TYPICALITY BASED ON SOFT AGGREGATIONS IN FUZZY OBJECT ORIENTED DATABASES

Gloria Bordogna and Gabriella Pasi
Istituto per le Tecnologie Informatiche Multimediali
CNR - Via Ampere 65, 20131 Milano, Italy
Tel. +39-0270643257/8 Fax: +39-0270643292
Email: fuzzir@itim.mi.cnr.it

Abstract

In this paper a method is proposed to compute typical objects for the classes in the scheme of a fuzzy object oriented database. Typical objects are considered as "representatives" of a fuzzy majority of the class instances. The instances of a class are represented in a topological space and the typical object is derived as "closest" to the fuzzy majority of the class instances.

1. INTRODUCTION

In the conceptual scheme of an Object Oriented Data Base a class identifies the structure shared by a collection of objects; the intensional definition of a class is the specification of its structure in terms of a collection of attributes and methods to operate on them. The extensional definition of a class is the set of its instances, i.e. objects corresponding to real entities of the considered application [1].
The concept of typicality in the OO data model has been introduced to characterize the typical attribute values of a class, thus defining a "virtual-object" which constitutes the class representative.
The primary use of typical objects is to support the association of default attribute values when the actual values are unknown. The use of typical objects makes the data model more informative than when using null values as pointers to unknown attribute values [2, 6, 10] . Moreover typical objects can be useful in database summarization, as they constitute a synthetic view of database contents [12]. Last but not least, in Fuzzy Object Oriented data models typical attribute values can be used to support the computation of the partial membership of objects to fuzzy classes [8].
Generally typical attribute values are defined a priori, in the phase of the data scheme generation, and stored as components of the intensional definition of classes: a limitation of this approach is that the objects stored successively as instances of a class may have attributes which are very different from those previously identified as the typical ones.
In this paper we propose a method to compute typical objects "a posteriori" of the database instantiation. We consider a fuzzy OO data model, the instances of which may have either precise or vague values, defined as trapezoidal possibility distributions over the attribute domain [4,5]. The proposed method is defined as an extension of a method previously defined to compute typical objects for crisp OO data bases [3].
In the literature, it has been outlined that the concept of typicality has a vague nature [7]; in our approach vagueness is modelled by defining vague typical attribute values as well as by computing the typical object as a representative of a fuzzy majority of the class instances. The concept of a fuzzy majority expressed by a linguistic quantifier such as *most of*, was first introduced in the context of fuzzy group decision making to identify a sub-group of consensual experts [9]; in our context this concept is inherited to identify a subset of homogeneous class instances.
In section 2 we synthetically describe the topological model to compute typical objects; in section 3 the representation of the class instances in the topological space are described and in section 4 the computation of typical objects is defined.

2. OVERVIEW OF THE PROPOSED APPROACH

The procedure for the computation of the typical object is based on the definition of a topological space. In the

crisp environment, for each class in the database scheme a D-dimensional space is defined, in which the instances are represented as points whose coordinate values are the precise attribute values [3]. The typical object is computed as a new point obtained by the aggregation of the instance points through a linguistic quantifier; the aggregation is performed separately on each coordinate (referring to a distinct attribute) of the points.
In the fuzzy extension proposed in this paper we consider that the attribute values may be vague, identified by linguistic labels associated with possibility distributions [4,5].
Let us try to figure out how an instance of a class looks like in the space. For sake of simplicity, let us first restrict our attention to a three dimensional space and to a class instance for which two attribute values are known precisely a_1, and a_2, while the third attribute has a vague value represented by a possibility distribution π_3. Now, let us assume that the possibility degrees are associated with grey levels; the value 1 is associated with the black, the value 0 with the white. In the space, the class instance identifies a segment shaded with different grey levels (see Figure 1).

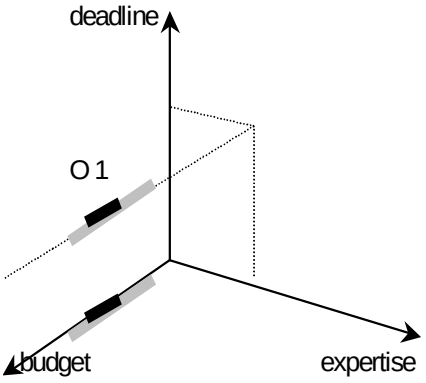

Figure 1: representation of a class instance.

