Importance weighted similarity based soft interpretation of crisp criteria queries *

Henrik Legind Larsen
Department of Computer Science,
Roskilde University, DK-4000 Roskilde
hll@ruc.dk

Abstract

We present a scheme for utilizing knowledge on user preferences and similarities in attribute domains for soft interpretation of multicriteria user queries. The knowledge on user preference is assumed to be represented for each user or user category by importance weights attached to the attributes. This knowledge defines a modification of the similarities in the attribute domain, thus providing a view into the information base. The view prescribes the soft interpretation of user queries. Numerical attributes and non-numerical term attributes are handled in a common framework, making the scheme applicable for information bases with any mix of the two kinds of attributes.

Keywords: flexible query answering, query interpretation, needs modeling, importance weighting, query enveloping.

1 Introduction

Despite the uncertainty of natural language, humans typically communicate their needs rather precisely using few words. This is possible because the communication takes places in a context that yields the intended meaning through removing ambiguities and exploiting knowledge on similarities and user preferences. While the query uttered by the user typically is in some crisp form, the user's information needs are, in general, modeled by fuzzy sets. We shall refer to the information needs as the *needs concept*, to distinguish it from the query uttered which we shall refer to as the *user query*. Several approaches to flexible querying systems (*e.g.* [2, 6]) assume that the users will describe their fuzzy needs through supplying fuzzification information. While ordinary, non-expert users do typically not give such information, it is needed by the system to determine whether an item is of potential interest, even if it does not satisfy the user query directly.

We present an approach to utilize application domain knowledge on similarities in attribute domains and on user preferences w.r.t. attributes for soft interpretation of user queries into needs concepts. The user preferences are represented by importance weights attached to each attribute for each user or user category. The known importance of an attribute for a user defines a modification of the similarities in the attribute domain. Thus, if the importance of an attribute is lower for some user, then the values in the attribute domain are more similar in the view of the user, in the sense, that other values become more acceptable as alternatives to the values specified by the user. The focus in this paper is on the scheme for representing and utilizing such knowledge in soft query interpretation. The knowledge may be acquired from several sources, such as human experts, published domain ontologies, and, last, but not least, data mining in the logs of the users' click streams, transactions, *etc.* However, the topic of knowledge acquisition is outside the scope of this paper.

* Work supported in part by the Danish Natural Science Research Council (Project No. 51-00-0376), and the IT University of Copenhagen.

The needs concept is derived through first determining the needs criterion for each criterion in the user query, and then joining the needs criteria into the needs concept. To access in information base systems supporting only non-fuzzy queries, we derive a (crisp) envelope [1] of the needs concept. The envelope is then converted into an equivalent access query that is applied to retrieve the set of objects that satisfies the needs concept above some threshold. These items are then compared to the needs concept to derive their membership of in that concept. Fig. 1 illustrates the general idea of soft interpretation and enveloping, while Fig. 2 shows the main process flow in applications for query answering systems.

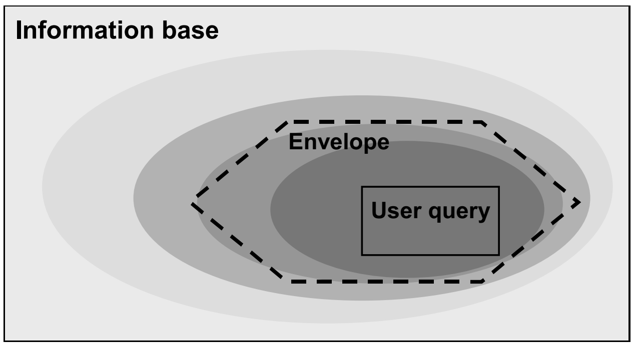

Figure 1: Soft query interpretation with envelope. The graded ovals illustrate the fuzzy needs concept.

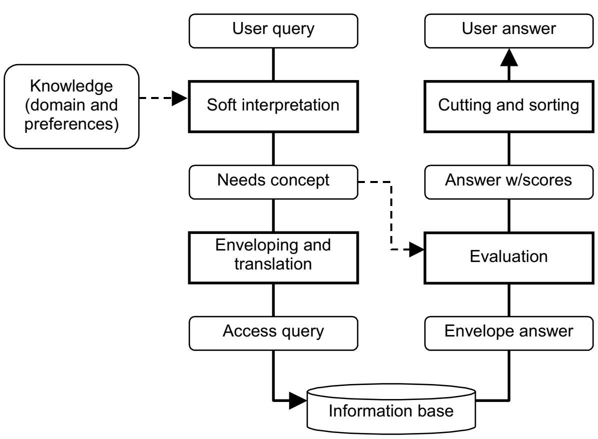

Figure 2: Flow in query answering systems applications of soft interpretation with enveloping.

