework such weights can be associated with the keywords. In doing so, any document is indexed with a set of terms (acting as topics), each of which has a weight reflecting the intensity of the topic in the document. Similarly, a user can put weights in his query. Such an approach is suggested in [1] where fuzzy sets are used and weights take their values in the unit interval.

More generally, when fuzzy set theory is used to model weights, a possible approach is to turn the classical division (and the underlying inclusion) into a graded one whose arguments are two fuzzy relations. This view is advocated in [2, 3, 4] and it turns out that different semantics are reached depending on the nature of the inclusion taken into account (implication or cardinality-based) and the role played by the weights attached to the divisor relation (importance or thresholds). In this paper, the idea is to investigate a new role for these weights when they act as "ideal values".

In section 2, the notion of a graded inclusion between fuzzy sets and that of the division of fuzzy relations are briefly recalled. Section 3 is devoted to the introduction of a new meaning of grades attached to elements appearing in the set acting as a model. It is pointed out that this view is connected with the idea of a rejection set, which, indeed, could also be of interest to enrich the possibilities offered when grades are importances or thresholds. The formalization of this approach is discussed in section 4. The need for a new set comparison operator is first shown and its definition in terms of a (double) implication is suggested. The conclusion summarizes the main points and contributions of the paper and opens some tracks for future research.

2 PREVIOUS SEMANTICS OF THE DIVISION OF FUZZY RELATIONS

As mentioned before, the division of two regular relations R(A, X) by S(A) where A and X represent sets of attributes whose underlying domains are finite, denoted by $div_{A \div A}(R, S)$, aims at determining the X-values connected in R with (at least) all the A-values appearing in S.

The extension of the relational division consists in the definition of this operation when fuzzy relations are involved. An example is the query "find the stores which have sold a moderate number of all medium-priced parts". Formally, the usual division operation is defined as:

$$x \in div_{A \div A}(R, S) \Leftrightarrow S \subseteq \Gamma^{-1}(x) \qquad (1)$$

where $\Gamma^{-1}(x) = \{a \mid (x, a) \in R\}$, and a natural extension, as suggested in [5] and [3], stems from formula 1, where the usual set inclusion is changed into a gradual inclusion of the fuzzy relation S in $\Gamma^{-1}(x)$(the fuzzy set of A-values connected with x in R). In this context, inclusion degrees based on fuzzy implications or fuzzy cardinalities may be considered and if $\Rightarrow_f$ (resp. *) denotes a fuzzy implication (resp. a triangular norm), the following generic definitions of the division of two fuzzy relations are obtained:

$$\mu_{div_{A \div A}(R, S)}(x) = min_{Domain(A)} (S(a) \Rightarrow_f R(x, a)) \qquad (2)$$

$$\mu_{div_{A \div A}(R, S)}(x) = (\Sigma_{Domain(A)} (S(a) * R(x, a))) / \Sigma_{Domain(A)} S(a) \qquad (3).$$

To get an interpretation (from a user point of view) of this operation (and thus of its result), the key point lies in a proper choice of the implication (resp. the triangular norm). If a Gödel-like implication ($a \Rightarrow_G b = 1$ if $a \leq b$, b (Gödel) or b/a (Goguen) or 1 - (a - b) (Lukasiewicz) otherwise) is chosen in formula 2, the contribution of the tuple <a, x> is 1 provided R(a, x) is over S(a) which means that S(a) is seen as a threshold to be attained. If Dienes implication ($a \Rightarrow_D b = max(1 - a, b)$) is used, S(a) plays the role of a level of importance. The contribution of the tuple <a, x> cannot equal 0 when S(a) is not completely important (strictly less than 1). Similarly, in formula 3, S(a) plays a role of threshold if the triangular norm is min and that of a level of importance for the product. Consequently, the interpretation of the division depends strongly on the type of inclusion degree and the role assigned to the weights in S-tuples, which is strongly tied to the implication or the triangular norm.

