

On flexible data representation and querying via extended fuzzy sets

Guy de Tré, Rita de Caluwe

Computer Science Laboratory
University of Ghent
Sint-Pietersnieuwstraat 41,
B-9000 Ghent, Belgium
{guy.detre,rita.decaluwe}@ugent.be

Janusz Kacprzyk, Sławomir Zadrozny

Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6,
01-447 Warsaw, Poland
{kacprzyk,zadrozny}@ibspan.waw.pl

Abstract

In search of semantic richer and more flexible database modelling and database querying techniques, different approaches based on fuzzy set theory have been developed. Among the most successful approaches are the possibilistic and similarity based models. More recently, extended possibilistic logic and extended fuzzy sets have been applied to further enrich flexible database models. In this paper it is presented how ‘IS’ predicates in flexible database queries can be evaluated in the presence of data that is modelled by such extended fuzzy sets. Furthermore, a comparison with the regular possibilistic database modelling approach is given.

Keywords: flexible querying, ‘IS’ predicates, fuzzy sets, extended fuzzy sets, IVFS, IFS, extended possibilistic truth values.

1 Introduction

A lot of information is only available in an imperfect form. Indeed, information could be imprecise, vague, uncertain, incomplete, inconsistent, etc. Traditional database models are not able to cope efficiently with such kinds of imperfectness. This observation was one of the main onsets for the development of more advanced, semantic richer database models now commonly known as ‘fuzzy’ databases. Among the most successful ‘fuzzy’ database models are the possibilistic and similarity based models. Overviews and descriptions of these models can be found in [?, ?, ?, ?, ?]. Some recent developments are described in [?].

This paper deals with two, more recent research advances in flexible database querying: the first one dealing with the applicability of various types of extended fuzzy sets in data representation and flexible database querying as exemplified by [?] and the second one dealing with the use of extended possibilistic truth values in a logical framework to support flexible querying [?]. More specifically, the focus of the paper is on the evaluation of so-called ‘IS’ predicates in the conditions of a flexible query under the assumptions that the data is modelled by interval-valued, intuitionistic or two-fold fuzzy sets (referred to, in what follows, as IVFS, IFS and TFS, respectively) and that the underlying logical framework is based on extended possibilistic truth values. New evaluation function for such extended fuzzy sets are presented and compared with the existing evaluation function for regular possibilistic data.

The remainder of the paper is organized as follows. In section 2 some preliminary definitions on various types of extended fuzzy sets and extended possibilistic truth values are given. Section ?? deals with the ‘IS’ predicates. A rationale for the use of various types of extended fuzzy sets in modelling of the data in a database and in a query is sought for. Both the traditional possibilistic and extended evaluation functions are described and compared.

2 Some Preliminaries

2.1 IVFS Sets

An interval-valued fuzzy set [?] over a universe of discourse U

$$F = \{ \langle u, \mu_F^l(u), \mu_F^u(u) \rangle \mid u \in U \}$$

is defined by two mappings $\mu_F^l, \mu_F^u : U \rightarrow [0, 1]$ such that

$$0 \leq \mu_F^l(u) \leq \mu_F^u(u) \leq 1, \forall u \in U$$

For each $u \in U$ the numbers $\mu_F^l(u)$ and $\mu_F^u(u)$ respectively represent the lower and upper bound on the degree of membership of u in F . Considering the special case where $\mu_F^l = \mu_F^u$, it can easily be seen that IVFS sets are generalizations of regular fuzzy sets.

If used to model imprecise data in a flexible database, a IVFS set can be assigned a possibilistic interpretation as to represent an extended possibility distribution (IVPD [?]). An IVPD might be associated with a database attribute A in which case it can be denoted by (π_A^l, π_A^u) and is characterized by the mappings π_A^l and π_A^u , such that $0 \leq \pi_A^l(u) \leq \pi_A^u(u) \leq 1, \forall u \in U$. Furthermore, it is assumed that $\pi_A^l(u)$ defines the lower bound on the degree of possibility that $A = u$ and $\pi_A^u(u)$ defines the upper bound on the degree of possibility that $A = u$. Thus, such a IVPD accounts for both imprecision of the data represented in a database and uncertainty connected with such a representation.

An IVFS set may be also used in a query in a case when membership degrees cannot be precisely assessed. For example, it may be convenient for a user to specify that the age of 30 is compatible with the concept of “young” to the degree between 0.4 and 0.6.

