RELATIVE AGGREGATION OPERATOR IN DATABASE FUZZY QUERYING

Cornelia TUDORIE, Severin BUMBARU, Luminita DUMITRIU

Department of Computer Science, University "Dunarea de Jos", Galati
Domneasca 111, 800008 Galati, Tel, Fax: 460182
email: Cornelia.Tudorie@ugal.ro, Severin.Bumbaru@ugal.ro, Luminita.Dumitriu@ugal.ro

Abstract: Fuzzy selection criteria querying relational databases include vague terms; they usually refer linguistic values form the attribute linguistic domains, defined as fuzzy sets. Generally, when a vague query is processed, the definitions of vague terms must already exist in a knowledge base. But there are also cases when vague terms must be dynamically defined, when a particular operation is used to aggregate simple criteria in a complex selection. The paper presents a new aggregation operator and the corresponding algorithm to evaluate the fuzzy query.

Keywords: Artificial Intelligence, Database, Fuzzy Logic, Fuzzy Queries

1. INTRODUCTION

The database querying access is usually limited by two main reasons:

- the rigid formal language syntax and
- the difficulty to realize and express precise criteria to locate the information

So, it is very useful to provide intelligent interfaces to databases, able to understand natural language queries and more important, able to interpret and evaluate imprecise criteria in queries.
Including vague criteria in a database query may have two advantages:

- the flexibility of the query expression
- the possibility to refine the results, assigning to each tuple the corresponding degree of criteria satisfaction.

We particularly focus on the possible vagueness of the selection criterion, which involve certain vague terms, currently used in natural language speaking. So, all the discussion in that direction is inspired from the usual necessities of our final database users, expressed in many various linguistic forms[1].

In a vague query, the selection criterion is no longer Boolean, so it can be more or less satisfied by the database tuples. Therefore, for each table row a satisfaction degree is estimated, which stands for a measure of its compatibility with the vague criterion.

Table 1

COMPANY

Name	. . .	Capital	Year	. . .
CONTRAST		7	2004	
GENERAL		130	1990	
ROLSIM		54	2000	
BAZAR		115	1997	
TIBCO		147	1992	
ELADA		120	1999	
LILITEX		130	1999	
EXPRESS		125	1996	
NARCIS		68	2001	
FLORIO		110	1999	
MINOTAUR		75	1994	

Example 1. If the Table 1, containing hypothetical companies, is considered, the response to the query

*Retrieve the **very big** companies, and **almost new***

[1] The study is made for a Romanian language interface for database fuzzy querying; the examples are translated.

This paper was recommended for publication by Sergiu CARAMAN

(*Care sunt firmele **relativ recente*** cu capital ***foarte mare**?)*

can be:

Table 2

Name	**. . .**	**Capital**	**Year**	**. . .**	**μ**
LILITEX		130	1999		1
ELADA		120	1999		0.3
BAZAR		115	1997		0.1

where the μ coefficient stands for the satisfaction degree of the query selection criterion for each database tuple.

The fuzzy set theory is already established as the adequate framework to model and to manage vague expressions, or in other words, to evaluate vague queries sent to relational databases (BADINS, 1995; BADINS, 1997; Bosc, *et al.*, 1993; Dubois and Prade, 1996; Pivert, 1991; Tudorie, 2003b).

The selection vague criterion may be very simple, but it may also be very complex. We can consider both the linguistic complexity, but the logical one too. The linguistic complexity of the criterion is coming from various categories of vague terms, and very often it has different semantic effects on the selection criterion, hence the logical complexity.
A review of several categories of linguistically terms with vague meaning, their fuzzy model and specific operations are presented in (Tudorie, 2003a), (Tudorie, 2003b), and many others.

Usually, when a vague query is processed, the definitions of possible vague terms (as fuzzy sets) must already exist in a knowledge base and the object qualification depends on these ones. But, depending on the query's complexity, there are cases when the fuzzy model must be dynamically provided and adapted to the particular context created by other vague criteria in the same.
For example, to compare the classical query:

*Retrieve the **new** companies*
(*Care sunt firmele **recente** ?*)

or even this one:

*Retrieve the **new** companies* **(and)** *having **big** capital*
(*Care sunt firmele **recente** cu capital **mare**?*)

with the following one:

*Retrieve the **new** companies* **within** *the **big*** ones
(*Care sunt firmele **recente** dintre cele cu capital **mare**?*)

The paper will demonstrate also that the evaluation procedure for the above queries are not the same.
The last one is a special type of query that will be discussed in the paper. A new operator, a not commutative one, for criteria aggregation in a selection query will be defined.