By generalizing this view to the extreme case in which all the attribute values are vague, a class instance identifies an ipercube with a black nucleous and surrounding shells shaded with different grey levels. In this general case, the grey level associated with a point of the ipercube is determined by a combination of the grey levels of its coordinate points. In the following, we will figure out a class instance as an ipercube in the space, even if this happen only as extreme case.
This graphical representation of the class instances is intuitive and makes it possible to identify the typical object of a class as the "virtual" ipercube closest to the fuzzy majority of the class ipercubes. It is determined by computing the possibility distributions characterizing the typical vague values of the attributes for each coordinate axis; this computation consists in the aggregation of the trapezoidal functions representing the attribute values of the class instances through an Induced Ordered Weighed Averaging operator reflecting the fuzzy majority [14]. This computation is based on the adoption of both a metrics for each coordinate of the space and a distance function in the space.
Further, a degree of consensus or agreement on the typical object among the instances in the fuzzy majority is evaluated: it is interpreted as an indicator of the *cohesion* of the class instances in the fuzzy majority. When the cohesion degree is low, i.e. below a given threshold, it means that the typical object is not a valid representative of the instances; in this case a new typical object representative of a stricter fuzzy majority of the instances can be computed.

3. A TOPOLOGICAL REPRESENTATION OF OBJECT INSTANCES

In database applications the available information is often characterized by *incompleteness*: a datum may be either ill-known (vague or uncertain), or completely unknown or it may not exist. Usually in databases when there is a lack of knowledge about the value of an attribute a *null value* is used as a place holder for the unknown value. Different kinds of null values have been introduced in the literature to characterize the different types of incompleteness [2,6,18].

In OOD models the computation of typical attribute values as prototypal values of the attributes of each class would offer a more infomative way than null values to represent unknown information.
Recently fuzzy set and possibility theory have been applied to provide a more flexible and accurate way to model incompleteness in data models. To this aim, several fuzzy Object Oriented data models have been proposed in the literature [4, 8, 11, 15]. Some of these models assume that at the schema level, typical attribute values are defined [8, 10].
We call typical object a representative instance of a structured class (a class having more attributes), while a representative instance of an attribute is a typical attribute value. A typical object is characterized by a set of typical attribute values.
The specification of typical attribute values is usually assumed as an a priori operation performed at the scheme definition level by an expert of the application domain. In [3] a method has been proposed to compute typical objects “a posteriori” of the database instantiation as representatives of a fuzzy majority of the class instances. To this aim, for each class in the scheme a topological space is defined with a dimension D given by:

$$D = \sum_{i=1}^{N} D_i$$

where:

- N is the number of the attributes of the class which are semantically significant in defining the typical object: for example the attributes having a numeric domain, such as the attribute age, or an ordinal domain, such as the attribute hear color. In fact, it makes sense to define a typical age of a population or a typical hear color, while it is a nonsense to define the typical name.
- D_i is the dimension of the space associated with the i-th attribute of the class. $D_i = 1$ for the attributes having a primitive domain. In this context only single valued attributes are considered. Further it is not allowed to choose as semantically significant the class attributes with have the class itself as domain.

In the crisp context, each coordinate of the topological space is associated with an attribute having a primitive domain. The instances of the class identify points whose coordinates are the precise values of the associated attributes.
In the fuzzy extension, both the instances of a class and the typical object are identified by ipercubes of the D-dimensional space, as it has been illustrated in section 2. Each axis of this space is associated with the basic domain of a "significant" attribute of the class; the attribute values of a class instance are defined by possibility distributions with a trapezoidal shape $(\alpha, \beta, \chi, \delta)$ on the attribute basic domain; when an attribute value is not vague (it is either precise or imprecise), we will assume that it is represented by a possibility distribution too (for a precise value $\alpha = \beta = \chi = \delta$; for an imprecise value $\alpha = \beta$ and $\chi = \delta$).