Example 1. Let us consider the query: "find the stores which have sold a moderate number of all medium-priced parts" addressed to the relations PARTS and SALES whose schema is given before. With the extensions displayed in tables 1 and 2, the results depending on the interpretation chosen by the user are:

for s1: $min(1 \Rightarrow_{Gö} .8, .4 \Rightarrow_{Gö} .2, .6 \Rightarrow_{Gö} 1) = .2$;
$min(1 \Rightarrow_{Gg} .8, .4 \Rightarrow_{Gg} .2, .6 \Rightarrow_{Gg} 1) =.5$;
$min(1 \Rightarrow_{Lu} .8, .4 \Rightarrow_{Lu} .2, .6 \Rightarrow_{Lu} 1) = 8$;
$min(1 \Rightarrow_{D} .8, .4 \Rightarrow_{D} .2, .6 \Rightarrow_{D} 1) = .6$;
(min(1,. 8) + min(.4,. 2) + min(.6, 1)) /
(1 + .4 + .6) = .8;
(1 * .8) + (.4 * .2) + (.6 * 1) / (1 +.4 +.6) = .74;

for s2: $min(1 \Rightarrow_{Gö} .5, .4 \Rightarrow_{Gö} 0, .6 \Rightarrow_{Gö} 0) = 0$;
$min(1 \Rightarrow_{Gg} .5, .4 \Rightarrow_{Gg} 0, .6 \Rightarrow_{Gg} 0) = 0$;
$min(1 \Rightarrow_{Lu} .5, .4 \Rightarrow_{Lu} 0, .6 \Rightarrow_{Lu} 0) =.4$;
$min(1 \Rightarrow_{D} .5, .4 \Rightarrow_{D} 0, .6 \Rightarrow_{D} 0) = .4$;
(min(1, .5) + min(.4, 0) + min(.6, 0)) /
(1 + .4 + .6) = .25;
(1 * .5) + (.4 * 0) + (.6 * 0) / (1 +.4 +.6) = .25 ♦

Table 1: Sample extension for the relation SALES

SALES	#store	#part	quantity	moderate
	s1	p1	30	.8
	s1	p2	99	.2
	s1	p3	40	1
	s2	p1	85	.5

Table 2: Sample extension for the relation PARTS

PARTS	#part	price	medium
	p1	100	1
	p2	300	.4
	p3	250	.6

3 ANOTHER ROLE OF THE GRADES IN THE DIVISOR

If the forementioned roles assigned to grades of the divisor are useful in some fields such as information or software component retrieval, another need appears in these domains. More precisely, the user may be interested in grading the elements of the divisor with weights that do not have a semantics of thresholds, but that act as "ideal values" (objectives to be attained as exactly as possible).

Let us illustrate this with an example in the context of information retrieval. A user is supposed to look for documents having a given "profile", i.e., addressing topic t1 at level l1, topic t2 at level l2 and so on. As with the two other interpretations (threshold and importance), it is assumed that the user knows the indexing language and the way levels (valued in the unit interval) are assigned during the indexing stage. For instance, a user interested in developing a database application with Java will choose the profile {<"databases", 1>; <"application development", 0.6>; <"Java", 0.8>}. With a threshold semantics, any document where the keyword "databases" (resp. "application development", ."Java") is assigned a grade greater than (or equal to) 1 (resp. 0.6, 0.8) is totally satisfactory and this is not the desired behavior, since when the desired "ideal" value of the profile is exceeded, the satisfaction should not be complete.

