

ON AN EXTENSION OF FUZZY AGGREGATE FUNCTIONS FOR DATABASES

Ana Aguilera¹
Ramon Mata-Toledo²
Alberto Subero¹
Morgan Monger³
Pranshu Gupta⁴

ABSTRACT

Today, Fuzzy sets and fuzzy logic are used in many applications areas such as control, instrumentation, resource management, and databases just to name a few. In this paper, the authors present an extension of fuzzy aggregate functions in the context of fuzzy databases. The paper is not intended as a complete examination of relational databases and the SQL language.

Keywords – Fuzzy logic, Fuzzy sets, Aggregate, Fuzzy operator, Fuzzy relations.

1. INTRODUCTION

The notion of fuzzy sets (Zadeh, 1965) has found ample and wide applications in a range of areas of finance, and control and management information systems. Fuzzy sets allow us to represent vague (or fuzzy) natural language concepts that may vary depending on the context in which they are used. Formally, the definition of a fuzzy set requires two basic components: a universe of discourse or domain and a function, called the membership function, which defines the “degree” to which a particular element of the domain belongs or not belongs to the set. More formally, we can define a fuzzy set, denoted by $\tilde{A}$, as follows:

$$\tilde{A} = \{(x, \mu_A(x)) \mid x \in U\}$$

where U is the universe of discourse and $\mu_A: U \rightarrow [0,1]$ is the membership function.

Fuzzy sets are an extension of the classic notion of sets (Bourbaki, 1968) called, for sake of understanding, crisp sets. A set A of the latter type can be defined, in terms of the membership function, as

$$A = \{(x, \mu_A(x)) \mid x \in U\}, \mu_A(x) = 0 \text{ if } x \notin A \text{ and } \mu_A(x) = 1 \text{ if } x \in A$$

Observe that in the case of crisp set, the membership function can only takes the extreme values of the interval $[0,1]$.

A Fuzzy set representation may vary even for similar concepts. Table 1, adapted from (Aguilera, 2011), shows examples of fuzzy sets where the domain is a discrete set of ages and the membership functions defined imprecise criteria such young, middle age, and old.

¹ Universidad de Carabobo (Valencia, Venezuela)

² James Madison University (Harrisonburg, Virginia, U.S.A.) matatora@jmu.edu

³ Ellucian, Inc. (Fairfax, Virginia, U.S.A.) gprashu@ksu.edu

⁴ DeSales University (Center Valley, Pennsylvania, U.S.A.) mongrmd@gmail.com

It can be observed directly from the table that depending on the function a person may have different degrees of membership for the same universe of discourse. Figure 1 shows the trapezoidal membership functions associated with the values shown in the Table 1.

Age	μ_{Young}	$\mu_{\text{Middle Age}}$	μ_{Old}
51	0	0.9	0.1
40	0	1	0
24	0.6	0.4	0
39	0	1	0
28	0.2	0.8	0
54	0	0.6	0.4
29	0.1	0.9	0
61	0	0	1
22	0.8	0.2	0
62	0	0	1

Table 1. Membership function values for ages

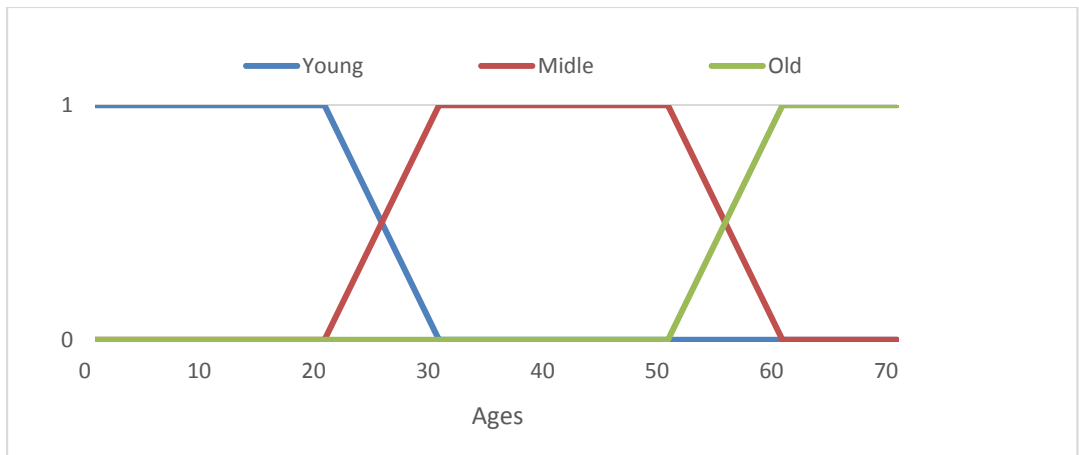


Figure 1. Membership function for ages

2. AGGREGATE OPERATORS OF FUZZY SETS

Several contributions on this area have appeared in the literature (Merigó, Fuller, Yager 2007). Here the authors define a new aggregate operators for fuzzy sets. They are as follows:

Definition 1. Let $\tilde{A} = \{(x, \mu_A(x)) \mid x \in U\}$ a fuzzy set with membership function $\mu_A(x)$. Let's define the **scalar cardinality of $\tilde{A}$** , denoted by $|\tilde{A}|$, as

$$|\tilde{A}| = \sum_{i=1}^n \mu_i(x)$$

Definition 2. The **contribution** denoted by w of an i^{th} element $(x_i, \mu_i(x))$ in a fuzzy set is defined as

$$w_i = x_i * \mu_i(x)$$

where $w_i \in \mathbb{R}$. The set $\{w_i\}$ is totally ordered. Therefore, the minimum and a maximum element can be defined. We will do this next.

Definition 3. Let $\tilde{A} = \{(x, \mu_A(x)) \mid x \in U\}$ be a fuzzy set. Let $(x_i, \mu_i(x)) \in \tilde{A}$ and let w_i be the contribution of this element.

$(x_i, \mu_i(x))$ can be define as

a) **Minimum of $\tilde{A}$** is w_i is minimum if $w_i \leq w_k (1 \leq k \leq n)$.

b) **Maximum of $\tilde{A}$** is w_i is maximum if $w_i \geq w_k (1 \leq k \leq n)$.

This way, we take as maximum the element that “contributes the most” to the set. Likewise, we define as minimum the element that contributes “the least” to the set. These definitions clearly extend the classic concept of maximum and minimum of the set when the degree of membership is 1.

Definition 4. Let $\tilde{A} = \{(x, \mu_A(x)) \mid x \in U\}$ be a fuzzy set; let w_i be the contribution of an i th element on $\tilde{A}$. The **weighed sum** in $\tilde{A}$, denoted by s , is defined as

$$s = \sum_{i=1}^n w_i$$

Definition 5. Let $\tilde{A} = \{(x, \mu_A(x)) \mid x \in U\}$ be a fuzzy set. Let s be a weighed sum and let $|\tilde{A}|$ be the cardinality of $\tilde{A}$. The weighed average, denoted by a , is defined as

$$a = \frac{s}{|\tilde{A}|}$$

3. RELATIONAL DATABASE CONCEPTS

Traditional databases (Codd, 1970) can be formally defined, in mathematical terms, as consisting of a set of mappings in a given relation schema (Maier 83). A *relation scheme*, R , is a set of attribute names $R = \{A_1, A_2, \dots, A_n\}$ where $D_i = \text{Domain}(A_i)$ $1 \leq i \leq n$ are arbitrary,

non-empty sets, finite or countably infinite. If $D = \bigcup_{i=1}^n D_i$, a *relation* r (or table) on R is

defined as a set of mappings $\{t_1, t_2, \dots, t_p\}$ from R to D where for each mapping $t_j \in r$, $t_j(A_i) \in D_i$ with $1 \leq j \leq p$. Each one of these mappings is called a *tuple* (or row). Following Maier (83) we have used mappings instead of the traditional concept of relations as subsets of the Cartesian Product of sets to avoid the explicit ordering of attribute names in the relational schema. We can extend the previous concept of relation to fuzzy sets with the understanding that for each for each mapping $t_j \in r$, $\mu(A_i) \in [0, 1]$.

4. AGGREGATE OPERATION IN RELATIONAL ALGEBRA

In this paper, and following the approach of Cox (1995), the authors also focus on the use of fuzzy logic in the formulation of queries against a standard relational database using relational algebra. By the latter term we mean a formal system for manipulating relations. Here, the use of fuzzy predicates aims to widen the qualifications specifications of the aggregate or group functions to include approximate boundaries. Paraphrasing Cox (1995) we want to obtain what we “intend” to retrieve instead of what we must retrieve due to the mechanical...nature of the crisp aggregate” functions. Within the scope of this paper we are interested in the extensions of the crisp group functions or aggregate such as Average (avg), Minimum (min), Maximum (max), SUM (sum), and COUNT (count). Each of these functions operates on a set of rows and produces a single result. In general, and in terms of relational algebra, we can define an aggregate operation as follows:

$$G_1, G_2, \dots, G_n \text{ G } F_1(A_1), F_2(A_2), \dots, F_n(A_n) (E) \text{ where}$$

