Context-Dependent Fuzzy Queries in SQLf

Claudia Jiménez[1], Hernán Álvarez[2], and Leonid Tineo[3,4]

[1] Department of Computer Science, National University of Colombia, Medellín, Colombia
csjimene@unalmed.edu.co
[2] Department of Processes and Energy, National University of Colombia, Medellín, Colombia
hdalvare@unalmed.edu.co
[3] Departamento de Computación, Universidad Simón Bolívar, Caracas, Venezuela
leonid@usb.ve
[4] Centro de Análisis, Modelado y Tratamiento de Datos, CAMYTD
Facultad de Ciencias y Tecnología, Universidad de Carabobo, Venezuela

Abstract. Fuzzy set theory has been used for extending the database language capabilities in order to admit vague queries. SQLf language is one of the most recognized efforts regarding this tendency, but it has limitations to interpret context-dependent vague terms. These terms are used as filtering criteria to retrieve database objects. This paper presents an improvement of SQLf language that adding inductive capabilities to the querying engine. This allows the discovering of the semantics of vague terms, in an autonomous and dynamic way. In the discovering of the meaning of vague terms, looking for more flexibility, our proposal considers different granularity levels in the fuzzy partitions required for the object categorization.

Keywords: Adaptive Fuzzy Systems, Flexible Querying, Fuzzy Database Technology, Fuzzy Partition.

1 Introduction

This work addresses the problem of representation and management of context-dependent vagueness as a key strategy to bring query languages even closer to natural language. Queries with imprecision, unlike to classical queries, allow users to express their information requirements, easily, as if they were interacting with an expert friend, who can understand all vague terms used in the query specification.

Several proposals have emerged extending the database SQL language in order to admit vague queries. These proposals are regularly based on fuzzy sets theory [2] [3] [9]. At present time, there is a rising interest for this field in research community, it promises success in development of emerging applications [4] [13] [16] [10] [18]. In those proposals, fuzzy sets representing vague terms are given by experts or users. However, in some cases, this subjectivity could affect the reliability of answer. This is because the meaning of vague terms might vary depending of particular context and not just according human perception.

Typically, the meaning of many vague adjectives used for qualifying objects changes according to the context being considered. For instance, an *expensive* price

R. Meersman et al. (Eds.): OTM 2012, Part II, LNCS 7566, pp. 763–779, 2012.

for a booking in Medellín could be considered *cheap* in other city like Vienna. Even the meaning of the *cheap* adjective may differ when hotels are restricted to be in specific zone of the city. Moreover, taking into account only the hotels that offer some specific amenities, the kind of hotels, which can be catalogued as *cheap*, could be a different class from the examples mentioned before. That is to say, there are many possible meanings for the same linguistic label and, it is necessary to delimit the context surrounding the vague terms in order to determine their appropriate interpretation.

The subjectivity problem in the definition and usage of fuzzy concepts, and the dependency between these concepts and the context, was recognized as an important difficulty to be solved in fuzzy database technology [4]. In this direction, some of the proposals try to consider the context-dependent vague terms by allowing users to directly specify the semantics of vague conditions of a query [1]. Such specification includes a new data definition sentence that specifies the fuzzy set representing the vague terms in a detailed manner. Additionally, a derivation formula can be applied to the sub-domain obtained from restricting a relation with other vague condition [1]. However, in these proposals, the subjectivity problem prevails. With static fuzzy sets definition, it is not possible to consider all probable contexts that can be specified by each query, nor the effect of time.

In order to overcome the subjectivity problem, and considering that the meaning of many vague terms is variable, we have conceived a data-driven model. This goal was achieved by giving inductive reasoning capabilities to the querying engine. This machine would identify the meaning of the labels by examining the theoretical fuzzy models. For this, we propose to obtain the values of corresponding parameters using percentiles derived from the contextual data producing reliable answers from flexible querying systems. In the discovering of the meaning of vague terms, this proposal, does not only considers the context delimited by each query, but also the different granularity levels in the object discrimination.

This paper is structured as follows: — Section 2 states presents the Theoretical Foundations of this work; — Section 3 describes the language extensions for Categorization Specification; — Section 4 explains the Vagueness Interpretation of context-dependent fuzzy queries; — Section 5 deals with operators for defining Complex Predicates; — Section 5 includes some experimental results using a benchmark database. Section 7 considers some Conclusions and Future Works.

2 Theoretical Foundations

In this work, we employ a fuzzy partition over some domain, to determine the semantics of a label L in based to the quantitative attributes of database objects. Therefore, we present the main concepts of fuzzy logic involved in this work as well as some statistics fundamentals.

Definition 1: *A* ***fuzzy set*** *is a generalization of conventional set concept. Let U be a specific domain, a fuzzy set A is determined by* $\mu_A: X \rightarrow [0,1]$*, called the* ***membership***

function. *For any x∈ U, the measure $\mu_A(x)$ is known as the* ***membership degree*** *of x. Thus,* ***a fuzzy set*** *is defined as a collection of ordered pairs:*

$$A = \{(x, \mu_A(x)) \mid x \in U\} \tag{1}$$

Definition 2: *In fuzzy set theory, the* ***empty set*** $\varnothing$ *and the* ***universal set*** *U are defined by the following axioms.*

$$\mu_\phi(x) = 0 \; \forall x \in U \quad \text{(Empty set)}$$
$$\mu_U(x) = 1 \; \forall x \in U \quad \text{(Universal set)} \tag{2}$$