2. CRISP AND LINGUISTIC DOMAIN FOR THE DATABASE ATTRIBUTES

Definition 1. The **linguistic label** is a word (usually coming from natural language) that designates a fuzzy entity.
A linguistic label may be assigned to a fuzzy set or fuzzy quantity, suggesting a vague term from usual language, typical to the application area where our model is working. The linguistic label may have various meaning, for example:

- a *gradual property* for an object (‚***good*** **mark**' for a student), when the membership degree of the fuzzy set is a function defined on an crisp attribute domain, generally a monotonous function; the value of the membership function expresses the *intensity of the property*: between the 0 degree (that is a not good mark), and the 1 degree (that is an absolutely good mark).
- a *qualification* for an object, that identifies a *category of objects* (‚***intelligent*** **student**'), when the membership function is not always monotonous and its value expresses the *closeness degree* of the current object to the object considered as *the representative one for that category*.

If more fuzzy sets are defined on the same referential domain, with different membership functions, then a ***fuzzy set family*** is formed. If a linguistic label is assigned to each fuzzy set, then the set of these labels may be the definition set of a ***linguistic variable***, and the labels are named ***linguistic values***.

Definition 2. The **linguistic variable** is a quadruple:

$(V, E(V), U, M)$ where

V is the name of the linguistic variable
$E(V)$ is a set of linguistic values for the linguistic variable V
U is the crisp referential domain of the linguistic variable V
M is a mapping $E(V) \rightarrow \mathcal{F}(U)$ that maps a fuzzy set on U for each linguistic values of V.

The choice of numerical representation of a linguistic term is seldom obvious. However, at the qualitative level, the term is well understood and well semantically placed with respect to other linguistic expressions. Thus, an order relation $\underline{\prec}$ on $E(V)$ is easy to define; for example:

little $\underline{\prec}$ *intermediate* $\underline{\prec}$ big

Obviously, $\underline{\prec}$ is a semantic order relation.

In most cases, the set of the linguistic values for a linguistic variable

$$E(V) = \{ e_i \}, i = 1,..., n_e$$

is an ordered set ($e_i \prec e_j$ for $i \leq j$) having an odd cardinality (like in figure 1). The intuitive order relation $\prec$ between linguistic values is the correspondent at the semantic level of a pre-order relation $\blacktriangleleft$, established between fuzzy sets defining the linguistic values (this relation is defined by Ulrich Bodenhofer in (Bodenhofer, 2000) and its ***interpretability*** is demonstrated in (Bodenhofer and Bauer, 2000)).

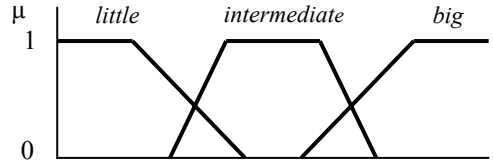

Fig. 1. Definitions of a set of linguistic values

It is easy to show that the relation at the linguistic level:

little $\prec$ *intermediate* $\prec$ big

is transferred at the numerical level:

$M(little) \blacktriangleleft M(intermediate) \blacktriangleleft M(big)$

Definition 3. Let be V a linguistic variable defined on the domain D of the table attribute A. The linguistic values of the V variable form the **linguistic domain** of the A attribute.

So, in a vague query context, the ***crisp domain*** (the domain attribute, according to the relational model theory) and the ***linguistic domain*** must be defined for each attribute.

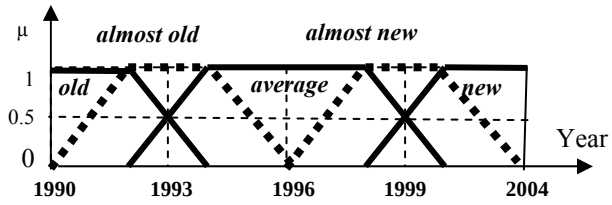

Fig. 2. The linguistic domain of the **Year** attribute

For example,
the crisp domain D=[1990, 2004] and
the linguistic domain
L={ ***old, almost old, average, almost new, new*** }
(L={ ***veche, destul de veche, nici veche nici nouă, destul de recentă, recentă*** })
might be associated to the attribute **Year** of the COMPANY table. They are drawn in Fig. 2.