E is any relational-algebra expression

$G_1, G_2 \dots, G_n$ is a list of attributes on which to group (can be empty)

Each F_i is an aggregate function

Each A_i is an attribute name

Table 3 shows the result of the query that counts “the number of instructors in each department” on the relation indicated in Table 2. Expressed in relational algebra, the query looks like this:

$$\text{Dept_name G COUNT(instructor)}$$

Fuzzy Relational Aggregate Operators

ID	Instructor	Dept_name	Salary	Age
17	Crick	Biology	72000	51
76	Katz	Computer science	75000	40
26	Srinivasan	Computer science	65000	24
12	Brandt	Computer science	92000	39
19	Kim	Electric engineering	80000	28
08	Wu	Finance	90000	54
31	Singh	Finance	80000	29
09	El Said	History	60000	61
44	Califieri	History	62000	22
23	Mozart	Music	40000	62

Table 2. Instructor relational table

Dept_name	count
Biology	1
Computer science	3
Electric engineering	1
Finance	2
History	2
Music	1

Table 3. Query result

Based on the definitions 1 through 5 we can now define the fuzzy aggregate relational database operators shown next.

5. COUNT

We can define this operator as follows:

$Count(L_i) = \sum_{t \in r \wedge t.A \in L_i} \mu_{\psi}(t)$ where L_i is an alias or label of the attribute $t.A$ of the tuple t of the relation r and ψ is a fuzzy condition over the relation.

If we apply this definition to the data set of Table 1 given above we will have result given in Table 4. That is the quantity of young, middle ages and old people in the relation instructor.

Label(age)	Count
young	0.8
middle-age	3.7
old	1.5

Table 4. Count

Bosc (2010) also introduced a variation of the aggregate function COUNT() on fuzzy sets defined over a fuzzy partition defined as follows:

$$count(L_i) = \sum_{t \in r} \min(\mu_{\omega}(t), \mu_{L_i}(t.A))$$

Where L_i is the label or alias of an attribute, $\mu_{\omega}(t)$ is a fuzzy condition, and $\mu_{L_i}(t.A)$ defines the fuzzy condition on the column or attribute A for the t tuples of the relation.

To illustrate the use of this definition, consider, for example, the membership function on the salary attributes defined as low, medium and high of Figure 2 and Table 6. For the query: “what is the distribution over the different age classes of the employees whose salary is medium?” the result is shown in the Table 7.

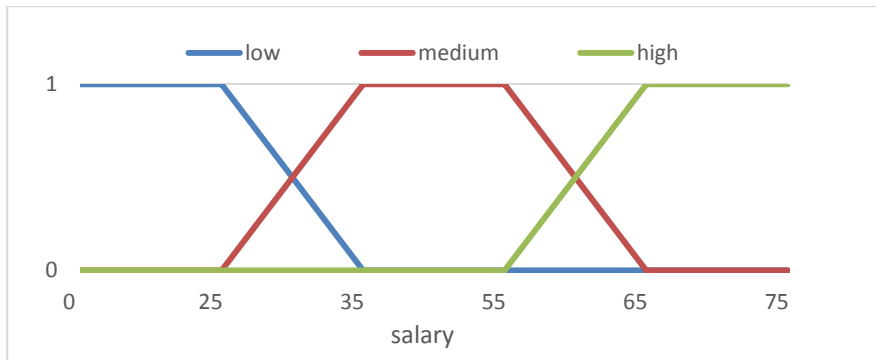


Figure 2. Membership function for salary

ID	Age	μ_{young}	$\mu_{\text{middle-age}}$	μ_{old}	Salary	μ_{low}	μ_{Medium}	μ_{High}
17	51	0	0.9	0.1	65	0	0	1
76	40	0	1	0	45	0	1	0
26	24	0.6	0.4	0	19	1	0	0
12	39	0	1	0	32	0.3	0.7	0
19	28	0.2	0.8	0	24	1	0	0
08	54	0	0.6	0.4	57	0	0.8	0.2
31	29	0.1	0.9	0	18	1	0	0
09	61	0	0	1	15	1	0	0
44	22	0.8	0.2	0	45	0	1	0
23	62	0	0	1	59	0	0.6	0.4

Table 6. Membership values for ages and salaries

6. MIN

We can define this operator as

MIN= (t.A, $\mu_{\psi}(t)$) where is (t.A, $\mu_{\psi}(t)$) is minimum according to the Definition 3 given above.

In other words, the minimum is the attribute, in a tuple, that contributes the least.