The special case where the ideal value specified for a grade is 0 deserves some particular attention. A specified zero grade means that "ideally" the associated element should be absent from the dividend. On the other hand, the elements which were not specified by the user, i.e., those which do not explicitely appear in the dividend relation S, also belong to S to a same degree 0 and thus will be considered undesirable as well. As a matter of fact, there is no way, using a single fuzzy set, to express different meanings for the zero in order to distinguish between the notions of undesirable elements and indifferent ones. For that purpose, two sets have to be provided: a fuzzy set S^+ of desired elements whose attached grade is positive and a crisp set S^0 of elements that are undesirable. The elements of Domain(A) that do not belong to the support of $S^+ \cup S^0$ are considered "indifferent" and they should not affect the result. In doing so, an element of S^0 will be treated as a limit case of an element of S^+ when the grade moves from a "small" value to zero and the behavior of the system is continuous, i.e., there is not a big difference in the result for these two situations.

It is worth noticing that, with the two other semantics (threshold or importance), the situation is quite different: the elements which do not appear in the divisor are all considered indifferent (cf. formulas 2 and 3). However, let us remark that a set of rejected elements could also be incorporated when importance or threshold semantics is chosen. In those cases, this set could be graded in order to express the effect in terms of importance or threshold of the (more or less) high presence of an undesired element.

Clearly, the notion of an "ideal" value played by a grade calls on a sort of resemblance for the matching mechanism (whereas the threshold semantics is more a matter of attainment). The operation (denoted f) used for this purpose must satisfy (at least) the following two properties:

- $f(a, a) = 1 \ \forall a \in [0, 1]$,
- $f(a, a + \alpha) = f(a, a - \alpha) \ \forall a, (a + \alpha), (a - \alpha) \in [0, 1]$.

The second expression states that the satisfaction is the same when a same quantity (α) is exceeding or lacking the "ideal" value a.

4 IDEAL VALUES AND THE DIVISION

In this section, a formalization of the behavior described before is proposed. More precisely, the principle of the approach is to start with formula 1 and to extend it in the spirit of what has been done for thresholds and importances.

First, let us remark that formula 1 can be rewritten as:

$$x \in div_{A \div A}(R, S) \Leftrightarrow S = (S \cap \Gamma^{-1}(x)) \qquad (4)$$

where the intersection used in the right-hand side is aimed at the removal of the values of domain(A) that are "indifferent". From this, one can move to the case where the relations are fuzzy (graded) ones. Here, S is indeed made of two sets: S^+ the fuzzy set of (more or less) desired elements and S^0 the crisp set of undesirable elements. R being a fuzzy relation, $\Gamma^{-1}(x)$ becomes:

$$\mu_{\Gamma^{-1}(x)}(a) = \mu_R(x, a).$$

For the sake of clarity, it is better to give an expression of the division in the spirit of formula (4) where the two components of S are treated separately, which yields:

$$\mu_{div_{A \div A}(R, S)}(x) = min(\mu_{\approx}(support(S^+) \cap \Gamma^{-1}(x), S^+), \mu_{\approx}(S^0 \cap \Gamma^{-1}(x), \varnothing) \qquad (5)$$

where ≈ stands for a graded equality between fuzzy sets. The first part of the formula is a direct adaptation from (4) which deals with the desired elements (S^+) and the second part expresses that the ideal grade of any undesirable element (specified in S^0) appearing in the dividend (R) is 0.

A rather natural way for defining the graded equality between fuzzy sets appearing in formula (5) is to generalize the usual equality seen as a double inclusion:

$$\mu_{\approx}(E, F) = min(\mu_{gr\text{-}inc}(E, F), \mu_{gr\text{-}inc}(F, E))$$

where gr-inc (- , -) denotes a graded inclusion between (fuzzy) sets. Here, the two lines used in formulas 2 and 3 for the inclusion (implication or cardinality-based inclusion) can be investigated in the perspective of the satisfaction of the properties mentioned at the end of section 3 (in particular, the symmetric behavior in case of excess or default). Different definitions of the inclusion are now reviewed.

If the grade of inclusion (of E in F) taken into account is based on the ratio between the cardinality of ($E \cap F$) and E:

$$\mu_{gr\text{-}inc}(E, F) = (\Sigma_X \, \mu_{E \cap F}(x)) / \Sigma_X \, \mu_E(x)$$

it is easy to prove that, in general, the desired behavior is not reached as shown in example 2.