3. A METHOD TO DISCOVER THE DEFINITION OF LINGUISTIC VALUES

In most applications, defining the linguistic values set of a linguistic variable covers almost uniformly the referential domain (figure 3). There are usually 3, or 5 or another odd number of linguistic values. This is the reason why a method for automatic discovering of the linguistic values definitions for a database attribute can be implemented (**FuzzyKAA** System, (Tudorie, 2003c)).

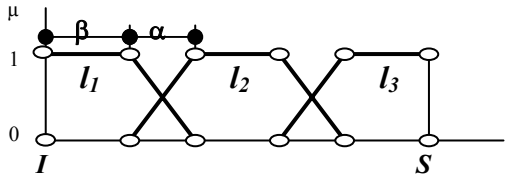

Fig. 3. A set of linguistic values on an attribute domain

The algorithm to obtain the definition for three linguistic values l_1, l_2, l_3 on a database attribute starts from the predefined values α and β, and the attribute crisp domain limits, *I* and *S*; these ones are coming from the database content.

$$M(l_1) = \begin{cases} 1 & , \quad I \leq t \leq I + \beta \\ 1 - \dfrac{t - (I + \beta)}{\alpha} & , \quad I + \beta \leq t \leq I + \beta + \alpha \\ 0 & , \quad t \geq I + \beta + \alpha \end{cases}$$

$$M(l_2) = \begin{cases} 0 & , \quad I \leq t \leq I + \beta \\ 1 - \dfrac{I + \beta + \alpha - t}{\alpha} & , \quad I + \beta \leq t \leq I + \beta + \alpha \\ 1 & , \quad I + \beta + \alpha \leq t \leq I + 2\beta + \alpha \\ 1 - \dfrac{t - (I + 2\beta + \alpha)}{\alpha} & , \quad I + 2\beta + \alpha \leq t \leq I + 2\beta + 2\alpha \\ 0 & , \quad t \geq I + 2\beta + 2\alpha \end{cases}$$

$$M(l_3) = \begin{cases} 0 & , \quad It \leq I + 2\beta + \alpha \\ 1 - \dfrac{I + 2\beta + 2\alpha - t}{\alpha} & , \quad I + 2\beta + \alpha \leq t \leq I + 2\beta + 2\alpha \\ 1 & , \quad t \geq I + 2\beta + 2\alpha \end{cases}$$

Using this method can have a great advantage: details regarding effective attribute domain limits, or distributions of the values, can be easy obtained thanks to directly connecting to the database.
The method can be very useful to develop the initial knowledge base, containing fuzzy definitions of vague terms in the application domain, but even later, to maintain the actuality of these definitions with the instantly database content.

Moreover, the method used by the FuzzyKAA System discovers the definitions of the linguistic values, distributed on the complete database attribute domain. But the same method can be used to dynamically define the linguistic values on attribute subdomains, depending on the context created by the

query selection criteria. We will be seeing more in the section V of the paper.

4. FUZZY CRITERIA AGGREGATION. THE TYPICAL FUZZY *AND*

In a classical (precise) query, the compound selection criterion is a logical expression containing comparisons and logical operators. In a vague context, the classic logical operators **AND**, **OR**, **NOT**, are extended to fuzzy aggregation connectives. They are able to compute a global satisfaction degree, starting from the satisfaction degrees of each vague selection criterion with respect a certain model of the fuzzy connectives.
Usually, the minimum and maximum functions stand for the fuzzy conjunctive and disjunctive connectives; the complement stands for the fuzzy negation connective. But there are many other propositions in the literature for defining aggregation connectives (Yager, 1991).
A particular list of various fuzzy models for the **AND** connective is present in (Dubois and Prade, 1996). They are corresponding to different linguistically expressions and of course to different logical meaning of the selection criterion.

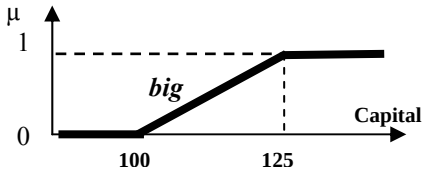

Fig. 4. The ***big*** linguistic value on the **Capital** attribute domain

Example 2. Let's take for example a query based on a complex vague criterion addressed to the COMPANY table:

*Retrieve the **new** companies* **(and)** *having **big** capital*

(*Care sunt firmele **recente** cu capital **mare**?*)

The result of this query evaluation, according to definitions in Fig. 2 and Fig. 4 and to the content of the COMPANY table, is:

Table 3

Name	Capital	Year	μ_{new}	μ_{big}	μ
ELADA	120	1999	0.5	0.8	0.5
LILITEX	130	1999	0.5	1	0.5
FLORIO	110	1999	0.5	0.4	0.4

The satisfaction degree of each linguistic value is computed for each table row and the ***min*** function is used to implement the fuzzy conjunction between them. The result contains the table rows having a significant global satisfaction degree.

