A note on fuzzy queries involving a global evaluation of a set of values satisfying a fuzzy property

Henri Prade
Lab. Langages et Systèmes Informatiques
Université Paul Sabatier, 118 route de Narbonne
31062 Toulouse Cedex - France

Let us consider a relation R in a relational database, involving two attributes A and B, as pictured in Table 1

R	...	A	...	B	...
	...	a_1	...	b_1	...
	...	a_2	...	b_2	...
	⋮	⋮	⋮	⋮	⋮
	...	a_n	...	b_n	...

Table 1

The attribute values a_i and b_j are supposed to be precisely known, i.e. they belong to the attribute domains of A and B respectively. The queries we consider in this note are of the form "What is the global evaluation f of the a_i's such that the corresponding b_i's satisfy the fuzzy property B". The global evaluations f we are more particularly interested in here are the average, the maximum of the minimum of the a_i's. Such queries when B is a crisp property can be easily handled by SEQUEL-like languages and thus it is desirable to treat them when B is fuzzy if we want to extend these query languages to all kinds of fuzzy/vague queries (see Hamon [5] for instance). Examples of such queries are "What is the minimum of the salaries of middle-aged people in the database ?" or even "What is the average of the high salaries of people in the database ?" (in this latter case A = B).

Let B be a fuzzy set defined on the domain of attribute $\mathcal{B}$. The b_i's are assumed to be reordered according to the decreasing values of $\mu_B(b_i)$, i.e.

$$\mu_B(b_1) \geq \mu_B(b_2) \geq \ldots \geq \mu_B(b_n) \quad (1)$$

Let B_α be the α-cut of B defined by $\forall \alpha \in (0,1]$, $B_\alpha = \{b_i, \mu_B(b_i) \geq \alpha\}$. Note that, due to (1), $B_\alpha = \{b_1,\ldots,b_k\}$ where k is such that $\mu_B(b_k) \geq \alpha$ and $\mu_B(b_{k+1}) < \alpha$ (we assume $\mu_B(b_{n+1}) = 0$ by convention).

Let $A(\alpha)$ be the set of values $\{a_1,\ldots,a_k\}$ corresponding to $B_\alpha = \{b_1,\ldots,b_k\}$, and $f[A(\alpha)] = f(a_1,\ldots,a_k)$. Then the fuzzy set N of the values of f applied to the a_i's whose corresponding b_i's are (more or less) in B, is given by

$$\mu_N(r) = \sup\{\alpha | f[A(\alpha)] = r\} \quad (2)$$

Note that $\mu_N(r) \neq 0$ only if $\exists\, \alpha \in (0,1]$, $f[a(\alpha)] = r$. For instance if $n = 5$, $\mu_B(b_1) = 1 = \mu_B(b_2)$, $\mu_B(b_3) = 0.8$; $\mu_B(b_4) = 0.5 = \mu_B(b_5)$, (2) gives $N = 1/f(a_1,a_2) + 0.8/f(a_1,a_2,a_3) + 0.5/f(a_1,a_2,a_3,a_4,a_5)$ where the grade of membership is before the '/' and '+' denotes the union of singletons. The fuzzy set N is not always normalized, i.e. when $\mu_B(b_1) < 1$, $\nexists r$, $\mu_N(r) = 1$; this corresponds to the fact that there is no b_i which completely belong to B. The meaning of (2) is clear ; depending on the membership threshold we consider, there are more or less a_i's which are taken into account in the evaluation by f. In the expression (2), it is assumed that the α-cuts of B are the only possible crisp representatives of the fuzzy set B ; <u>all</u> the elements with a membership degree greater or equal to α must be considered in any crisp view of B of level α. It is why quantities like $f(a_1,a_2,a_3,a_5)$ or $f(a_1,a_2,a_3,a_4)$ for instance, do not appear in the above example.

<u>N.B.1</u>. However as pointed out in [3], it would be possible to have a slightly different understanding of the fuzzy set B : the crisp set S is a representative of B if and only if $B_1 \subseteq S \subseteq s(B)$ (where $s(B) = \{b, \mu_B(b) > 0\}$) ; then the suitability of S for representing B is computed as $\inf\{\mu_B(b), b \in S\}$. In this view, the set of crisp representatives includes and is larger than the set of α-cuts. □

N.B.2. The expression (2) is quite similar to the first definition of the fuzzy cardinality of a finite fuzzy set proposed by Zadeh (see [3] and [9] for discussions) ; this definition is recovered for $a_i = 1$, $\forall$ i and $f = \Sigma$.□

Thus the fuzziness of B induces a fuzzy set of possible answers $\mu_N(r)/r$ for the query, instead of one value when B is crisp. It would be desirable to summarize this information in a more concise, but still significant, way. It seems that this can be done at least in two different kinds of way.