Definition 3: *Let U be a numeric domain, $x_1 \leq x_2 \leq x_3 \leq x_4 \in U$, we define the* ***linear shape membership functions****:* ***trapezoidal****(x_1, x_2, x_3, x_4) as the membership function $\mu: X \rightarrow [0,1]$ given in (3); we also define* ***left shoulder****(x_2, x_3, x_4) as trapezoidal$(-\infty, x_2, x_3, x_4)$ and* ***right shoulder****(x_1, x_2, x_3) as trapezoidal$(x_1, x_2, x_3, +\infty)$. (see* ***Fig. 1****)*

$$\mu(x) = \begin{cases} 0 & when \quad x < x_1 \\ \frac{x - x_1}{x_2 - x_1} & when \quad x_1 \leq x < x_2 \\ 1 & when \quad x_2 \leq x \leq x_3 \\ \frac{x_4 - x}{x_4 - x_3} & when \quad x_3 < x \leq x_4 \\ 0 & when \quad x_4 < x \end{cases} \tag{3}$$

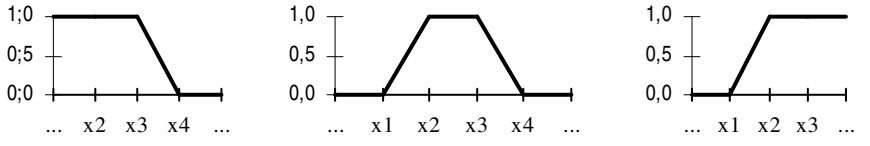

Fig. 1. Linear shape membership functions: From left to right: *left shoulder(x_2, x_3, x_4)*, *trapezoidal(x_1, x_2, x_3, x_4)* and *right shoulder(x_1, x_2, x_3)*

Definition 4: *A* ***linguistic variable*** *is associated to an attribute or concept that is measurable in quantitative terms but it can be described by a set of linguistic terms. A linguistic variable is characterized by a quintuple (X, T(X), U, G, S), where X is the name of the variable considered, T(X) denotes the term-set that can be used as linguistic values for X on the universe of discourse U, G is composed by grammatical rules that generate new terms in T(X) and S is conformed by the semantic rules associating each label L with its meaning S(L) by means of a fuzzy set.*

Definition 5: *A* ***cognition frame*** $\langle U, \mathcal{F}, \preceq \rangle$ *is defined as family of fuzzy sets* $\mathcal{F}$ *defined over the same universe of discourse U. The* ***granularity level*** *K of a cognition frame is the cardinality of* $\mathcal{F}$ *. It results convenient for any fuzzy theory application [15] to establish a total ordering* $\preceq$ *over the frame of cognition. We denote* A_i *to the i-th fuzzy set in the frame* $\mathcal{F}$ *for* $i \leq K$.

When we transform a quantitative variable into a categorical one, its domain is mapped to a cognition frame because presents the meaning of the each labeled class or category, represented by one fuzzy set in the frame, according to some specific context and granularity level. Fig. 2 shows a cognition frame composed by three fuzzy sets, established to represent the "low", "medium" and "high" terms for labeling the variable linguistic *X*. It can be highlighted that, for all cognition frames, $A_i \preceq Aj$, if $i < j \leq k$.

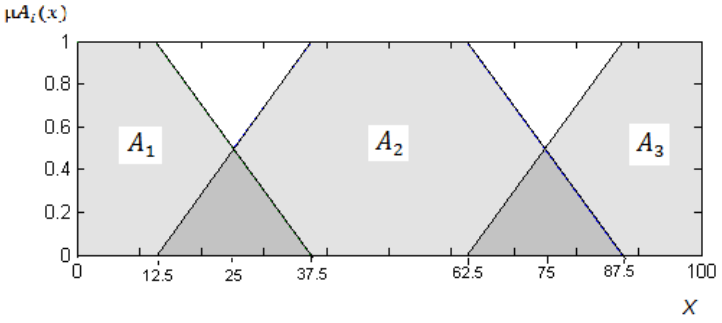

Fig. 2. Cognition Frame with *k* =3, an example

Definition 6: *A frame of cognition is called a* ***fuzzy partition*** *if, and only if:*

$$\forall A \in \mathcal{F} \; A \neq \phi \tag{4}$$

$$\bigcup_{A \in F} A = U \tag{5}$$

The definition of fuzzy partition generalizes classical mathematical partition concept. This kind of partitions differs from classical because fuzzy partition does not demand the exclusivity condition. That is, it does not demand that the intersection between any pair of sets in $\mathcal{F}$ must be empty.

The basic goal of a fuzzy partition is to recognize the differences between groups or classes that can describe in algebraic or graphical form to gain a better understanding of a particular context. In order to make such partition, it is required a representation of all contextual data to continue differentiating one of each *k* classes to be considered in the object categorization.

From data abstraction and knowledge derived from a fuzzy partition, emerges what is called *information granules*, characterized by their interpretability. However, this

cannot be guaranteed until that a set of constraints is imposed to the fuzzy partition. In literature, several constraints have been proposed, although there is no agreement on which constraints should be adopted [11]. We opted by selecting the following restrictions in order to guarantee an appropriate logical structure in the fuzzy partition required for discovering the semantics of vague terms.