The classical form for the conjunctive selection criterion is:

CCSV **::=** CSV **AND** CSV

CSV **::=** <*attribute*> <*linguistic value*>

where the fuzzy model for the conjunction is

$$\mu_{AND}(t) = min(\mu_1(t), \mu_2(t))$$

and t is a database tuple.
About the linguistic values' fuzzy definitions for a database context, it's easy to apply the method presented in the previous section, with the remark that the linguistic values family covers the whole crisp attribute domain as discourse universe.

5. A NEW AGGREGATION OPERATOR. DYNAMIC DEFINITION OF LINGUISTIC VALUES

Our study has found a new class of queries, where two fuzzy criteria are combined in a complex selection criteria such that a second fuzzy criterion is applied on a subset of database rows, already selected by the fist one. We assume that the secondly applied fuzzy criterion is expressed by a linguistic value of a database attribute, which is a gradual property; not an absolute property, but a relative one.
In this case, modeling the linguistic domain of the second attribute requires taking into account not the whole crisp attribute domain, but a limited subset, characteristic to the first criterion-selected database rows.

Example 3. Let's consider as example, the following query addressed to the COMPANY table:

*Retrieve the **new** companies* **within** *the **big** ones*

(*Care sunt firmele **recente*** **dintre** *cele cu capital **mare**?*)

The query evaluation procedure respects the following steps:

Table 4

Name	Capital	Year	μ_{big}
GENERAL	130	1990	1
TIBCO	147	1992	1
LILITEX	130	1999	1
EXPRESS	125	1996	1
ELADA	120	1999	0.8
BAZAR	115	1997	0.6
FLORIO	110	1999	0.4

1. The selection criterion ***big*** **capital** is evaluated, taking into account the definition in the Fig. 4; a tuple group as intermediate result is obtained, where the condition $\mu_{big}>0$ is satisfied (Table 4).

2. The interval containing the register year for the selected companies forms the **Year** sub-domain

[1990, 1999]; this is the one considered later, instead of [1990, 2004].

3. The linguistic value set {***old, almost old, average, almost new, new***} will partition this sub-domain (Fig.5 – the new definitions are labeled in capital letters).

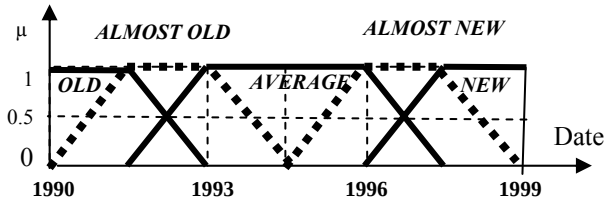

Fig. 5. Linguistic values defined on a sub-domain

4. The global criterion satisfaction degree will result for each tuple.

Table 5

Name	*Capital*	*Year*	μ_{big}	μ_{NEW}	**μ**
GENERAL	130	1990	1	0	0
TIBCO	147	1992	1	0	0
LILITEX	130	1999	1	1	1
EXPRESS	125	1996	1	0	0
ELADA	120	1999	0.8	1	0.8
BAZAR	115	1997	0.6	0.66	0.6
FLORIO	110	1999	0.4	1	0.4

The new aggregation operation can be:

CCSV **::=** CSV **WITHIN** CSV

CSV **::=** <*attribute*> <*linguistic value*>

where the fuzzy model for the conjunction is

$$\mu_{WITHIN}(t) = min(\mu_{1/2}(t), \mu_2(t))$$

t is a database tuple and

$\mu_{1/2}$ is the satisfaction degree of the first criterion relative to the second one.