In Section 2, we recall the characteristics of importance weighting functions for criteria in conjunctive queries. Based on ideas introduced in [10], we then, in Section 3, propose importance-weighted fuzzification schemes for numerical criteria represented by trapezoidal membership functions. In Section 4, we outline a scheme for importance-weighted fuzzification of criteria on non-numerical attributes, *i.e.* attributes whose domain is represented by a set of terms, each terms representing some concept by a word or a phrase. In Section 5, we describe the aggregation of scores of an object in the criteria into the overall score in the needs concept, and characterize the user answer determined by this measure. We conclude in Section 6.

2 Importance weighting of criteria

Consider a user criterion V *is* C where V is a variable for the value of some attribute of the objects represented in the information base, and C is a crisp constraint on the attribute domain D, the set of allowable values of the attribute. A soft interpretation of the criterion V *is* C transforms it into V *is* $\tilde{C}$ where $\tilde{C}$ is the fuzzy subset (of D) of values that are similar to values in C. Thus, $\tilde{C}$ must be a fuzzy superset of C, *i.e.* , for all $x \in D$, $\mu_{\tilde{C}}(x) \geq \mu_C(x)$. Since the degree, to which some value in D is considered similar to (*i.e.* an acceptable replacement for) values in C, is affected of the importance of satisfying the criterion, this importance must be considered in the fuzzification.

We assume that a similarity measure on a domain D is modeled by a fuzzy relation $S(D, D)$, though such a relation is only represented explicitly for non-numerical attributes. The soft interpretations of V *is* C may then be considered as a two-step process. First, C is fuzzified by the initial similarity into $\tilde{C} = \cup_{x \in D}\{S(x, c)/x\}$ where $S(x, c)$, with $c \in C$, is the membership degree of x in $\tilde{C}$. Then, $\tilde{C}$ is modified by importance weight $w \in [0, 1]$ learned for the user.

A mapping $f : [0, 1] \times \mathcal{F}(D) \to \mathcal{F}(D), (w, \tilde{C}) \mapsto \tilde{C}_w$ is defined as an importance weighting transformation for fuzzy criteria in multicriteria queries, if it satisfies:

$$\begin{array}{lll} 1. & \tilde{C}_u \subseteq \tilde{C}_w & \text{if } u > w \\ 2. & \mu_{\tilde{C}_u}(x) \leq \mu_{\tilde{C}_u}(y) & \text{if } \mu_{\tilde{C}_w}(x) < \mu_{\tilde{C}_w}(y) \\ 3. & \tilde{C}_0 = D & \\ 4. & \tilde{C}_1 = \tilde{C} & \end{array} \tag{1}$$

If $g : [0,1]^2 \to [0,1]$ is an importance weighting function for conjunctive queries, *i.e.* if g satisfies [13]:

$$\begin{array}{lll} 1'. & g(u,a) \leq g(w,a) & \text{if } u > w \\ 2'. & g(w,a) \leq g(w,b) & \text{if } a < b \\ 3'. & g(0,a) = 1 & \\ 4'. & g(1,a) = a & \end{array} \tag{2}$$

then the transformation defined by $\mu_{\tilde{C}_w}(x) = g(w, \mu_{\tilde{C}}(x))$ is seen to be an importance weighting transformation as defined by (1). A family of functions satisfying (2) is defined by [4, 5, 10] $g(w,a) = (1-w) \oplus a$, where $\oplus$ is a t-conorm such as the algebraic sum $\oplus_{\text{sum}}$, defined by $a \oplus_{\text{sum}} b = a + b - ab$. Adopting this operator, we obtain $g_1(w,a) = (1-w) \oplus_{\text{sum}} a = (1-w) + a - (1-w)a$, or:

$$g_1(w,a) = 1 - w(1-a) \tag{3}$$

An function satisfying (2) but not in this family, is [11] :

$$g_2(w,a) = a^w \tag{4}$$

3 Numerical criteria

We shall here consider numerical criteria, on the form V *is* C where C is described by an interval $[a,b] \subset \mathbb{R}, a \leq b$. Thus, in this crisp form, the criterion is satisfied completely for all values of V in $[a,b]$, and not satisfied at all for any other value. The criterion is softened into a fuzzy criterion representing the needs criterion through an appropriate extension of C into a fuzzy interval $\tilde{C}$ that, in general, is unbalanced, reflecting a greater flexibility (or acceptability) on the one side of $[a,b]$ than on the other. Fig. 3 illustrates the representation of numerical needs criteria by trapezoidal membership functions. The knowledge on the attribute includes the relative flexibilities α and β ($\alpha, \beta \geq 0$) of, respectively, the left and the right fuzzy parts (spreads) of the fuzzy interval representing the initial softening $\tilde{C}$. We shall assume that α and β characterize the flexibilities for