A first -quite intuitive- technique is to use the weighted mean

$$w(N) = \frac{\sum_r \mu_N(r).r}{\sum_r \mu_N(r)} \qquad (3)$$

A slightly different expression which might be also considered is

$$w'(N) = \frac{\sum_i f(a_1,\ldots,a_i).\mu_B(b_i)}{\sum_i \mu_B(b_i)} \qquad (3')$$

A second, perhaps more subtle, technique is to compute the lower and/or the upper expected value attached to N. Let $\mu_N(r_j)$ be abbreviated by μ_j for $j = 1,q$ (μ_N is non-zero only for a finite number of r_j's). Then the lower expectation $E_*(N)$ and the upper expectation $E^*(N)$ are respectively defined by

$$E_*(N) = \sum_{j=1}^{q} r_j.(\max_{k\leq j} \mu_k - \max_{k<j} \mu_k) \qquad (4)$$

$$E^*(N) = \sum_{j=1}^{q} r_j.(\max_{k\geq j} \mu_k - \max_{k>j} \mu_k) \qquad (5)$$

where the r_j's are ordered increasingly, i.e.

$$r_1 \leq r_2 \leq \ldots \leq r_q \tag{6}$$

The reader is referred to [3] and [4] for rationales about these quantities. It can be proved for instance that the upper expectation of the fuzzy cardinality of a fuzzy set (when suitably defined) is nothing but its scalar cardinality while the lower expectation is the cardinality of the 1-cut. (See [3] and [6]).

When the μ_k's are decreasing, i.e.

$$\mu_1 \geq \mu_2 \geq \ldots \geq \mu_q \tag{7}$$

the expressions (4) and (5) can be simplified into

$$\begin{cases} E_*(N) = r_1 \quad \text{if } \mu_1 = 1 \\ E^*(N) = \sum_{j=1}^{q} r_j.(\mu_j - \mu_{j+1}) \\ \qquad = r_1 + \sum_{2}^{q} \mu_j.(r_j - r_{j-1}) \quad \text{if } \mu_1 = 1 \end{cases} \tag{8}$$

with $\mu_{q+1} = 0$ by convention. When the μ_k's are increasing, i.e.

$$\mu_1 \leq \mu_2 \leq \ldots \leq \mu_q \tag{9}$$

The expressions (4) and (5) yield

$$\begin{cases} E_*(N) = \sum_{j=1}^{q} r_j.(\mu_j - \mu_{j-1}) \\ \qquad = r_q - \sum_{j=1}^{q-1} \mu_j.(r_{j+1} - r_j) \quad \text{if } \mu_q = 1 \\ E^*(N) = r_q \quad \text{if } \mu_q = 1 \end{cases} \tag{10}$$

with $\mu_0 = 0$ by convention.

These results are now applied to the cases where f is the maximum operation, the minimum operation and the average operation.

i) f = max

We have $\alpha \leq \beta \Rightarrow A(\alpha) \supseteq A(\beta) \Rightarrow \max[A(\alpha)] \geq \max[a(\beta)]$. Then it can be checked that when the r_i's are increasingly ordered (i.e. (6) holds), the corresponding $\mu_i = \mu_N(r_i)$ are decreasing (i.e. (7) holds). Thus (8) applies. Let us consider the simple example given in Table 2.

R	A	B	$\mu_B(b_i)$
	8	b_1	1
	10	b_2	0.8
	7	b_3	0.6
	11	b_4	0.5

Table 2

(8) yields
$$E_*(N) = 8$$
$$E^*(N) = 8 + 0.8(10-8) + 0.5(11-10) = 10.1$$

while (3) gives
$$w(N) = \frac{8 + 10\text{x}0.8 + 11\text{x}0.5}{1 + 0.8 + 0.5} \simeq \frac{21.5}{2.3} = 9.34$$

and (3') gives
$$w'(N) = \frac{8 + 10\text{x}0.8 + 10\text{x}0.6 + 11\text{x}0.5}{1 + 0.8 + 0.6 + 0.5} \simeq \frac{27.5}{2.9} = 9.48$$

Note that it appears that $w'(N)$ is not a suitable summarizer since if we add pairs $(a_k, \mu_B(b_k))$ such that $a_k \leq 10$ $0.5 < \mu_B(b_k) \leq 0.8$, we increase $w'(N)$ whatever the values of the a_k's, which is paradoxical !

