A New Proposal of Aggregation Functions: The Linguistic Summary

Ignacio Blanco[1], Daniel Sánchez[2], José M. Serrano[2], and María A. Vila[2]

[1] Universidad de Almería, Departamento de Lenguajes y Computación, Camino de Sacramento s/n ,01071 Almería , SPAIN, iblanco@ual.es

[2] Universidad de Granada, Depto. de Ciencias de la Computación e I.A., E.T.S.I de Informática, Periodista Daniel Saucedo Aranda s/n, 18071 Granada SPAIN
daniel@decsai.ugr.es, jmserrano@decsai.ugr.es, vila@decsai.ugr.es
http://frontdb.decsai.ugr.es

Abstract. This paper presents a new way of giving the summary of a numerical attribute involved in a fuzzy query. It is based on the idea of offering a linguistic interpretation, therefore we propose to use a flat fuzzy number as summary. To obtain it, we optimize any index which measures the relation between the fuzzy bag (which is the answer to the fuzzy query) and the fuzzy number. Several indices should be considered: some of them are based on linguistic quantified sentences, other ones are founded on divergence measures. The method can be also used to summarize other related fuzzy sets such as fuzzy average, maximum, minimum etc.

1 Introduction

Aggregation functions are widely used in database querying. They also play a central role in data warehousing and data mining issues. Several "classical" aggregation functions are oriented to give a kind of summary for the bag of values which is the result of any database query. If we are dealing with numerical attributes the arithmetic average is the function usually employed for this purpose.

If the query involves any imprecise property, the result is a fuzzy bag of values. The current summarizing procedure consists of using the alpha cut representation, computing the attribute values average for each alpha cut and giving a "fuzzy set of means". The main drawbacks of this approach are that the result is not easily understandable and that using it in additional comparisons, arithmetic operations etc. is almost impossible. Consequently this approach does not seem to be useful for addressing problems such as nested fuzzy queries, data warehousing with imprecise dimension values etc.

For these reasons , we propose to use a flat fuzzy number as the summary of a fuzzy bag of numeric values. This fuzzy number has a direct linguistic interpretation and can be easily compared and operated. To obtain this "linguistic

T. Bilgiç et al. (Eds.): IFSA 2003, LNAI 2715, pp. 127–134, 2003.

summary", we propose to optimize any index which measures the relation between the considered fuzzy bag and the fuzzy number. Several indices should be considered: some of them are based in linguistic quantified sentences, other ones are founded in divergence measures.

We will present the problem in the next section, the following one is devoted to the mathematical formulation which leads to a constrained optimization problem. An optimization procedure is proposed in the next section where a real example is also offered. The paper finishes with some concluding remarks and the references.

2 Problem Presentation

Although the results we will present can be applied in many situations, the problem we face arises from a fuzzy query process. Frequently, the output of such a process consists of several pairs $\{(v_i, \alpha_i)\}$ where $i \in \{1, 2..n\}$; $\alpha_i \in [0, 1]$ and $v_i \in D_A$ being $D_A \subset R$ the domain of any numeric attribute. This kind of results appears when we ask for any numerical attribute value of tuples verifying an imprecise property. Some examples are the "salary of young people" or the "price of apartments located near to the beach".

It is important to remark that the pairs $\{(v_i, \alpha_i)\}$ are neither a set, nor a fuzzy set. In the output of a query like the examples above offered, there may be pairs (v_i, α_i) , (v_j, α_j) that verify $v_i = v_j$ and $\alpha_i = \alpha_j$. The mathematical structure corresponding to this situation is that of "fuzzy bag". This concept is a generalization of the fuzzy set and bag ones and it has been mainly developed by R. Yager ([YA96])

Let us consider now that the querying process has a summary objective. That is, the user does not want to know detailed values, but some kind of approximation which gives him a general idea about what is the value of an attribute for those items verifying some imprecise property. By continuing the above examples, we would ask for: "the approximate salary of young people" or for "the approximate price of apartments located near to the beach".

Additionally, it could be necessary to obtain a more specific aggregation function such as the average, the maximum or the minimum, and to consider queries such as: "the average salary of young people" or "the maximal price of apartments located near to the beach".

These last cases of specifics aggregations have been previously dealt with by E. Rudensteiner and L. Bic ([RU89]) and D.Dubois and H. Prade ([DP90]), by giving fuzzy sets as query solution. The central idea of these approaches is the following: starting from the fuzzy bag output $\widetilde{B} = \{(v_i, \alpha_i)\}$, we obtain the sequence $\{\beta_j\}, j \in \{1...m\}$; $\beta_j < \beta_{j+1}$; $\beta_m = 1$ of the different membership values which can appear in $\widetilde{B}$.