Constraint 1: (Proper ordering) *A frame of cognition* $\langle U, \mathcal{F}, \preceq \rangle$ *should be properly ordered. The order of fuzzy sets reflects the order of membership values:*

$$\forall 1 \le i < j \le K : \mu_{A_i}(x) = 1 \wedge \mu_{Aj}(y) = 1 \rightarrow x < y \tag{7}$$

Constraint 2: (Coverage or completeness) *a frame of cognition* $\langle U, \mathcal{F}, \preceq \rangle$ *must be complete, meaning that each element of the universe U belongs at least to one fuzzy set in the frame:*

$$\forall x \in U \; \exists A \in \mathcal{F} \therefore \mu_A(x) > 0 \tag{8}$$

Constraint 3: (Complementarity or Σ-criterion) *for each element of the Universe U, all membership values of the frame must sum up one:*

$$\forall x \in U : \sum_{A \in F} \mu_A(x) = 1 \tag{9}$$

Constraint 4: (Distinguishable granules) *any fuzzy set in the frame of cognition must be well distinguishable from the remaining fuzzy sets.*

Constraint 5: (Justifiable number) *In order to guarantee meaningful interpretations, the number of fuzzy sets in a frame should not be too high, preferably less than 7.*

Constraint 6: (No more than two) *for any element in U, no more than two fuzzy sets in* $\mathcal{F}$ *could give a nonzero membership degree:*

$$\forall x \in U : |\{A \in \mathcal{F} : \mu_A(x) > 0\}| \le 2 \tag{10}$$

Constraint 7: (Full acceptance) *whenever a membership value decreases from 1, the other value increases from zero:*

$$\forall x \in U \;\; \forall A \in \mathcal{F} \; \mu_A(x) < 1 \rightarrow \exists B \in \mathcal{F}, B \neq A, \mu_B(x) > 0 \tag{11}$$

Constraint 8: (Consistent partition) *a frame of cognition* $\langle U, \mathcal{F}, \preceq \rangle$ *is a consistent partition if* $\exists x \in U$ *such that:*

$$\forall A \in \mathcal{F}(\mu_A(x) = 1 \rightarrow \forall B \in \mathcal{F}, B \neq A, \mu_A(x) = 0) \tag{12}$$

Constraint 9: (Normality) *a fuzzy set A should be normal, i.e. exists at least one element (called prototype) with full membership:*

$$\forall A \in \mathcal{F} \quad \exists x \in U : \mu_A(x) = 1 \tag{13}$$

Constraint 10: (Convexity) *a fuzzy set A should be convex i.e. the membership values of the elements belonging to any interval are not lower than the membership values at the interval's extremes:*

$$\forall a, b, x \in U : a \le x \le b \rightarrow \mu_A(x) \ge \min\{\mu_A(a), \mu_A(b)\} \tag{14}$$

Given previous constraints, overlapping is allowed only for adjacent classes. Additionally, any element in the overlapping area has complementary membership degrees for both overlapping classes. The intersection cross point is at the membership degree level of 0.5.

Classical Statistics has assumed that most data distributions can be adjusted to the Gaussian or normal probabilistic model. A partition of the universe of discourse based on the normal distribution has the advantage of relying on only two parameters: the mean and the variance. But, when there are biases or asymmetries in data distribution trying to represent all data with only two parameters is insufficient. Therefore, when the data distribution does not present a behavior that can be considered normal, have been proposed the so-called *non-parametric* models. This term does not mean that such models have no parameters, but the number and nature of these can be flexible and not fixed in advance. It means that the data distribution need not be defined a priori because the available data are used to determine it. This is the reason why this type of models is also called *distribution-free* models. Within this category, the most commonly models used to describe the distribution of a data collection are the frequency histograms, the cumulative frequency polygons, the kernel density estimation and percentile-based models [12].

Choosing an appropriate number of percentiles, it is obtained a model representing the shape and location of the data distribution, better than the others mentioned before, since they can also represent the biases or asymmetries in the distributions and have a limited number of parameters. In addition, a percentile-based model assures the fulfillment of convexity constraint (10) for the fuzzy sets derived from it, which is not the case with the frequency histograms or the kernel density estimations.

Particularly, a very effective way to describe a collection of statistical data is *the 5-number summary* composed by the minimum value, maximum value and the three quartiles of the distribution of the sorted data. The quartiles Q_i divide the density distribution of the sorted data into four areas, each with 25% of the data. From this summary statistics, it is usually for building the box and whiskers plot (or box plot), one of the most informative graphical models used to represent collections of data [8]. Fig. 3 shows an example of a box plot.

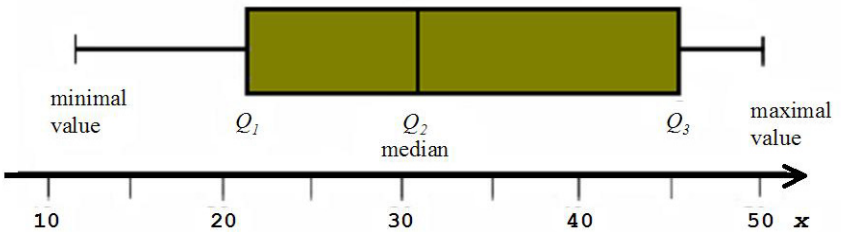

Fig. 3. Box and Whisker plot example

By examining the length of the whiskers compared to the box length, we can see whether the dataset has light, normal or heavy tails. Comparing the lengths of the whiskers it can be determined whether the distribution of the data is skewed or symmetric. With the aim of generalization, the next definition is given.