2.2 IFS Sets

An intuitionistic fuzzy set[?], referred to later as IFS set ¹, over a universe of discourse U

$$F = \{ \langle u, \mu_F(u), \nu_F(u) \rangle \mid u \in U \}$$

¹As there is currently a dispute on the appropriateness of the original name going on we prefer to use this abbreviation

is defined by two mappings $\mu_F, \nu_F : U \rightarrow [0, 1]$ such that

$$0 \leq \mu_F(u) + \nu_F(u) \leq 1, \forall u \in U$$

For each $u \in U$ the numbers $\mu_F(u)$ and $\nu_F(u)$ respectively represent the degree of membership and the degree of nonmembership of u in F . Considering the special case where $\nu_F = 1 - \mu_F$, it can easily be seen that the IFS fuzzy sets are generalizations of regular fuzzy sets. Moreover, taking $\mu_F^l = \mu_F$ and $\mu_F^u = 1 - \nu_F$, an IFS set may be formally treated as an IVFS set. Further study of this similarity is beyond the scope of this paper. In what follows, we are interested only in the interpretation of both types of sets in the context of database querying.

If used to model imprecise data in a flexible database, a IFS fuzzy set can be assigned a possibilistic interpretation, i.e., imply an extended possibility distribution (IPD). A IPD might be associated with a database attribute A in which case it can be denoted by $(\pi_{\mu_A}, \pi_{\nu_A})$ and is characterized by the mappings π_{μ_A} and π_{ν_A} , such that $\pi_{\mu_A}(u) + \pi_{\nu_A}(u) \leq 1, \forall u \in U$. Furthermore, it is assumed that $\pi_{\mu_A}(u)$ defines the degree of possibility that $A = u$ and $\pi_{\nu_A}(u)$ defines a degree of explicit “impossibility” that $A = u$. In case of an ordinary possibilistic distribution π , such a degree of impossibility is equal $1 - \pi(u)$, i.e., is fully determined by $\pi(u)$. The notion of impossibility degree seems to be quite intuitive, however it may be cast in the ordinary possibilistic context by the observation that $\nu(u)$ in IFS set substitutes for $1 - \mu(u)$ in the usual fuzzy set. The latter, and thus the former too, might be interpreted as the necessity that $A \neq u$.

Let us illustrate the idea of using IFS related extended possibility distribution to represent the data in a database. Let us assume a database table collecting information on some crimes believed to be committed by somebody from a list of suspects. Additionally, we assume that each crime was committed by a single person. Thus, we have an attribute indicating a perpetrator. Its value might be a IPD implied by an IFS set of suspects. Such an IFS set, F , might emerge, e.g., from the following procedure. A group of experts studied the data on the crimes and characteristics of par-

ticular suspects under consideration and voted for each suspect:

- “yes” if an expert finds given suspect a possible perpetrator of given crime
- “no” if an expert finds impossible that given suspect committed given crime
- an expert may abstain from voting if he or she cannot decide neither “yes” nor “no”

Then, the proportion of “yes” and “no” votes may be used to assess the membership functions μ and ν , respectively, of the IFS set F . Finally, the corresponding IPD distribution indicates for each suspect how possible (π_{μ_F}) and impossible (π_{ν_F}) perpetrator of given crime he or she is.

An IFS set finds a more intuitive applicability in a query to indicate which values of an attribute are preferred and which are to be avoided. The memberships of the former are represented by μ_{π_A} , while of the latter by ν_{π_A} .

2.3 TFS Sets

A two-fold fuzzy set F [?] over a universe of discourse U is defined as a pair of two fuzzy sets $F = (P, A)$ such that $\text{support}(P) \subseteq \text{core}(A)$. As $0 \leq \mu_P(u) \leq \mu_A(u) \leq 1$, taking $\mu_F^l = \mu_P$ and $\mu_F^u = \mu_A$ or $\mu_F = \mu_P$ and $\nu_F = 1 - \mu_A$, a TFS set may be formally treated as a special case of IVFS or IFS set, respectively. However, in what follows, we are interested only in the interpretation of these sets in the context of database querying.

An TFS set may be used in a query to indicate which values of an attribute are allowed (acceptable, not rejected) and which (from among them) are really preferred, cf. [?]. The membership degrees of these values correspond to μ_A and μ_P in case of the former and the latter values, respectively.