4. TYPICAL OBJECT COMPUTATION

The procedure for the computation of the typical object of a class is the following. First, for each class a subset of its instances is selected randomly as a sample set; it is assumed that the class instances are represented by this sample set. The sample set cardinality (K) strongly influences the time needed for the computation of the typical object.
An evaluation matrix can be defined in which the rows refer to the K instances of the class and the columns refer to the D attributes (i.e. the space dimension). The value $\pi_{ij} = (\alpha_{ij} , \beta_{ij} , \chi_{ij} , \delta_{ij})$ of the matrix is the trapezoidal possibility distribution of the j-th attribute for the i-th instance.

Class C	A1	A2	A3	A4	A5
Object 1					
Object 2		π_{22}			
Object 3				π_{34}	

Table 1: matrix representing the instances of a class

The typical object of a class is computed so as to take into account a fuzzy majority Q, such as "*most*" [9, 13], of the K class instances: it can be also figured out as an ipercube $\underline{o}$ in the D-dimensional space of the attributes closest to the ipercubes in the fuzzy majority Q.
The rows of the decision matrix, are used to compute the typical object $\underline{o}$.
The aggregation function is formalized by an IOWA operator of dimension K, with the weighting vector W automatically defined so as to reflect the semantics of the quantifier Q [13]. First, by following Zadeh [16], the membership function of the fuzzy subset representing the relative monotone non decreasing quantifier is defined $Q : [0,1] \rightarrow [0,1]$. In fact, relative monotone non decreasing quantifiers model the semantics associated with the concept of a fuzzy majority which increases as the number of the elements in the fuzzy majority come close to *all*.
Then the K elements $w_i \in [0,1]$ of the weighting vector W are obtained as:

$$w_i = Q(i/K) - Q((i-1)/K) \quad \forall i=1 \ldots K$$

We remind that an Induced Ordered Weighted Average operator (IOWA) allows to generalize the OWA so as the non linear component requiring the reorder of the arguments can be defined based on a reorder vector R [14].
In our context, since the possibility distributions are represented by trapezoidal fuzzy numbers $(\alpha, \beta, \chi, \delta)$, we apply the IOWA separately to the bags:
$(\alpha_1 ,..., \alpha_K)$, $(\beta_1,...,\beta_K)$, $(\chi_1,...,\chi_K)$, $(\gamma_1,...,\gamma_K)$, by considering a common reorder vector R.
By considering the aggregation of the $(\alpha_1 ,..., \alpha_K)$, the IOWA operator is defined as follows:

$$IOWA(\langle r_1, \alpha_{1j}\rangle, ..., \langle r_K, \alpha_{K\,j}\rangle) = \sum_{i=1}^{K} b_i w_i$$

in which the value b_i is the α_j that has associated the i-th largest of the r values.
The reorder vector R is common to both the aggregations on all the axis, and to the aggregations of all the four bags along each axis. Its elements are determined on the

basis of the euclidean distance of the instances from their average instance.
The average instance is obtained by applying an average operator to the values (α, β, χ, δ), in each column of the decision matrix.
A distance measure between possibility distributions is adopted : let us indicate by $d_h(\pi_{ih}.\pi_{jh})$ the distance on the h-th axis [17]. Then, the Euclidean distance between two class instances d (o_i o_j) is obtained as:

$$d(o_i\ o_j)=\sqrt{d_1(\pi_{i1}.\pi_{j1})^2+\ d_2(\pi_{i2}.\pi_{j2})^2+\ ..\ d_D(\pi_{iD}.\pi_{jD})^2}$$

The K instances are then ranked based on R in decreasing order of their euclidean distance from the average instance distance $d(o_i , o_{Average})$.
By using this reorder criterion the typical attribute values are determined more heavily by the instance attribute values which are "closer" to the average value of the attributes. The typical attribute values computed in this way are seen as representatives of Q attribute values, and the typical object $\underline{o}$ is the representative of Q instances.
The degree of cohesion of the class instances in the fuzzy majority with respect to the typical object is computed as follows:

- first, the similarity between each instance o_i and the typical object $\underline{o}$ is evaluated; we interpret this evaluation as the computation of the *degree of cohesion of the instance with respect to Q instances*; the similarity measure between pairs of instances is defined as the complement of their Euclidean distance normalized with respect to the maximum:

$$sim(o_i\ \underline{o}) = 1 - \frac{d(o_i\,\underline{o})}{\max_{j=1,\ldots K}(d(o_j\,\underline{o}))}$$

- second, the *degree of cohesion among Q instances* is computed; this is performed by aggregating the similarity degrees computed in the previous step by using the OWA operator associated with Q.

When the degree of cohesion is not satisfactory, i.e. it is below a given threshold, it is assumed that the typical object is not a valid representative of the considered majority. In this case, one can try to compute a new typical object as representative of a stricter fuzzy majority.

References

1. E. Bertino, and L. Martino, Object-Oriented Database Management Systems: Concepts and Issues. *IEEE Computer* 4, 33 ,1991.
2. J. Biskup, A formal approach to null values in database relations, in *Advances in Database Theory*, H. Gallaire & J.M. Nicolas eds., 1, Plenum Press, 299, 1981.
3. G. Bordogna and G. Pasi, "Typicality based on soft Aggregations in Crisp Object Oriented Database", in proc. *of NAFIPS'99*, New York, 9-11 June, 1999.
4. G.Bordogna and G. Pasi, D. Lucarella, A Fuzzy Object Oriented Data Model for Managing Vague and Uncertain Information, to appear in *International Journal of Intelligent Systems*, 1999.
5. P.Bosc, and H. Prade, An introduction to fuzzy set and possibility theory-based treatment of soft queries and uncertain or imprecise databases. *Report IRIT/93-57-R*, 1993.
6. E. F. Codd, Missing Information (applicable and inapplicable) in relational databases, *SIGMOD record*, 15, 53, 1986.
7. D. Dubois, H. Prade, J-P. Rossazza, Vagueness, Typicality and Uncertainty in Class Hierarchies, *Int. Journal of Intelligent Systems*, 6, 167, 1991.
8. R.George, B.P. Buckles, F.E. Petry, Modelling Class Hierarchies in the Fuzzy Object-Oriented Data Model, *Fuzzy Sets and Systems* 60(3), 1993.
9. J. Kacprzyk, M. Fedrizzi and H. Nurmi, Fuzzy Logic with Linguistic Quantifiers in Group Decision Making and Consensus Formation, in An Introduction to Fuzzy Logic Applications in Intelligent Systems, R.R. Yager and L.A. Zadeh eds., Kluwer, 263-280, 1992.
10. G. Pasi , and R. R. Yager, Calculating Attribute Values using Inherithance Structures in Fuzzy Object Oriented Data Models, to appear in *IEEE Transactions on Systems, man and Cybernetics*, 1999.
11. R. De Caluwe, *Fuzzy and Uncertain Object-Oriented databases, Concepts and Models*, World Scientific, Singapore, 1997.
12. R.R. Yager, A new approach to Summarization of data: Information Science, 28, 69-86, 1982.
13. R.R. Yager, Interpreting Linguistically Quantified Propositions, *International Journal of Intelligent Systems*, 9, 541, 1994.
14. R.R Yager., D. Filev, Operations for Granular Computing: Mixing Words and Numbers, in proc. of *FUZZIEEE*, 123, 1998.
15. N.Van Gyseghem, R.de Caluwe, R. Vandenberghe, UFO: Uncertainty and Fuzziness in an Object-Oriented model, *II IEEE Int. Conf. on Fuzzy Systems*, 1, 489 (San Francisco, USA, 1993.
16. L.A. Zadeh, A computational Approach to Fuzzy Quantifiers in Natural Languages, *Computing and Mathematics with Applications.* 9, 149, 1983.
17. L.A. Zadeh, Similarity relations and fuzzy orderings. Information Science, 3, 177-200, 1971
18. R. Zicari, Incomplete Information in Object-Oriented Databases, *SIGMOD RECORD*, 19, 33-40 1990.