Definition 6*: Formally, a* ***percentile*** P_q *is a point in the domain of a quantitative variable, under which there is a percentage q of the values of a sorted dataset.*

3 Categorization Specification

SQLf [2] is a fuzzy query language for relational databases that allows the use of fuzzy condition anywhere SQL allows a Boolean one. SQLf definition was updated till standard SQL:2003 [5] allowing fuzziness in: nesting, grouping, algebra operators, integrity constraints, views, data manipulation operations, deductive rules, constraints, assertions rules, functions, procedures, triggers, objects, multiset data type, enhancements to SQL-invoked routines, table creation and the merge statement.

A fuzzy condition may involve different linguistic terms. SQLf enables users to define their own terms. User-defined linguistic terms allowed, in SQLf language, are fuzzy predicates, modifiers, comparators, connectors and quantifiers. All they have a semantics model according to fuzzy set theory and the corresponding syntax in SQLf. For example, the syntax for specifying a fuzzy predicate identified by a *name* and defined by the membership function *trapezoidal*(x_1,x_2,x_3,x_4) in the universe of quantitative *domain*, the SQLf sentence would be as the following:

```
CREATE FUZZY PREDICATE name ON domain AS (x1,x2,x3,x4)
```

We propose here to extend SQLf for allowing definition and use of context dependent linguistic terms. Semantic for such terms would be based on the concept of fuzzy partition. We will name such terms with *labels* to conform a cognition frame. For this, proposed SQLf sentence is as follows:

```
CREATE FUZZY CATEGORIZATION label1, label2, … , labelK ON
attribute1,… ,attributen AS CONTEXT DEPENDENT
```

In this data definition statement, $label_1$, $label_2$, … , $label_K$ are identifiers for the linguistic terms specified in the categorization. These are intended to be respectively, interpreted by fuzzy predicates A_1, A_2, … , A_K in a fuzzy partition $\langle U, \mathcal{F}, \preceq \rangle$ of granularity K. Universe U would be determined by the context at querying time. Each $attribute_j$ is the object identifier of a column attribute, a designator including the table identification followed by a dot and the column name in such table. The specification $attribute_1$,… , $attribute_n$ tells that $label_1$, $label_2$, … , $label_K$ are valid fuzzy predicates for $attribute_1$,… ,$attribute_n$. The final clause `AS CONTEXT DEPENDENT` remarks the fact that actual model of each label would be inferred in according to context at querying time. We limit the number of label in a categorization to be between 2 and 6. General theoretical models for categorization are given in Table 1 These models are based on statistical concept of percentile.

In order to illustrate the proposed definition of fuzzy categorization, let us consider the tables `Department` and `Employee` with the following schemes:

```
Department(depID, name, locality, headID, budget)

Employee(empID, name, birth, studyLevel, salary, depID)
```

Suppose we want to perform fuzzy queries on these tables considering different levels of budget for departments as well as salaries and study levels of employees. For considered attributes we would like to consider just three categories, namely low, middle and high classes. So, we may specify it in our extension of SQLf by the sentence:

```
CREATE FUZZY CATEGORIZATION low, middle, high ON
Employee.studyLevel, Employee.salary, Department.budget
AS CONTEXT DEPENDENT
```

Remark that the same labels could apply to different attributes. Semantics of each label would be adjusted to the context of corresponding attribute values at query time. With previous definition, we might address SQLf query:

```
SELECT d.depID, d.budget, e.empID, e.salary
FROM Department AS d, Employee AS e
WHERE d.locality='Medellin' AND d.depID=e.depID AND
d.budget=low AND e.salary=high AND e.studyLevel=low
```

As categorization semantics is context dependent and it would be determined just at query time, we allow online definition of categorizations. We propose three variants for these specifications. First is as follows:

```
WITH FUZZY CATEGORIZATION label1, label2, … , labelK
query
```

This is a categorization defined just for performing the specific query. The semantics of this statement is similar to that of `CREATE FUZZY CATEGORIZATION` but labels are not stored in the database catalogue. They are just used in the query. The

clause ON does not appear, labels could be used with any quantitative attribute in the query. For example, we might address the query

```
WITH FUZZY CATEGORIZATION low, middle, high
SELECT d.depID, d.budget, e.empID, e.salary
FROM Department AS d, Employee AS e
WHERE d.locality='Medellin' AND d.depID=e.depID AND
d.budget=low AND e.salary=high AND e.studyLevel=low
```

Second variant is to define just a label but telling the querying system the ordering position *i* of such label in the cognition frame as well as its granularity *K*.

```
WITH FUZZY LABEL label AS i IN CATEGORIZATION OF K
query
```

One example of this kind of fuzzy queries would be the following:

```
WITH FUZZY LABEL low AS 1 IN CATEGORIZATION OF 3
SELECT * FROM cars WHERE trademark = 'Ford' AND hp=low
```

Third and last variant of online labels categorization definition is to the label just when it is used in the query condition. In this case, the scope is local for each single condition. We must tell the querying system the ordering position *i* of such label in the cognition frame as well as its granularity *K*. The syntax of such conditions is:

```
Attribute e.*, d.* = label AS i IN CATEGORIZATION OF k
```

Using this variant, we might address the query

```
SELECT d.depID, d.budget, e.empID, e.salary
FROM Department AS d, Employee AS e
WHERE d.locality='Medellin' AND d.depID=e.depID AND
d.budget=low AS 1 IN CATEGORIZATION OF 3 AND
e.salary=high AS 4 IN CATEGORIZATION OF 4 AND
e.studyLevel=low AS 1 IN CATEGORIZATION OF 5
```

In this example, despite the same label is used for different attributes, each use conform a different cognition frame according to their own granularity level.