In view of the relation between TFS fuzzy sets and both IVFS and IFS sets the former might be also used to represent data in a database. Using the interpretations of IVPD and IPD, respectively, one may treat twofold possibility distribution (TPD) as a special case such that:

- lower possibility π_A^l may be greater than 0 only when the upper possibility π_A^u is equal 1; when treating TPD as a special case of IVPD,
- impossibility degree of an element u , $\pi_{\nu_F}(u)$, have to be 0 if its possibility, $\pi_{\mu_F}(u)$, is greater than 0; when treating TPD as a special case of IPD (in terms of the example on crimes database it means that it is not allowed to have “yes” and “no” votes for the same suspect)

2.4 Extended Possibilistic Truth Values

The concept ‘extended possibilistic truth value’ (EPTV) [?] is defined as a (normalized) possibility distribution over the universal set $I^* = \{T, F, \perp\}$ of truth values, where T represents ‘True’, F represents ‘False’ and \perp represents an ‘undefined’ truth value. An EPTV can be used to express the result of the evaluation of (belief in) a proposition p and is then denoted by $\tilde{t}^*(p)$. In general

$$\tilde{t}^*(p) = \{(T, \mu_{\tilde{t}^*(p)}(T)), (F, \mu_{\tilde{t}^*(p)}(F)), (\perp, \mu_{\tilde{t}^*(p)}(\perp))\}$$

where $\mu_{\tilde{t}^*(p)}(T)$ represents the possibility that proposition p is true, $\mu_{\tilde{t}^*(p)}(F)$ represents the possibility that proposition p is false and $\mu_{\tilde{t}^*(p)}(\perp)$ represents the possibility that (some of) the elements of p are not applicable, undefined or not supplied.

In this way, EPTVs provide an epistemological representation of the truth of a proposition, which allows to reflect the knowledge about the actual truth and additionally allow to explicitly deal with those cases where the truth value of a proposition is (partly) undefined.

An overview of some special values of EPTVs is given in table ???. As an example, consider the modelling of an unknown truth value by the possibility distribution $\{(T, 1), (F, 1)\}$, which denotes that it is completely possible that the proposition is true (T), or it is also completely possible that the proposition is false (F).

New propositions can be constructed from existing propositions, using generalized logical operators. An unary operator ‘ \sim ’ is provided for the

Table 1: Special cases.

$\tilde{t}^*(p)$	Interpretation
$\{(T, 1)\}$	p is true
$\{(F, 1)\}$	p is false
$\{(T, 1), (F, 1)\}$	p is unknown
$\{(\perp, 1)\}$	p is undefined
$\{(T, 1), (F, 1), (\perp, 1)\}$	no information

negation of a proposition and binary operators ‘ $\tilde{\wedge}$ ’, ‘ $\tilde{\vee}$ ’, ‘ $\tilde{\Rightarrow}$ ’ and ‘ $\tilde{\Leftrightarrow}$ ’ are respectively provided for the conjunction, disjunction, implication and equivalence of propositions. The arithmetic rules to calculate the EPTV of a composite proposition and the algebraic properties of extended possibilistic truth values are presented in [?].

3 ‘IS’ predicates in flexible querying

In flexible database querying so-called ‘IS’ predicates can be used to compare stored values (or labels) with labels provided by the user. In general an ‘IS’ proposition is of the form

$$A \text{ IS } L$$

where A is the name of a database attribute and L is a linguistic term that is given by the user. Examples of ‘IS’ propositions are: ‘*Price* IS cheap’, ‘*Age* IS young’, ‘*Duration* IS long’ and ‘*Weight* IS heavy’. Hereby, *Price*, *Age*, *Duration* and *Weight* are database attributes and ‘cheap’, ‘young’, ‘long’ and ‘heavy’ are linguistic terms that represent values within the domain associated with the attribute and are modelled by means of fuzzy set theory.

From the possibilistic logic viewpoint, the evaluation of an ‘IS’ proposition should be interpreted as follows [?]. L is a fuzzy set and π_A is a possibility distribution being a value of the attribute A at given record. $p = A \text{ IS } L$ is treated as a proposition of a multi-valued logic, thus possessing a fuzzy set of models, M . On the other hand, π_A is a possibility distribution on the space of interpretations. An interpretation is meant here as an assignment of a value to the attribute A . Then, evaluation of the proposition p corresponds to the computing of the possibility measure (under π) of the set M and its complement, meant as possibility

that proposition p is true and false, respectively. Next we discuss how it might be adopted for the case of three logical values and various types of extended fuzzy sets.