By obtaining the corresponding α-cut of $\widetilde{B}$ we can generate a crisp bag sequence $\{B_{\beta_j}\}$, and the application of the considered aggregation function $f(.)$ (average, maximum, minimum etc.) to each crisp bag gives us a sequence of pairs $\{(u_j, \beta_j)\}$ where $\forall j \in \{1, ..., m\}$ $u_j = f(\beta_j)$. This sequence can be easily transformed in a fuzzy set by using:

$$\forall v \in D_A \ \mu_f(v) = \begin{cases} sup\{\beta_j / u_j = v\} & \text{if } \exists j \in \{1, ...n\} \quad v = u_j \\ 0 & \text{otherwise} \end{cases}$$

However, in our opinion, offering a fuzzy bag or a fuzzy set as a solution for a summary query, may be meaningless, since a non expert user could be not able to understand it.

Additionally, the result of the query could be the input for another processes, such as to "nested queries", for example to ask for: "workers whose salary is greater than the average salary of young people". If the "average salary" is a fuzzy bag or a fuzzy set, solving this query is very difficult (if not impossible), since there is not a comparison procedure for general fuzzy sets.

For all these reasons we propose to give a fuzzy number as the result of a summary query either when we consider the whole original fuzzy bag, or we take any aggregation function which produces an intermediate fuzzy set.

It should be remarked our problem should be viewed a particular case that of the linguistic aproximation (LA) one. This problem has been mainly dealed with in the context of fuzzy control or fuzzy expert systems (see [[ko99],[Wha01]) and most of the proposed methods assume there is a set of label previously established as well as the item to be approximated is a fuzzy set. Since we start from different context (fuzzy querying processes) our approach is also quite different

3 Mathematical Formulation of the Problem

According to consideration above, our problem is focused in the following way:

"Given a fuzzy numerical bag $\widetilde{B}$ with support $[a, b] \in \mathbf{R}$ to obtain a fuzzy number $\widetilde{F}$ defined on the same support such that: "$\mathcal{M}(\widetilde{B}, \widetilde{F})$*is minimal*" where $\mathcal{M}$ stands for some kind of measure of distance, divergence, opposite to an approximation measure etc.. between $\widetilde{B}$ and $\widetilde{F}$.

Since the fuzzy number should actually be an approximation we may to consider it is a trapezoidal fuzzy number, which is easier to interpret and to compute. Therefore the problem can be formulated as follows:

$$\begin{aligned} &\text{Minimize} && \mathcal{M}(\widetilde{B}, \widetilde{F}(m_1, m_2, a_1, a_2)) \\ &\text{Subject to} \\ & && h_k(m_1, m_2, a_1, a_2) \leq 0 \ \forall k \in \{1, ..l\} \end{aligned} \tag{1}$$

where:

- $\widetilde{F}(m_1, m_2, a_1, a_2)$ is a trapezoidal fuzzy number with support $[m_1 - a_1, m_2 + a_2]$ and mode $[m_1, m_2]$
- $\forall k \in \{1, ..l\}$; $h_k(m_1, m_2, a_1, a_2) \leq 0$ stands for any constraint associated with the fuzzy number such as $a_1 \geq 0$ or $m_1 \leq m_2$. A more specific formulation of these constraints will offered in the follwing section 3.1.

It is clear that the optimization problem given in (1) has (m_1, m_2, a_1, a_2) as variables and we have to obtain them by solving it. However before to do it, we must to concrete the problem elements.

3.1 The Measure $\mathcal{M}$

$\mathcal{M}$ should be either a divergence measure or the opposite to a compatibility measure. We have considered both possibilities.