4 Vagueness Interpretation

Context-dependent vagueness interpretation requires some additional tasks than usual translation because of the inductive process required for discovering the fuzzy set models. After the lexical and syntactical analysis, that allows identifying the pattern that the query matches, begins the semantic analysis of the statement. It involves three important tasks: context delimitation, vagueness inference and evaluation.

In context delimitation, the linguistic context that specifies the current vague query must be derived from the base table specification in query. The context would be the relation defined operationally by the Cartesian product of tables in the `FROM` clause filtered by crisp conditions in `WHERE` clause (if any).

In previous section's first example, context is delimited by the table specification:

```
FROM Department AS d, Employee AS e
WHERE d.locality='Medellin' AND d.depID=e.depID
```

In this case, on one hand, the label *low* in the expression `d.budget=`*low* will be interpreted in the context of all values for `budget` attribute in current instance of `Department` table restricted to those rows where `locality` attribute has the value `'Medellin'`. On the other hand, the same label *low* but in expression `e.studyLevel=`*low* will be interpreted in a different context. We restrict the current instance of `Employee` relation to those rows where `depID` attribute matches with those of `Department` table where actual value of `locality` attribute is `'Medellin'`. The context defined for restringing the employee relation by the expression `e.studylevel =` *low*, will be conformed by all current tuples who meet this condition in the derived `Employee` (employees from Medellin). In similar way, we obtain the context for label *high* in expression `e.salary =` *high*.

After the context is delimited, the vagueness inference process begins. In this process the vague terms semantics, in terms of fuzzy sets, will be estimated. The parameter estimators of those fuzzy membership functions are some percentiles values extracted from the contextual data.

The core of the whole process of interpretation is the semantic analysis where it can be found the meaning of labels used to restrict some database objects. From this process, a fuzzy partition of the contextual data is obtained.

The fuzzy partition determines a particular model to each linguistic label according to the context and granularity level considered in the object categorization Since the context is only determined when the query is specified, the fuzzy partition can not be done earlier. Depending on these variables, the inference machine discovers a concrete meaning for a label attached to a linguistic variable.

The semantic rules defined for finding the context dependent meaning of a label appears in Table 1. Different granularity levels in the fuzzy partition are considered. Fuzzy sets models defining labels are with linear shape membership functions: trapezoidal, left shoulder and right shoulder.

The automatic generation of information granules from available data gives to inference machine the capability to perform data mining tasks for acquiring knowledge that can be transferred to users. To do that, user previously must specify the categorization and query engine must implement mechanisms for the inference of parameters in theoretical models according to context.

Table 1. Theoretical models for the interpretation of context-dependent vague terms. First column *K* is the granularity of the categorization. Second one, *i* is the order of the label in the categorization, Third column determines the semantics for the *i-th* label. Each P_q is the *q*-th percentile for the attribute values in the context.

K	*I*	**Membership function**
2	1	left shoulder(P_0, $P_{37.5}$, $P_{62.5}$)
	2	right shoulder ($P_{37.5}$, $P_{62.5}$, P_{100})
3	1	left shoulder (P_0, $P_{12.5}$, $P_{37.5}$)
	2	trapezoidal($P_{12.5}$, $P_{37.5}$, $P_{62.5}$, $P_{87.5}$)
	3	right shoulder ($P_{62.5}$, $P_{87.5}$, P_{100})
4	1	left shoulder (P_0, $P_{15.6}$, $P_{28.1}$)
	2	trapezoidal($P_{15.6}$, $P_{28.1}$, $P_{43.8}$, $P_{56.3}$)
	3	trapezoidal($P_{43.8}$, $P_{56.3}$, $P_{71.9}$, $P_{84.4}$)
	4	right shoulder ($P_{71.9}$, $P_{84.4}$, P_{100})
5	1	left shoulder (P_0, P_5, P_{15})
	2	trapezoidal (P_5, P_{15}, P_{30}, P_{40})
	3	trapezoidal (P_{30}, P_{40}, P_{60}, P_{70})
	4	trapezoidal (P_{60}, P_{70}, P_{85}, P_{95})
	5	right shoulder (P_{85}, P_{95}, P_{100})
6	1	left shoulder(P_0, P_{10}, P_{18})
	2	trapezoidal(P_{10}, P_{18}, P_{28}, P_{36})
	3	trapezoidal(P_{28}, P_{36}, P_{46}, P_{54})
	4	trapezoidal(P_{46}, P_{54}, P_{64}, P_{72})
	5	trapezoidal(P_{64}, P_{72}, P_{82}, P_{90})
	6	right shoulder(P_{82}, P_{90}, P_{100})

5 Complex Predicates

Disjunction and conjunction are usually interpreted by *min* and *max* respectively. This is because the minimal degree of membership is a continuous triangular norm, and the

maximal value is a continuous triangular co-norm defined in the standard fuzzy theory [14]. However, *max* function does not generate convex fuzzy sets and this constraint has to be met in order to obtain interpretable granules and to preserve the concept of universal set. Therefore, in this work, it is proposed to represent the OR connective by the sum of membership degrees corresponding to each concatenated simple condition when the specified classes come from the same frame of cognition. In this case, sum operator is also a triangular co-norm because of complementarity constraint. Fig. 4 evidences the convenience of using *sum* rather than *max* function.