3.1 Possibilistic modelling approach

In the possibilistic approach, attribute values are represented by possibility distributions. With the understanding that π_A is the possibility distribution function characterizing the value of attribute A with domain dom_A and μ_L is the membership function of a fuzzy set that models the label L that is given by the user, the EPTV of the proposition ‘ $A \text{ IS } L$ ’ is defined by²

- $\mu_{\tilde{t}^*(‘A \text{ IS } L’)}(T) = \sup_{x \in dom_A} \min(\pi_A(x), \mu_L(x))$
- $\mu_{\tilde{t}^*(‘A \text{ IS } L’)}(F) = \sup_{x \in dom_A \setminus \{\perp_A\}} \min(\pi_A(x), 1 - \mu_L(x))$
- $\mu_{\tilde{t}^*(‘A \text{ IS } L’)}(\perp) = \min(\pi_A(\perp_A), 1 - \mu_L(\perp_A))$

Hereby, it has been explicitly assumed that each domain dom_A contains a special value ‘undefined’, represented by \perp_A , that is used to model the inapplicability of a ‘regular’ domain value.

As an example, the calculation of the EPTV for the proposition ‘around₃₀ IS young’ for an attribute ‘Age’ is illustrated in Fig. ???. Hereby, $\mu_{\tilde{t}^*(‘around_{30} \text{ IS } young’)}(T)$ is shortly written as μ_T and $\mu_{\tilde{t}^*(‘around_{30} \text{ IS } young’)}(F)$ as μ_F . $\mu_{\tilde{t}^*(‘around_{30} \text{ IS } young’)}(\perp) = 0$.

3.2 Modelling approach based on extended fuzzy sets

In a modelling approach based on extended fuzzy sets both, data and query representations might employ these extensions. We will focus here on IFS sets here, which due to their interpretation seem to provide for the most challenging context. Thus, we assume that, on the one hand, the values of the attribute A of an ‘IS’ proposition could be modelled by a IPD $(\pi_{\mu_A}, \pi_{\nu_A})$. On the other hand, the linguistic term L might be modelled by means

²This definition is an extension of the one originally given in [?].

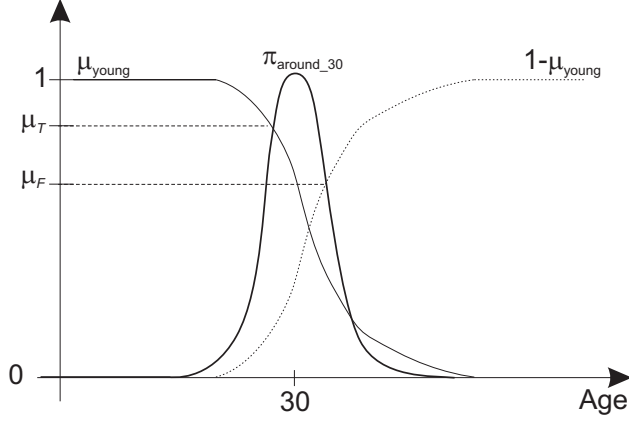


Figure 1: Example of possibilistic approach.

of an IFS set with membership function μ_L and nonmembership function ν_L .

Let us first start with a regular possibility distribution π_A being the value of an attribute A . In such cases, the EPTV of the proposition ‘ A IS L ’ is defined by

- $\mu_{\tilde{t}^*(‘A \text{ IS } L’)}(T) = \sup_{x \in \text{dom}_A} \min(\pi_A(x), \mu_L(x))$
- $\mu_{\tilde{t}^*(‘A \text{ IS } L’)}(F) = \sup_{x \in \text{dom}_A \setminus \{\perp_A\}} \min(\pi_A(x), \nu_L(x))$
- $\mu_{\tilde{t}^*(‘A \text{ IS } L’)}(\perp) = \min(\pi_A(\perp_A), \nu_L(\perp_A))$

In such a case, a counterpart of the example in Fig. ?? is given in Fig. ?. Information about nonmembership of elements denoted by the linguistic term L is now given in a more general way by the nonmembership function ν_L , instead of by the more constrained complement $1 - \mu_L$. This allows for example, as illustrated in the figure, to consider ages between 30 and 50 as neither young, nor not young. Again, $\mu_{\tilde{t}^*(‘around_30 \text{ IS } young’)}(T)$ is shortly written as μ_T and $\mu_{\tilde{t}^*(‘around_30 \text{ IS } young’)}(F)$ as μ_F . $\mu_{\tilde{t}^*(‘around_30 \text{ IS } young’)}(\perp) = 0$.