Two other remarks are interesting:

i. The secondly evaluated gradual property has not an absolute crisp support, but a relative one. Its linguistic label expresses a property relative to the restricted tuple subset and it abandons its absolute original meaning. Therefore, a more suggestive linguistic expression of the new query type, may be:

*Retrieve the **most recent** companies within the **big*** ones

(*Care sunt firmele **mai recente** dintre cele cu capital **mare**?*)

In the precise query context, the ***most recent*** criterion applied on a tuples group is equivalent to the aggregation MAX function and returns the (only one) company having the greatest registration date. But in the imprecise context, the ***big*** company group is a fuzzy set too (a membership degree is attached to each company). The greatest date can correspond here to any company, but not necessary to the biggest one. So, a ranked list of the companies satisfying the ***recent*** criterion, but also taking into account their ***big capital*** degree, is the most adequate response to the above query. In other words, we consider this query can be assimilated with the initially discussed one (*Retrieve the **new** companies* **within** *the **big** ones*), so it can be submitted to the same evaluation procedure. Moreover, it may be even more suggestive for the database user, and semantically adequate to the response in table 5.

ii. The new aggregate operator is not commutative, that is the inversion of the two criteria leads to a different query answer. For example, let's compare the query:

*Retrieve the **new** companies* **within** *the **big*** ones

(*Care sunt firmele **recente** dintre cele cu capital **mare**?*)

with the following one:

*Retrieve the **big** companies* **within** *the **new*** ones

(*Care sunt firmele cu capital **mare** dintre cele **recente**?*)

6. CONCLUSIONS

A particular type of vague query sent to a database is discussed in this paper. A new operator to formalize it and a procedure for evaluate it is proposed. The main idea is to dynamically define sets of linguistic labels on limited attribute domains, determined by previous fuzzy selections. After a comparative analysis including other similar query types, well known as fuzzy model, we conclude that: the evaluation procedure, proposed by this paper, provides an accurate model for the discussed vague expression, with respect to query semantic.

7. REFERENCES

Project BADINS. (1995). Bases de données multimédia et interrogation souple. *Rapport d'activité scientifique*, Institut de recherche en informatique et systèmes aléatoires, Rennes

Project BADINS. (1997). Bases de données multimédia et interrogation souple. *Rapport d'activité scientifique*, Institut de recherche en informatique et systèmes aléatoires, Rennes, 1997

Bodenhofer, U. (2000). A general framework for ordering fuzzy alternatives with respect to fuzzy orderings", *8th Int. Conf. on Information Processing and Management of Uncertainty (IPMU 2000)*, Madrid, pp. 1071-1077

Bodenhofer, U., Bauer, P. (2000). Towards an Axiomatic Treatment of „Interpretability"", *6th International Conference on Soft Computing*, Iizuka, pp. 334-339

Bosc, P., Lietard L., Pivert, O. (1993). On the Interpretation of Set-Oriented Fuzzy Quantified Queries and their Evaluation in a Database Management System, In J. Komorowski, Z. W. Ras (eds.), Lecture Notes in Artificial Intelligence, **689**, Springer-Verlag, pp. 209-218

Dubois, D., Prade, H. (1996). Using fuzzy sets in flexible querying: Why and how?, In H. Christiansen, H.L. Larsen, T. Andreasen (eds.), *Workshop on Flexible Query-Answering Systems*, pp. 89-103

Herrera-Viedma, E. (2000). An Information Retrieval System with Ordinal Linguistic Weighted Queries Based on Two Weighting Semantics, *8th International Conference on Information Processing and Management of Uncertainty in Knowledge-Bases Systems(IPMU 2000)*, Madrid

Kacprzyk, J.., Zadrozny, S. (2001). Computing with words in intelligent database querying: standalone and Internet-based applications, *Information Sciences*, **134**, Elsevier, pp.71-109

Pivert, O. (1991). Contribution à l'interrogation flexible de bases de données - Expression et évaluation de requêtes floues, T*hèse*, Université de Rennes I

Tudorie, C. (2003a). Cercetări privind aplicarea tehnicilor de inteligență artificială pentru interogarea bazelor de date. Scientific Rapport, University ’Dunărea de Jos’, Galați

Tudorie, C. (2003b). Vague criteria in relational database queries. In *Bulletin of “Dunarea de Jos” University of Galați*, **III/2003**, pp. 43-48

Tudorie, C. (2003c). Contribuții la realizarea unei interfețe inteligente pentru interogarea bazelor de date. *Scientific Report*, University ’Dunărea de Jos’, Galați, 2003

Tudorie, C., Dumitriu, L. (2004). How are the Attribute Linguistic Domains Involved in Database Fuzzy Queries EvaluationVague criteria in relational database queries. In *Scientific Bulletin of “Politehnica” University of Timisoara*, 49(63), pp. 61-64

Yager, R.R. (1991). Connectives and quantifiers in fuzzy sets, In *Fuzzy Sets and Systems*, **40-1**, Elsevier Science, pp 39-75