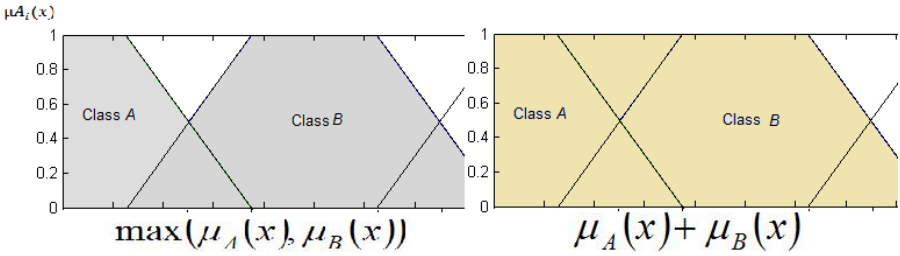

Fig. 4. Disjunction operators for fuzzy sets. Sum operator (right side), and *max* function (left side)

In multidimensional case, when the aggregation is conformed by simple restrictions over different domains, the sum is no longer an appropriate operator. Consequently, the average of different membership degrees of simple conditions is adopted for representing the disjunction. That is, a LOWA (Linguistic Ordered Weighted Averaging) operator [6] used to evaluate the global membership function of the tuple t_i to the disjunction of p labeled classes Li is given by:

$$\mu L_1(ti)\ OR\ \mu L_2(ti)...OR\ \mu Lp(ti) = \sum_{i=1}^{p} \frac{\mu Li(ti)}{p} \forall ti \in R \tag{15}$$

In flexible querying systems it is also likely required to know the compatibility degree of objects with some linguistic label defined by a linear combination of restrictions. In this linear combination each restriction may have a different weight or importance indicator β_i considered during the derived term translation. These weightings must sum up the unit (equivalent to one hundred percent) in order to be a LWA (Linguistic Weighted Averaging) operator as a generalization of the LOWA operator [6].

$$\sum_{i=1}^{p} \beta_i = 1,\ \beta_i \in [0,1] \forall i = 1,..., p \tag{16}$$

To determine the global membership degree of a tuple t_i in relation R with the expression derived from a linear combination of restrictions, a weighted average membership over individual labeled classes is computed. That is, the model that determines the degree of belonging to the linear combination is:

$$\mu expression(ti) = \sum_{i=1}^{p} \beta i \mu conditioni(ti) \; \forall ti \in R, \beta i \in [0,1] \wedge \sum_{i=1}^{p} \beta i = 1 \tag{17}$$

The extended syntax of the query to include aggregates originated by a linear combination of simple conditions, vague or specific, includes a new operators * and +. With these operators, we propose this new syntax for fuzzy conditions:

$\beta 1 * condition\ 1 + \beta 2 * condition\ 2\ [+...]$

To show an example of a query that follows this pattern, the following statement shows the attractive cars, defined as a linear combination of *high* miles per gallon (mpg) , *high* horsepower (hp) and *low* weight, considering the weights 0.4, 0.4 and 0.2, respectively.

```
WITH FUZZY CATEGORIZATION low, middle, high
SELECT name FROM cars WHERE
0.4*(mpg=high) + 0.4*(hp=high) + 0.2*(weight=low)
```

6 Experimental Results

Once the theoretical models for fuzzy sets were defined for each label in a cognition frame, experimental tests were performed with real benchmark databases for the evaluation of the fulfillment of all the restrictions imposed to have a proper logical structure of reasoning and to see how the context determines the meaning of labels. The first database is called Pima Indians Diabetes Database originally owned by National Institute of Diabetes and Digestive and Kidney Diseases, available at [7]. This database has 768 women records: 500 tested negative and 268 tested positives. The number of attributes registered in the database is eight, but to illustrate our proposal for finding the meaning of context-dependent fuzzy terms, we present the fuzzy set models only for two attributes: the number of times pregnant and plasma glucose concentration at two hours in an oral glucose tolerance test.

Because of the precedence of the reference database, it can be suspected that both chosen variables as examples, could be different depending if the person tested positive or negative to diabetes test. So, we decided to find that could be a *low* or *high* number of times pregnant, considering the test results and only these two categories. According to the formulas proposed in Table 1, for $K = 2$, we obtained the fuzzy set membership functions in Table 2 and Table 3, to represent the variables for both positive and negative tested. Fig. 5 and Fig. 6 graphically illustrate these models.

Table 2. Models of context-dependent vague terms for the number of times pregnant in Diabetes Database

Context	Left shoulder	p0	p0	p37.5	p62.5	Right shoulder	p37.5	p625	p100	p100
Tested Negative (n=500)	Low	0	0	2	4	High	2	4	13	13
Tested Positive (n=268)	Low	0	0	3	6	High	3	6	17	17

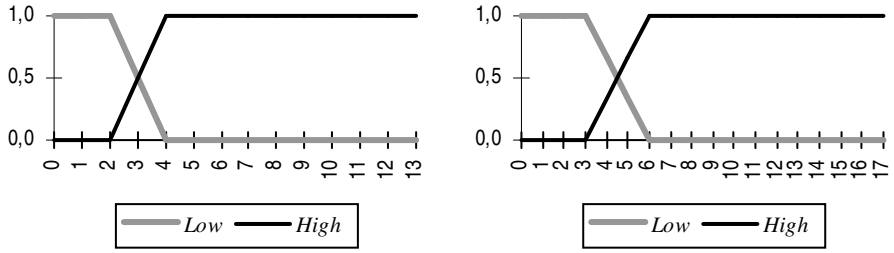

Fig. 5. Cognition frames for number of time pregnant in Diabetes Database. Left: Models for context "Tested Negative". Right: Models for context "Tested Negative".