Example 2. Let us take three fuzzy sets E, F and G defined as follows:

E = {.1/a + .7/b}, F = {.1/a + .5/b}, G = {.1/a + .9/b}.

The expected behavior should lead to the validity of the equality: $\mu_{\approx}(E, F) = \mu_{\approx}(E, G)$. Let us compute these values from those of the various inclusion degrees involved, assuming the minimum for the intersection:

$\mu_{gr\text{-}inc}(E, F) = 0.75$, $\mu_{gr\text{-}inc}(F, E) = 1$, $\mu_{\approx}(E, F) = 0.75$,
$\mu_{gr\text{-}inc}(E, G) = 1$, $\mu_{gr\text{-}inc}(G, E) = 0.8$, $\mu_{\approx}(E, G) = 0.8$ ♦

Let us now look at (fuzzy) implication-based grades of inclusion:

$$\mu_{gr\text{-}inc}(E, F) = min_{x \in X} \, \mu_E(x) \Rightarrow_f \mu_F(x)$$

where $\Rightarrow_f$ is a fuzzy implication. Dienes ($a \Rightarrow_D b = max(1 - a, b)$), Gödel ($a \Rightarrow_{Gö} b = 1$ if $a \le b$, b otherwise) or Goguen ($a \Rightarrow_{Gg} b = 1$ if $a \le b$, b/a otherwise) have a result which is not a linear combination of the inputs and then, they are by nature unable to reflect the desired behavior as illustrated in example 3 hereafter.

Example 3. Let us take the same fuzzy sets E, F anf G as those of example 2. The computation of the equality degree between E and F on the one hand and E and G on the other yields:

- with Dienes implication:

$\mu_{gr\text{-}inc}(E, F) = 0.5$, $\mu_{gr\text{-}inc}(F, E) = 0.7$, $\mu_{\approx}(E, F) = 0.5$,
$\mu_{gr\text{-}inc}(E, G) = 0.9$, $\mu_{gr\text{-}inc}(G, E) = 0.7$, $\mu_{\approx}(E, G) = 0.7$

- with Gödel implication:

$\mu_{gr\text{-}inc}(E, F) = 0.5$, $\mu_{gr\text{-}inc}(F, E) = 1$, $\mu_{\approx}(E, F) = 0.5$,
$\mu_{gr\text{-}inc}(E, G) = 1$, $\mu_{gr\text{-}inc}(G, E) = 0.7$, $\mu_{\approx}(E, G) = 0.7$

- with Goguen implication:

$\mu_{gr\text{-}inc}(E, F) = 5/7$, $\mu_{gr\text{-}inc}(F, E) = 1$, $\mu_{\approx}(E, G) = 5/7$,
$\mu_{gr\text{-}inc}(E, G) = 1$, $\mu_{gr\text{-}inc}(G, E) = 7/9$, $\mu_{\approx}(E, G) = 7/9$ ♦

Let us now move to the case of Lukasiewicz implication ($a \Rightarrow_L b = 1$ if $a \leq b$, $1 - (a - b)$ otherwise) and consider the key situation where an element x has a grade a in E and $(a - \alpha)$ in F. For this element x, one has:

$$\mu_E(x) \Rightarrow_L \mu_F(x) = 1 - \alpha \text{ and } \mu_F(x) \Rightarrow_L \mu_E(x) = 1.$$

The "symmetric" situation would be when the grade of x is a in E and $(a + \alpha)$ in F, for which one gets:

$$\mu_E(x) \Rightarrow_L \mu_F(x) = 1 \text{ and } \mu_F(x) \Rightarrow_L \mu_E(x) = 1 - \alpha.$$

We observe that in both cases the "local" result tied to x is the same (namely $(1 - \alpha)$), which is the expected one.