Example query: “what is the minimum age for young instructors?” According to Table 8 the minimum is MIN = (29, 0.1)

Label(age)	Count
young	0.8
middle-age	2.5
old	1

Table 7. Count defined by Bosc (2010)

Age	μ_{Young}	w_i
51	0	0
40	0	0
24	0.6	14.4
39	0	0
28	0.2	5.6
54	0	0
29	0.1	2.9
61	0	0
22	0.8	17.6
62	0	0

Table 8. The contribution for the attribute age in a fuzzy set of young

7. MAX

We can define this operator as

MAX= (t.A, $\mu_{\psi}(t)$) where is (t.A, $\mu_{\psi}(t)$) is maximum according to the Definition 3 given above.

In other words, the minimum is the attribute, in a tuple, that contributes the most.

Example query: “what is the maximum age for young instructors?” According to Table 8 the maximum is

$$\text{MAX} = (22, 0.8)$$

8. SUM

We can define this operator as

$$\sum_{i=1}^n w_i \text{ where } w_i \text{ is the contribution defined according to the Definition 4 given}$$

above

9. AVG

We can define this operator as the SUM divided by the Cardinality (Definition 1 given above)

10. IMPLEMENTATION

A first implementation attempt of the fuzzy aggregate COUNT was performed using the database management system PostgreSQL. The catalog was modified first to include the new aggregate functions. The catalog of a relational database management system stores the metadata of the relational schema such as all table related information (indexes, attributes, constraints, privileges, etc). The parser was modified next. Here a new type of node was created. This new node is called A Partition and contains all relevant data of the partition such as is its domain, their labels, among other things. A pointer (UsingClause) to a list of nodes partition was added to the SelectStmt node. Within the module planner the tree search is verified and, from it, a fuzzy plan is created. Additional plans are also created within this module to calculate the necessary cardinalities to evaluate the fuzzy aggregate functions. The run-time manager uses the different execution plans in a recursive manner to obtain the required sets. The run-time manager also calculates the degrees of membership according to the fuzzy predicates that were defined.

An example of query for a fuzzy aggregate with fuzzy partition in a SQL expression is shown in Figure 3.

lab_marcha=# select label(edad), count(*), count_p(*), count_prel(*) from estudio where peso = gordo group by label(edad) using partition(edad) = {nino, adolescente};			
label(edad)	count	count_p	count_prel
adolescente	2	0.0600006	0.000766781
nino	4	3.5	0.00253531
(2 rows)			
lab_marcha=# select label(edad), count(*), count_p(*), count_prel(*) from estudio where peso = flaco group by label(edad) using partition(edad) = {maduro, viejo};			
label(edad)	count	count_p	count_prel
maduro	15	11.8182	0.248092
viejo	3	1.7	0.0977011
(2 rows)			
lab_marcha=# select label(edad), count(*), count_p(*), count_prel(*) from estudio where talla = bajo group by label(edad) using partition(edad) = {nino, adolescente, maduro, viejo};			
label(edad)	count	count_p	count_prel
adolescente	136	77.75	0.99361
nino	1432	1333.5	0.965954
maduro	60	46.8182	0.982825
viejo	15	8.9	0.511494
(4 rows)			

Figure 3. Query result in PostgreSQL

11. ACKNOWLEDGEMENTS

The authors appreciate the financial support of the Fulbright program at James Madison University, USA and the CDCH of Carabobo University, Venezuela to Dr. Ana Aguilera.

12. REFERENCES

- Bosc, P. ; Pivert, O. On a fuzzy group-by clause in SQLf IEEE International Conference on Fuzzy Systems (FUZZ), Pp. 1-6, 2010.
- Bourbaki, Nicolas. Éléments of Mathématique, Théorie des Ensembles, Chap 2. Hermann, Paris, 1968
- Codd, E.F. A Relational Model of Data for Large Shared Data Banks. CACM 13:6, June 1970, 377-387
- Cox, Earl. Fuzzy Logic for Business and Industry. Charles River Media, Inc. 1995.
- Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8, 338–353. doi:10.1016/S0019-9958(65)90241-X
- Merigó, J.M; Casanova, M. Generalized Aggregation Operators and Fuzzy Numbers in a Unified Model between the Weighted Average and the OWA Operator. http://www.fcee.urv.es/sigef/english/congressos/congres15/039_Merigo_Casanovas.pdf. Date unknown.
- Fuller, R. OWA Operators in Decision Making <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.22.7664>
- Yager, R. R. Final Report: Information Fusion and Aggregation for Cooperative Systems. <http://www.dtic.mil/dtic/tr/fulltext/u2/a487284.pdf>. 2007