N.B.3. Besides we always have $E_*(N) \leq w(N)$ when (8) applies, but the inequality $w(N) \leq E^*(N)$ may not hold. Consider the following counter-example proposed in Table 3.

R	A	B	$\mu_B(b_i)$
	8	b_1	1
	9	b_2	0.2
	9.5	b_3	0.1

Table 3

Indeed, we obtain

$$E^*(N) = 8 + 0.2\ (9-8) + 0.1\ (9.5-9) = 8.250$$

$$w(N) = \frac{8 + 9\times0.2 + 9.5\times0.1}{1 + 0.2 + 0.1} = \frac{10.75}{1.3} = 8.269$$

□

When the $\mu_B(b_i)$ are increased, the μ_k's are increased and $E^*(N)$ increases linearly as indicated by (8). $E^*(N)$ gives a scalar estimate of the maximum of the a_i's such that the corresponding b_i's are representative elements of B ; $E_*(N)$ is a lower bound which is attached to the b_i's which undisputedly belong to B. The fuzziness of B induces an uncertainty on the answer, represented by the pair $(E_*(N), E^*(N))$; when B is crisp we have $E_*(N) = E^*(N)$ (this is true whatever f). The meaning of w(N) remains less clear and its behavior is not always completely satisfying as indicated in the following example given in Table 4.

R	A	B	$\mu_B(b_i)$
	8	b_1	1
	9	b_2	0.9

Table 4

Then we get $E^*(N) = 8 + 0.9 \times (9-8) = 8.9$ and $w(N) = \frac{16.1}{1.9} \simeq 8.47$.
We observe that when $\mu_B(b_2) \to 1$, $E^*(N) \to 9$ while $w(N) \to 8.5$, i.e. $E^*(N) \to f(a_1,a_2) = \max(a_1,a_2)$ which is intuitively satisfying ; contrastedly $w(N) \to \frac{a_1 + a_2}{2}$.

ii) f = min

We have $\alpha \leq \beta \Rightarrow A(\alpha) \supseteq A(\beta) \Rightarrow \min[A(\alpha)] \leq \min[A(\beta)]$. Then it can be checked that when the r_i's are increasingly ordered (i.e. (6) holds), the corresponding $\mu_i = \mu_N(r_i)$ are increasing (i.e. (9) holds). Thus (10) applies. Now $E_*(N)$ gives a scalar estimate of the minimum of the a_i's such that the corresponding b_i's are representative elements of B ; $E^*(N)$ is an upper bound obtained if we only consider the b_i's such that $\mu_B(b_i) = 1$. For instance, in the example of Table 2, we get

$E_*(N) = 8 - 0.6\ (8-7) = 7.4$ and $E^*(N) = 8$.

It can be seen that $w(N)$ suffers the same drawbacks as when f = max.

iii) f = arithmetic mean

Then there is no monotonicity property of the μ_i's with respect to the r_i's. Then we have to use (4) and (5) directly. Let us consider the following example where the arithmetic mean r_i and the corresponding μ_i's are given in Table 5.

$\mu_i = \mu_N(r_i)$	0.7	1	0.2	0.5
r_i	8	9	10	11
	r_1	r_2	r_3	r_4

Table 5

Then we obtain $w(N) = \frac{221}{24} \simeq 9.2$;

$$\begin{aligned} E_*(N) &= r_1.(0.7 - 0) + r_2(1-0.7) + r_3(1-1) + r_4(1-1) \\ &= r_2 - 0.7\ (r_2-r_1) = 8.3\ ; \\ E^*(N) &= r_1(1-1) + r_2(1-0.5) + r_3(0.5-0.5) + r_4(0.5-0) \\ &= r_2 + 0.5(r_4-r_2) = 9 + 0.5\text{x}2 = 10. \end{aligned}$$

Note that r_3, whose membership degree μ_3 is smaller than μ_2 and μ_4, does not appear in the computation. This behavior is general, as it can be checked on (4) and (5). Only the "convex part" of N, here $0.7/r_1 + 1/r_2 + 0.5/r_4$ is taken into account ; (a fuzzy set F defined on an ordered domain, is convex on its support $s(F) = \{r, \mu_F(r) > 0\}$ if and only if $\forall\ (x,y,z) \in s(F)^3$, $x \leq y \leq z \Rightarrow \mu_F(y) \geq \min(\mu_F(x), \mu_F(z))$). See [3].