Table 3. Models of context-dependent vague terms for plasma glucose concentration in Diabetes Database

Context	Left shoulder	p0	p0	p37.5	p62.5	Right shoulder	p37.5	p625	p100	p100
Tested Positive (n=268)	Low	0	0	129	152	High	129	152	199	199
Tested Negative (n=500)	Low	0	0	100	115.9	High	100	115.9	197	197

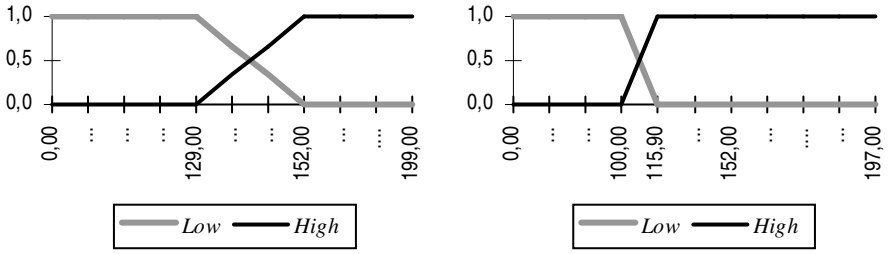

Fig. 6. Cognition frames for glucose concentration in Diabetes Database. Left: Models for context "Tested Negative". Right: Models for context "Tested Negative".

The other database chosen as reference contains information about some features of 398 cars also available in [7]. Some attributes of the cars in the dataset are trade mark, acceleration limit and hp (motor horsepower).

We created a table called `cars` and populate it with the dataset. For the experimentation, we defined the following categorization:

```
CREATE FUZZY CATEGORIZATION low, middle, high ON
cars.hp, cars.acceleration AS CONTEXT DEPENDENT
```

Using previous categorization specification described in the table, we have can formulate the following four queries.

```
SELECT * FROM cars WHERE hp =low
SELECT * FROM cars WHERE trademark = 'Ford' AND hp =low
SELECT * FROM cars WHERE trademark = 'Chevrolet' AND hp
= low
SELECT * FROM cars WHERE acceleration > 16 AND hp = low
```

The difference in previous queries is the context. We can observe that despite we have defined linguistic labels for acceleration attribute; we do not use them in the queries. The fourth query uses this attribute but with a crisp condition, delimiting the context. The semantics of the linguistic label low is inferred for each query, obtaining the membership functions shown in Fig. 7.

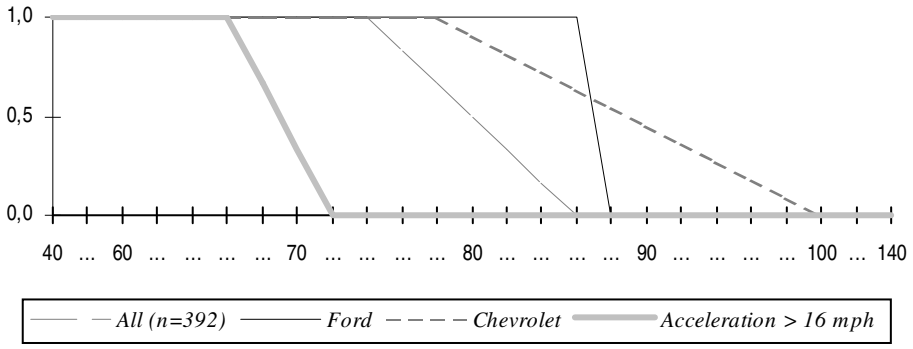

Fig. 7. Membership functions inferred for fuzzy label low on hp (horsepower) attribute of cars table, considering different contexts

The proposed interpretation for categorization (Table 1) meets all constraints Defined for a fuzzy partition. It may be formally proved in a very strike way. We put this in evidence in the experimentation with the dataset. First, we defined a categorization considering two classes for hp attribute and taking as context the whole dataset. We repeat the experience with three classes or categories.

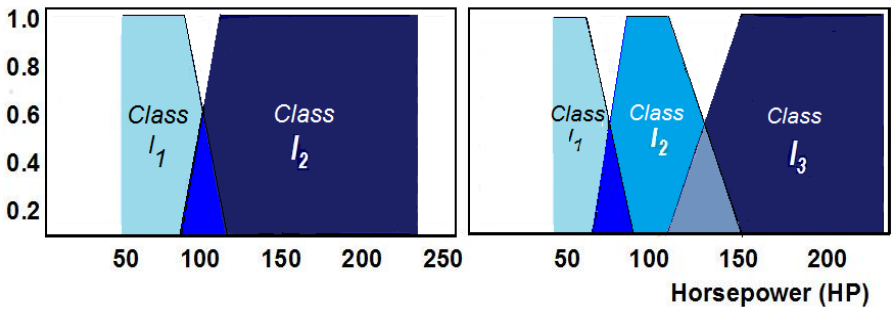

Fig. 8. Two inferred fuzzy partitions for cars horsepower

Fig. 8 shows the two different cognition frames to which the system gets adjusted in this experimentation. It is easy to see in these examples two consistent partitions and, any pair of membership degrees for each value in the domain always sums up one.