Now, let us consider the case where both data and query are based on IFS sets. Thus, L is, as previously, an IFS set and the value of A is an IPD $(\pi_{\mu_A}, \pi_{\nu_A})$. In such cases, the EPTV of the proposition ‘ A IS L ’ is defined by

- $\mu_{\tilde{t}^*(‘A \text{ IS } L’)}(T) = \sup_{x \in \text{dom}_A} \min(\pi_{\mu_A}(x), \mu_L(x))$

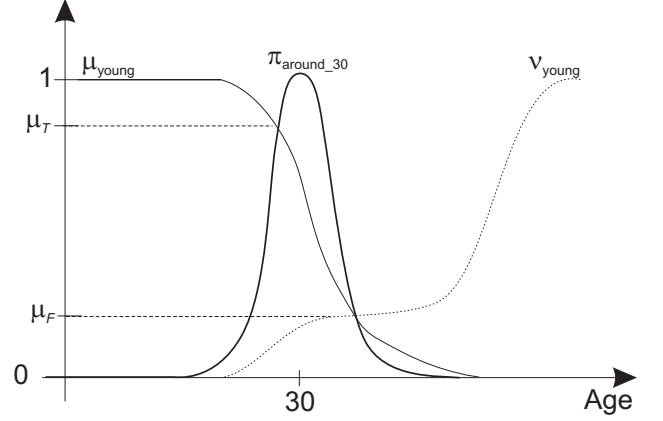


Figure 2: Example of IFS-based approach.

- $\mu_{\tilde{t}^*(‘A \text{ IS } L’)}(F) = \sup_{x \in \text{dom}_A \setminus \{\perp_A\}} \min(\pi_{\mu_A}(x), \nu_L(x))$
- $\mu_{\tilde{t}^*(‘A \text{ IS } L’)}(\perp) = \max(\min(\pi_{\mu_A}(\perp_A), \nu_L(\perp_A)), \sup_{x \in \text{dom}_A \setminus \{\perp_A\}} \min(1 - \pi_{\mu_A}(x) - \pi_{\nu_A}(x), \mu_L(x)))$

Basically, the formula for $\mu_{\tilde{t}^*(‘A \text{ IS } L’)}(\perp)$ has been essentially extended. Let us illustrate the rationale on an example referring to the previously introduced database on crimes. Let us assume that for the crime coded C1 all experts abstained when asked if John is a possible perpetrator. Thus, we have $\pi_{\mu_A}(John) = \pi_{\nu_A}(John) = 0$. Then, if the query asks if John is a potential perpetrator of the crime C1 the answer, according to the previous formulae, would be $\mu_{\tilde{t}^*(‘A \text{ IS } L’)}(T) = 0$ and $\mu_{\tilde{t}^*(‘A \text{ IS } L’)}(\perp) = 0$, the value of $\mu_{\tilde{t}^*(‘A \text{ IS } L’)}(F)$ depending on the π_{μ_A} for other suspects. The new formula yields $\mu_{\tilde{t}^*(‘A \text{ IS } L’)}(\perp) = 1$ what seems reasonable, taking into account that experts were not in a position to decide the John’s case.

As described above, EPTVs are suited to model the satisfaction of ‘IS’ propositions in a flexible way in both the possibilistic and extended approaches.

As shortly illustrated in Fig. ??, the use of IFS sets allows for more flexibility in the expression of nonmembership and thus can be used in database applications where such a flexibility is required.