Again the fuzziness of B induces an uncertainty about the answer, which is conveniently summarized by the pair of lower and upper expectations $(E_*(N), E^*(N))$; it gives an idea of the variability of the answer with respect to the different possible crisp interpretations of B ; this cannot be cap-red by the single number $w(N)$.

N.B.4. It can be observed that $E^*(N)$ in (8) (as well as $E_*(N)$ in (10)) is of the form

$$\sum_{j=1}^{q} m(N_j).f[N_j] \qquad (11)$$

with $m(N_j) = \mu_j-\mu_{j+1}$ (resp. $m(N_j) = \mu_j-\mu_{j-1}$) ; $f[N_j] = r_j$ and $N_j = \{r_1,\ldots,r_j\}$ (resp. $N_j = \{r_j,\ldots,r_q\}$). m is nothing but the basic proba-bility assignment in Shafer'sense [8], attached to the membership function μ_N (see [2]). The expression (11) is still equal to

$$\sum_\alpha m^*(B_\alpha).f[A(\alpha)] \qquad (12)$$

where m^* is the basic probability assignment attached to μ_B ; i.e. $m^*(B_\alpha) = \alpha-\beta$ with $B_\alpha = \{b_1,\ldots,b_k\}$ and $B_\beta = \{b_1,\ldots,b_k,b_{k+1}\}$, where $\mu_B(b_k) = \alpha$ and $\mu_B(b_{k+1}) = \beta$. If $\exists\ \alpha, \alpha'$ with $\alpha > \alpha'$ such that $f[A(\alpha)] = f[A$ the equality between (11) and (12) holds since $r(\alpha-\alpha')+r(\alpha'-\beta) = r(\alpha-\beta)$. Th expression (12), which is also an expectation (since $\sum_\alpha m^*(B_\alpha) = 1$), can be in the general case as another definition of a possible scalar answer when B is fuzzy ; however it is a single number which in general differs both from $E_*(N)$ and from $E^*(N)$ (e.g. for f = arithmetic mean). The expression (12) is used in [1] in another application context. □

N.B.5. The approach presented here can be extended to the case where our knowledge of the values of attribute A are pervaded with fuzziness and where the b_i's remain precisely known. Indeed formulas (4) and (5) can be straigh-forwardly generalized when the r_j's are fuzzy real numbers (the r_j's can st be computed since operations such as 'max', 'min' or the arithmetic mean are defined for fuzzy numbers). When the b_i's are also fuzzily known we have to distinguish between the items which are more or less possibly B and those wh more or less necessarily B ; see [6,7]. Then, the approach can be applied to the possibility degrees and the necessity degrees separately. □

[1] Dubois, D., Jaulent, M.C. (1986) A statistical approach to the analysis and the synthesis of fuzzy regions. In Tech. Rep. n° 244, LSI, Univ. P. Sabatier, Toulouse.

[2] Dubois, D., Prade, H. (1982) On several representations of an uncertain body of evidence. In : Fuzzy Information and Decision Processes (M.M. Gupta, E. Sanchez, eds.), North-Holland, 167-181.

[3] Dubois, D., Prade, H. (1985) Fuzzy cardinality and the modeling of imprecise quantification. Fuzzy Sets & Systems, 16, 199-230.

[4] Dubois, D., Prade, H. (1986) The mean value of a fuzzy number. Fuzzy Sets & Systems, to appear.

[5] Hamon, G. (1986) Extension d'un langage d'interrogation de base de données en vue de l'utilisation de questions imprécises. Thèse de Docteur-Ingénieur Univ. de Rennes, juin 1986.

[6] Prade, H. (1984) Lipski's approach to incomplete information data bases restated and generalized in the setting of Zadeh's possibility theory. Information Systems, 9, 27-42.

[7] Prade, H., Testemale, C. (1984) Generalizing database relational algebra for the treatment of incomplete/uncertain information and vague queries. Information Sciences, 34, 115-143.

[8] Shafer, G. (1976) A Mathematical Theory of Evidence. Princeton University Press, Princeton, USA.

[9] Zadeh, L.A. (1983) A computational approach to fuzzy quantifiers in natural languages. Computers and Mathematics with Applications, 9, 149-184.