Divergence Measures The divergence between two fuzzy sets can be established in an axiomatic way and any fuzzy measure of this divergence weights their distance. A general study of this can be found in [MO02]. Among other divergence examples, this article presents the Hamming distance between fuzzy sets. We will use this measure by considering:

$$\mathcal{M}(\widetilde{B},\widetilde{F}) = (\sum_{i=1}^{n} \|\alpha_i - \widetilde{F}(v_i)\|)/n$$

where $\widetilde{F}(.)$ denotes the membership function of $\widetilde{F}$

Compatibility Measures Since we have already defined a distance-based measure, a new approach is necessary if we want to consider truly different measures. Therefore we will consider that a good approximation to the fuzzy bag $\widetilde{B}$ could be the fuzzy number $\widetilde{F}$ which maximize the accomplishment degree of the quantified sentence

$$"Q \text{ elements of } \widetilde{B} \text{ are } \widetilde{F}" \tag{2}$$

where Q stands for any linguistic quantifier such as "most", "almost all", "all" etc. And we will define:

$$\mathcal{M}(\widetilde{B},\widetilde{F}) = \textit{oposite of}\{\text{accomplishment degree of (2)}\} \tag{3}$$

A wide study on the evaluation of quantified sentences and their relations with the different approaches to the fuzzy cardinal can be found in [DSV99].

The basic idea of this paper is that the accomplishment degree of the sentence: "Q *of* $\widetilde{D}$ *are* $\widetilde{A}$" can be obtained by means of the degree of matching between the quantifier ,(defined as a fuzzy set on [0,1]), and the "relative cardinal" of $\widetilde{A}$ with respect to $\widetilde{D}$, given also by a fuzzy set on [0,1]. The paper proposes two methods for computing this cardinal and the matching degree:

– *Possibilistic method*
Let $M(\widetilde{A}) = \{\alpha \in [0,1] | \exists x_i \text{ such that } \widetilde{A}(x_i) = \alpha\}$, and:

$$M(\widetilde{A}/\widetilde{D}) = M(\widetilde{A} \cap \widetilde{D}) \cup M(\widetilde{D})$$

$$CR(\widetilde{A}/\widetilde{D}) = \left\{ \frac{|(\widetilde{A} \bigcap \widetilde{D})_\alpha|}{|\widetilde{D}_\alpha|} \text{ such that } \alpha \in M(\widetilde{A}/\widetilde{D}) \right\}$$

the relative cardinality of $\widetilde{A}$ with respect to $\widetilde{D}$ is defined as:

$$\forall c \in CR(\widetilde{A}/\widetilde{D}) \ : \ ES(\widetilde{A}/\widetilde{D}, c) = \max \left\{ \alpha \in M(\widetilde{A}/\widetilde{D}) | c = \frac{|(\widetilde{A} \bigcap \widetilde{D})_\alpha|}{|\widetilde{D}_\alpha|} \right\}$$

The matching degree with any quantifier Q is given by:

$$ZS_Q(\widetilde{A}/\widetilde{D}) = \max_{c \in CR(\widetilde{A}/\widetilde{D})} \min(ES(\widetilde{A}/\widetilde{D}, c), Q(c))$$

– *Probabilistic Method*
Let us consider that the set $M(\widetilde{A}/\widetilde{D})$ defined in the above paragraph is ordered, that is $M(\widetilde{A}/\widetilde{D}) = \{\alpha_1, >, ..., > \alpha_{m+1}\}$ with $\alpha_1 = 1$ and $\alpha_{m+1} = 0$. Let

$$C(\widetilde{A}/\widetilde{D}, \alpha_i) = \frac{|(\widetilde{A} \bigcap \widetilde{D})_\alpha|}{|\widetilde{D}_\alpha|}$$

Then the relative cardinality of $\widetilde{A}$ with respect to $\widetilde{D}$, $ER(\widetilde{A}/\widetilde{D})$ is defined as:

$$\forall c \in CR(\widetilde{A}/\widetilde{D}) \ : \ ER(\widetilde{A}/\widetilde{D}, c) = \sum_{c=C(\widetilde{A}/\widetilde{D}, \alpha_i)} (\alpha_i - \alpha_{i+1})$$

The matching degree with any quantifier Q is given by:

$$GD_Q(\widetilde{A}/\widetilde{D}) = \sum_{c \in CR(\widetilde{A}/\widetilde{D})} ER(\widetilde{A}/\widetilde{D}, c) \times Q(c)$$

Verifying that both methods are usable in the fuzzy bag case is straightforward, since the involved processes are union, intersection, and computing the cardinal of an $\alpha-$cut are defined also for fuzzy bags.

Therefore, chosen any quantifier Q, we can use $ZS_Q(\widetilde{B}/\widetilde{F})$ or $GD_Q(\widetilde{B}/\widetilde{F})$ as the compatibility degree appearing in (3).