The nonmembership related part π_{ν_A} of the IPD $(\pi_{\mu_A}, \pi_{\nu_A})$ for A , has not been used for the calculation of the EPTV $\tilde{t}^*(‘around_30 \text{ IS } young’)$. In

fact, we are only interested in the truth value of ‘around 30’ being ‘young’ (and ‘not young’), not in the truth value of ‘not around 30’ being ‘young’ (and ‘not young’). The nonmembership part has been used in the another example only for the calculation of $\mu_{\tilde{t}^*(\text{‘}A \text{ IS } L\text{’})}(\perp)$. It seems it requires further studies if it should be taken into account also in other cases. For example, in cases where it might be reasonable to calculate the EPTV of propositions of the form ‘ $NOT(A) \text{ IS } L$ ’, π_{ν_A} can be used as follows

- $\mu_{\tilde{t}^*(\text{‘}NOT(A) \text{ IS } L\text{’})}(T) = \sup_{x \in \text{dom}_A} \min(\pi_{\nu_A}(x), \mu_L(x))$
- $\mu_{\tilde{t}^*(\text{‘}NOT(A) \text{ IS } L\text{’})}(F) = \sup_{x \in \text{dom}_A \setminus \{\perp_A\}} \min(\pi_{\nu_A}(x), \nu_L(x))$
- $\mu_{\tilde{t}^*(\text{‘}NOT(A) \text{ IS } L\text{’})}(\perp) = \min(\pi_{\nu_A}(\perp_A), \nu_L(\perp_A))$

4 Conclusions

In this paper, it is illustrated how extended fuzzy sets and the corresponding possibility distributions can be used in flexible database querying. More specifically, the focus is on the evaluation of propositions with ‘IS’ predicates, under the assumption that an underlying logical framework that is based on EPTVs is used and IFS sets are employed to represent data nad query. Both the classical possibilistic and extended evaluation methods are described and compared with each other. This subject definitely requires further study and is in line with the current trend to exploit both positive and negative knowledge (bipolarity).

References

- [1] K. Atanassov, “Intuitionistic fuzzy sets”, *Fuzzy Sets and Systems* **20** (1986) 87–96.
- [2] G. Bordogna and G. Pasi (eds.), *Recent Issues on Fuzzy Databases* (Physica-Verlag, Heidelberg, Germany, 2000).
- [3] P. Bosc and J. Kacprzyk (eds.), *Fuzziness in Database Management Systems* (Physica-Verlag, Heidelberg, Germany, 1995).
- [4] R. De Caluwe (ed.), *Fuzzy and Uncertain Object-Oriented Databases: Concepts and Models* (World Scientific, Singapore, 1997).
- [5] G. de Tré, “Extended Possibilistic Truth Values”, *International Journal of Intelligent Systems* **17** (2002) 427–446.
- [6] G. De Tré and R. De Caluwe, “Modelling Uncertainty in Multimedia Database Systems: An Extended Possibilistic Approach”, *Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **11** 1 (2003) 5–22.
- [7] D. Dubois and H. Prade, “Twofold fuzzy sets and rough sets - Some issues in knowledge representation”, *Fuzzy Sets and Systems* **23** (1987) 3–18.
- [8] D. Dubois and H. Prade, *Possibility Theory* (Plenum Press, New York, USA, 1988).
- [9] D. Dubois and H. Prade, “Possibility theory, probability theory and multiple-valued logics: A clarification”, *Annals of Mathematics and Artificial Intelligence* **32** (2001) 35–66.
- [10] D. Dubois and H. Prade, “Bipolarity in flexible querying”, in: T. Andreasen et al. (Eds.), *FQAS 2002*, LNAI 2522 (Springer-Verlag, Berlin, Heidelberg, 2002), 174–182.
- [11] P. Grzegorzewski and E. Mrówka, “Flexible Querying via Intuitionistic Fuzzy Sets”, in: *Proceedings of the 3rd Int. Conference in Fuzzy Logic and Technology — Eusflat 2003* (Zittau, Germany, 2003) 228–231.
- [12] Z. Ma (ed.), *Advances in Fuzzy Object-Oriented Databases: Modeling and Applications* (IDEA Group Publishing, Hershey, USA, 2005).
- [13] M. Oussalah, “Interval possibility measures: basic concepts and conditioning”, *Kybernetes*, **32** 3 (2003) 317–342.
- [14] F.E. Petry, *Fuzzy Databases: Principles and Applications* (Kluwer Academic Publishers, Boston, USA, 1996).
- [15] I. B. Türksen, “Interval valued fuzzy sets based on normal forms”, *Fuzzy Sets and Systems* **20** (1986) 191–210.

- [16] A. Yazici and R. George, *Fuzzy Database Modeling* (Physica-Verlag, Heidelberg, Germany, 1999).