7 Conclusions and Future Works

In this work, we present an extension of SQLf language with the aim the user can formulate queries with linguistic labels whose meaning depends on current data context. The possibility to interpret vague queries in multiple ways, depending on their linguistic context, as proposed here, allows the constructions of more flexible and reliable query-answering systems. This proposal allows different levels of granularity in the categorization required for the interpretation of such terms, looking for more flexibility. This feature is provided in the sense of allowing online linguistic labels definition at querying time with different variants. Semantics of our extension has been formally defined based in a statistical nonparametric model, fuzzy set theory and the constraints defining well formed fuzzy partitions. Proposed syntax follows the style of standard SQL and SQLf extension. In our extension we provide also new ways of combining fuzzy conditions. The OR operator is overloaded for being interpreted as the sum when combined conditions are in the same cognition frame. These new conditions may be combined with linear combinations. In this way, we have made a contribution to vague querying that is a field with rising interest at present time and promising success in near future. We have focused our extension in SQLf due to great development of this language. Nevertheless, concepts and techniques presented here may be adopted in any existing fuzzy query language. In future works we would present corresponding extensions to other linguistic terms, such as modifiers, comparators, qualifiers and quantifiers, in order to be adjusted to data context. We think it would be possible to store some statistics in system metadata in order to keep advantage of them for meaning inferences, with the intention to improve the performance of the query systems. It could also be subject of future work.

Acknowledgments. This work is done for the glory of God. “Now this is our boast: Our conscience testifies that we have conducted ourselves in the world, and especially in our relations with you, with integrity and godly sincerity. We have done so, relying not on worldly wisdom but on God’s grace. For we do not write you anything you cannot read or understand. And I hope that, as you have understood us in part, you will come to understand fully that you can boast of us just as we will boast of you in the day of the Lord Jesus.” (2 Corinthians 1:12-14, New International Version NIV).

References

1. Bordogna, G., Psaila, G.: Customizable flexible querying for classical relational databases. In: Galindo, J. (ed.) Handbook of Research on Fuzzy Information Processing in Databases, pp. 191–217. Idea Group Inc., IGI (2008)

2. Bosc, P., Pivert, O.: SQLF: A relational database language for fuzzy querying. IEEE Transactions on Fuzzy Systems 3(1), 1–17 (1995)
3. Galindo, J., Medina, J.M., Pons, O., Cubero, J.C.: A Server for Fuzzy SQL Queries. In: Andreasen, T., Christiansen, H., Larsen, H.L. (eds.) FQAS 1998. LNCS (LNAI), vol. 1495, pp. 164–174. Springer, Heidelberg (1998)
4. Galindo, J.: Introduction and trends in fuzzy logic and fuzzy databases. In: Galindo, J. (ed.) Handbook of Research on Fuzzy Information Processing in Databases. Idea Group Inc., IGI (2008)
5. Goncalves, M., González, C., Tineo, L.: A New Upgrade to SQLf: Towards a Standard in Fuzzy Databases. In: Proc. of DEXA 2009 Workshops (2009)
6. Herrera, F., Herrera-Viedma, E.: Aggregation operators for linguistic weighted information. IEEE Trans. on Systems, Man and Cibernetics 27, 268–290 (1997)
7. Information Technology Laboratory (ITL), Statistical reference data sets archives, http://www.itl.nist.gov/div898/strd/general/dataarchive.html
8. Johnson, R., Wichern, D.: Applied Multivariate Statistical Analysis. Pearson (2008)
9. Kackprzyk, J., Zadrozny, S.: Computing with words in intelligent database querying: standalone and Internet-based applications. Inform. Sciences 134, 71–109 (2001)
10. Ma, Z.M., Yan, L.: A Literature Overview of Fuzzy Conceptual Data Modeling. Journal of Information Science and Engineering 26(2), 427–441 (2010)
11. Mencar, C.: Theory of Fuzzy Information Granulation: Contributions to Interpretability Issues. Doctoral Thesis. Universidad de Bari. Italia (2004), http://www.di.uniba.it/~mencar/download/research/tesi_mencar.pdf
12. Neter, J., Wasserman, W.: Applied Linear Regression Analysis. Wiley, New York (2001)
13. Pivert, O., Bosc, P.: Fuzzy Preference Queries to Relational Databases. Imperial College Press (2012)
14. Trillas, E., Alsina, C., Pradera, A.: On a class of fuzzy set theories. In: IEEE, Fuzzy Systems Conference International (2007), http://ieeexplore.ieee.org
15. Tudorie, C.: Qualifying objects in classical relational databases. In: Galindo, J. (ed.) Handbook of Research on Fuzzy Information Processing in Databases, pp. 218–249. Idea Group Inc., IGI (2008)
16. Yager, R.: Soft Querying of Standard and Uncertain Databases. IEEE Transactions on Fuzzy Systems 18(2), 336–347 (2010)
17. Zadeh, L.A.: The concept of linguistic variable and its application to approximate reasoning. Information Science 8(3), 199–249 (1975)
18. Zhao, F., Ma, Z.M.: Vague Query Based on Vague Relational Model. In: Yu, W., Sanchez, E.N. (eds.) Advances in Computational Intelligence. AISC, vol. 61, pp. 229–238. Springer, Heidelberg (2009)