Finally, we propose to define the division of fuzzy relations where the grades of the divisor act as "ideal" values (i.e, to be attained as closely as possible) according to formula (5) where the fuzzy equality between (fuzzy) sets is defined as:

$$\mu_{\approx}(E, F) = min(min_{x \in X}\ \mu_E(x) \Rightarrow_L \mu_F(x),\ min_{x \in X}\ \mu_F(x) \Rightarrow_L \mu_E(x)).$$

Example 4. Let us illustrate the result obtained in the context of the retrieval of documents (see section 3) with:

S^+ = {1/database, 0.7/application development, 0.8/Java}
S^0 = {C, C++}

meaning that the user is interested in documents dealing with database application development in Java, but not with C or C++. If the documents:

d1 = {0.8/database, 1/application development, 1/Java, 0.4/Pascal, 0.2/C}
d2 = {1/database, 0.4/application development, 0.7/Java, 0.6/C, 0.4/C++}

are in the archive, their matching degree with respect to the query (S^+ and S^0) are:

- for d1: min($\mu_{\approx}$({0.8/database, 1/application development, 1/Java}, {1/database, 0.7/application development, 0.8/Java}), $\mu_{\approx}$({0.2/C}, ∅)) = min(0.7, 0.8) = 0.7,

- for d2: min($\mu_{\approx}$({1/database, 0.4/application development, 0.7/Java}), {1/database, 0.7/application development, 0.8/Java}), $\mu_{\approx}$({0.6/C, 0.4/C++}, ∅)) = min(0.7, 0.4) = 0.4 ♦

The approach adopted here with Lukasiewicz implication has the advantage of keeping close to the general framework suggested for extending the division to fuzzy relations, which makes use of fuzzy implications. The counterpart is the rigidity of the solution due to the general shape of the comparison which, for each element x of the divisor (including S^+ and S^0), computes:

$$min(\mu_{divisor}(x) \Rightarrow_L \mu_{dividend}(x),\ \mu_{dividend}(x) \Rightarrow_L \mu_{divisor}(x)) = 1 - abs(\mu_{divisor}(x) - \mu_{dividend}(x)).$$

Then, it could then be of interest to investigate the extent to which such a function can be parameterized while keeping the connection with a double inclusion relying on an implication(-like) operator. In doing so, one could expect in particular a maximal satisfaction when grades are close but not necessarily identical and/or a minimal satisfaction for pairs of grades differing from less than 1.

5 CONCLUSION

In this paper, a new semantics is proposed for the division of fuzzy relations. This approach is complementary to the two other interpretations suggested in previous papers. The idea is to offer a unique framework in which the three possible interpretations take place and it turns out that graded inclusion and equality between fuzzy sets are central tools. The new semantics, in which the grades of the divisor are seen as ideal values to attain as closely as possible, can be modelled using Lukasiewicz implication which offers appropriate properties. The extension of this solution in order to reach a richer range of attitudes in the framework of implication-like operators is a matter for future work.

References

[1] G. Bordogna, P. Carrara and G. Pasi (1991). Query term weights as constraints in fuzzy information retrieval. *Information Processing & Management*, 27, pp 15-26.

[2] P. Bosc, G. Bordogna and G. Pasi (1996). Fuzzy inclusion in database and information retrieval query interpretation. *Proc. 6th International Conference on Information Processing and Management of Uncertainty (IPMU'96)*, Granada (Spain), July 1996, pp 1171-1176.

[3] P. Bosc, O. Pivert and H. Prade (1997). Flexible queries in relational databases - The example of the division operator. *Theoretical Computer Science*, 171, pp 281-302.

[4] P. Bosc, E. Damiani and M-G. Fugini (1998). Dynamic service identification in a CORBA-like environment. *Proc. 1st International Workshop on Innovative Internet Information Systems (IIIS'98)*, Pisa (Italy), pp 161-175.

[5] D. Dubois and H. Prade (1996). Semantics of quotient operators in fuzzy relational databases, *Fuzzy Sets and Systems*, 78, pp 89-93.