3.2 The Constraint System

The constraints defined on the four parameter of $\widetilde{F}$: (m_1, m_2, a_1, a_2) can be divided in two classes:

Those oriented to assure that $\widetilde{F}$ is a fuzzy number, which are:

$$\begin{array}{ll} -a_1 & \leq 0 \\ -a_2 & \leq 0 \\ m_1 - m_2 & \leq 0 \end{array} \tag{4}$$

And those oriented to fit the fuzzy number $\widetilde{F}(m_1, m_2, a_1, a_2)$ inside the $\widetilde{B}$ area. If $[a, b]$ is the minimal interval including the $\widetilde{B}$ support and $[c, d]$ the minimal interval including the $\widetilde{B}$ mode, we impose:

$$\begin{array}{ll} -m_1 + a_1 + a & \leq 0 \\ m_2 + a_2 - b & \leq 0 \\ m_1 - m_2 - (d - c) & \leq 0 \end{array} \tag{5}$$

Additionally, to control the imprecision of the obtained solution, we will impose that the fuzziness of $\widetilde{F}$ should be least than or equal to that of the $\widetilde{B}$:

$$\mathcal{F}(\widetilde{F}) - \mathcal{F}(\widetilde{B}) \leq 0 \tag{6}$$

We propose the Delgado et al.'s fuzziness measure (see [DVV98]) as fuzzy measure. The adaptation of this measure to the case of $\widetilde{B}$ is given by:

$$\mathcal{F}(\widetilde{B}) = \sum_{\beta_j \leq 1/2} (R(\beta_j) - L(\beta_j))(\beta_j - \beta_{j-1}) - \sum_{\beta_j > 1/2} (R(\beta_j) - L(\beta_j))(\beta_j - \beta_{j-1})$$

where the sequence $\{\beta_j\}$ has been defined in the section 2 and

$$\forall \alpha \in [0,1]\ ;\ R(\alpha) = \sup\{x | x \in \widetilde{B}_\alpha\} \text{ and } L(\alpha) = \inf\{x | x \in \widetilde{B}_\alpha\}$$

The following property can be also found in [DVV98].

Property 1. The fuzziness of any trapezoidal fuzzy number $\widetilde{F}(m_1, m_2, a_1, a_2)$ is given by $m_2 - m_1 + (a_2 - a_1)/2$

This property gives us the constraint (6):

$$1/2a_2 - 1/2a_1 + m_2 - m_1 - \mathcal{F}(\widetilde{B}) \leq 0 \tag{7}$$

4 The Optimization Process

According to the formulation given above, the problem of obtaining a summary for a fuzzy query numerical answer can be focused in solving the following optimization problem:

$$\begin{array}{ll} \text{Minimize} & \mathcal{M}(\widetilde{B}, \widetilde{F}(m_1, m_2, a_1, a_2)) \\ \text{Subject to} & \end{array} \tag{8}$$

$$\begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & -1 & 1 \\ -1/2 & 1/2 & -1 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ m_1 \\ m_2 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \\ -a \\ b \\ (d-c) \\ \mathcal{F}(\widetilde{B}) \end{pmatrix} \leq \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Clearly, (8) is a constrained optimization problem, with a linear constraint system and with an objective function whose properties can vary depending on what alternative is chosen among those ones presented in the Section 3.1. Moreover, if we take some compatibility-based measure we have no formal expression for the objective of (8), although we can evaluate this objective at every point. Therefore, if we want to design an optimization process general enough to cover all possible objectives, it is necessary to use a direct search optimization method which only needs the objective function values.

For these reasons we have chosen a penalty method ([FC68]) to deal with the constraints and a direct search method to solve the unconstrained problems associated to the penalty approach, concretely we have used the Hooke-Jeeves search method. ([HJ66]).

This optimization procedure provides us with a way to get the trapezoidal fuzzy number which approximates the fuzzy bag, by considering the different kind of measures and even, in the case of compatibility-based measures, different linguistic quantifiers. The following example illustrates how this procedure can be used and its usefulness in order to provide a summarized answer.

Example 1. We have considered the marks obtained by a group of students in the subjects "Databases" and "Programming Languages". Our goal is now to know what is the mark in "Programming Languages" of those students which have attained a "good mark" in "Databases". We have applied the optimization procedure to the fuzzy bag direct result obtained from the query and the fuzzy sets average, maximum and minimum. The chosen objective has been the probabilistic compatibility measure and the linguistic quantifier "most", whose membership function is $\forall x \in [0,1]\ ;\ Q(x) = x$. The results appears in the figure

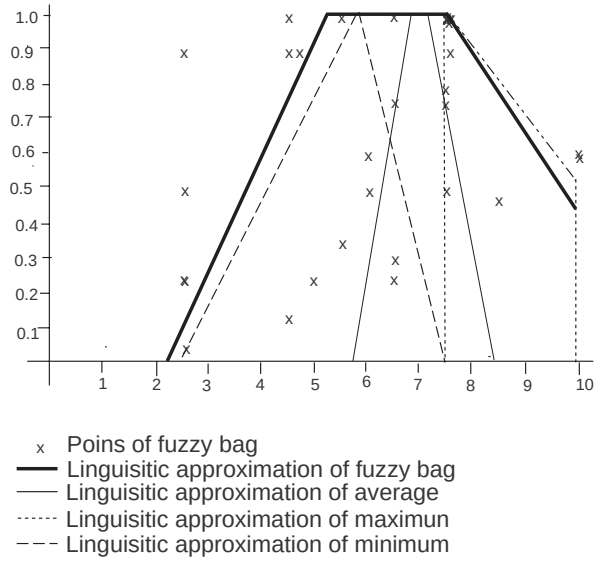

Fig. 1. Results of example

Now it is possible to offer linguistic interpretations of the results such as:

- The marks in "Programming Languages" of students which are good in "Databases" are "more or less" between 5 and 7.5.

- The average of marks in "Programming Languages" of students which are good in "Databases" is more or less between 6.90 and 7.15.

- The maximum of marks in "Programming Languages" of students which are good in "Databases" is a bit greater than 7.5 and less than 10.

- The minimum of marks in "Programming Languages" of students which are good in "Databases" is around 5.5.

5 Concluding Remarks

We have presented a new way of giving the summary of a numerical attribute involved in a fuzzy query. This is based on the idea of offering a linguistic interpretation as result and it can be applied either to the initial fuzzy bag or to the fuzzy set that results from considering the average, maximum, minimum or, in general, every aggregation function.

A procedure to obtain this linguistic summary is also presented. It is founded in solving an constrained optimization problem whose objective function can adopt different formulations depending on what is the measure we choose.

A real example show us how this new approach can be more understandable for a non expert user, besides the possibility of using this result in an additional process such as a nested query.

However this is only an initial proposal. Deeper studies are necessary in order to improve the approach, concretely:

• We are carrying out an experimental analysis in order to decide on the more suitable objective function, by considering divergence or compatibility-based measures and, for this second case, by using different linguistic quantifiers.

• All this process is going to be implemented in the prototype of a fuzzy database system, developed by our working group for tuning the optimization algorithm and for adapting it to solve real problems.

The results of these studies will be offered later on in a forthcoming work.

References

[DVV98] M. Delgado,M.A. Vila,W Voxman "A fuzziness measure for fuzzy numbers: Applications" *Fuzzy Sets and Systems 94* pp.205-216 1998.

[DSV99] M.Delgado, D. Sanchez and M.A. Vila " Fuzzy cardinality based evaluation of quantified sentences" *Int. Journal of Approximate Reasoning* Vol. 23 pp. 23-66 2000.

[DP90] D. Dubois, H. Prade, "Measuring properties of fuzzy sets: a general technique and its use in fuzzy query evaluation" *Fuzzy Sets and Systems* 38, 137-152, 1990.

[FC68] Fiacco A. V. and McCormick G.P. *Non Linear Programming: Sequential Unconstrained Minimization Techniques* J. Wiley New York 1968.

[HJ66] Hooke R., Jeeves T.A. "Direct Search of Numerical and Statistical Problems" *J. ACM* 8 212-229 1966.

[[ko99] Kowalczyk R. "On numerical and linguistc quantifcation in linguistic approximation" *Proceeding of IEEE- SMC'99* pp. vol. 5 pp. 326-331 1999

[MO02] S. Montes, I. Couso, P. Gil, C. Bertoluzza "Divergence measure between fuzzy sets" *Int. Journal of Approximate Reasoning* V. 30 pp. 91-105, 2002.

[RU89] E.A. Rudensteiner, L. Bic, "Aggregates in possibilistic databases" *Proceedings of the Fifteenth Conference on Very Large Database* (VLDB'89), Amsterdam (Holland), 287-295, 1989.

[Wha01] Whalen T., Scott B. "Empirical comparison of techniques for linguistic approximation" *Proceeding of 9th IFSA Congress* Vol. 1 pp.93-97 2001

[YA96] R.R. Yager "On the Theory of Bags" *Int. Journal on General Systems* V. 13 pp. 23-37 1986