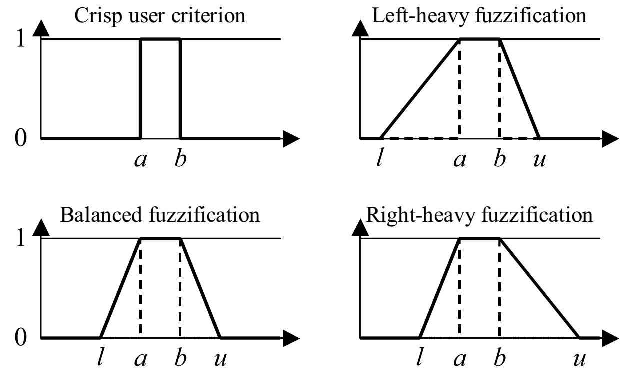

Figure 3: Fuzzification forms for needs criteria

trapezoidal membership function by $\alpha = (a-l)/a$ and $\beta = (u-b)/b$. Hence, the support of $\tilde{C}$ is:

$$\text{supp}(\tilde{C}) = [l,u] = [a(1-\alpha), b(1+\beta)] \tag{5}$$

and $\tilde{C}$ is characterized by the membership function $\mu_{\tilde{C}}$, represented by the trapeze $\phi(l,a,b,u)$, as follows:

$$\mu_{\tilde{C}}(x) = \begin{cases} 1 & a \leq x \leq b \\ \frac{x-l}{a-l} & l \leq x \leq a \\ \frac{u-x}{u-b} & b \leq x \leq u \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Let us first consider the left spread, $[l,a]$. By (5) and (6) we have:

$$\mu_{\tilde{C}}(x) = \frac{x - a(1-\alpha)}{a - a(1-\alpha)} = \frac{x - a(1-\alpha)}{a\alpha} \tag{7}$$

Through weighting $\mu_{\tilde{C}}(x)$ by the importance $w \in [0,1]$, applying the importance weighting function g_1 (3), $\tilde{C}$ is transformed by $\mu_{\tilde{C}_w}(x) = g_1(w, \mu_{\tilde{C}}(x))$:

$$g_1(w, \mu_{\tilde{C}}(x)) = 1 - w(1 - \frac{x - a(1-\alpha)}{a\alpha}) \tag{8}$$

We shall now determine $\text{supp}(\tilde{C}_w) = [l_w, u_w]$ as follows. Since l_w is seen to be the value of x for which $\mu_{\tilde{C}}(x) = 0$, it is found as solution to that equation, namely $l_w = a(1-\alpha/w)$. Similarly, we find $u_w = b(1+\beta/w)$. Hence, for $w \in]0,1]$:

$$\text{supp}(\tilde{C}_w) = [a(1 - \frac{\alpha}{w}), b(1 + \frac{\beta}{w})] \tag{9}$$

Thus, $\mu_{\tilde{C}_w}$ is represented by $\phi(l_w, a, b, u_w)$, and defined by (6) for $l = l_w$ and $u = u_w$. Comparing (9) to (5), we see that importance weighting with the weight w modifies α and β by the factor $1/w$. It is easily verified that this transformation is an importance weighting transformation for fuzzy criteria (1). In referring to an object ω in an information base Ω, we shall use the notation $\mu_{\tilde{C}_w}(\omega)$ for satisfaction of $\tilde{C}_w$ by the value of the object attribute constrained by $\tilde{C}_w$.

We have above assumed that the relative flexibilities α and β represent the case for $w = 1$. However, since α and β are derived from experts or through data mining, they are likely to reflect an importance, $w_d \in]0, 1]$, typically attached to criteria on the attribute, and thus implicitly applied as default by the initial softening. We may then take account for this in the model above by replacing in (9) the factor $1/w$ of each of the relative flexibilities by w_d/w.

For the access query, we obtain an envelope of $\tilde{C}_w$, as the δ-cut $\tilde{C}_w^{(\delta)}$, $\delta \in]0, 1]$, defined by: $\mu_{\tilde{C}_w^{(\delta)}}(x) = 1$ if $\mu_{\tilde{C}_w}(x) \geq \delta$, and 0 otherwise. Hence,

$$\tilde{C}_w^{(\delta)} = [l_w + \delta(a - l_w), u_w - \delta(u_w - b)]$$

Fig. 4 illustrates the envelope for a choice of δ.

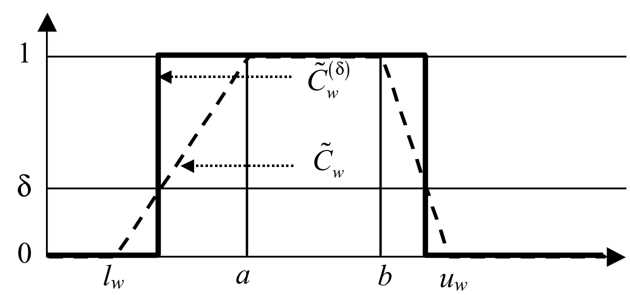

Figure 4: The δ-cut of $\tilde{C}_w$, denoted $\tilde{C}_w^{(\delta)}$

4 Representation of non-numerical needs criteria

Typically, non-numerical attribute domains have the form of a set (or vocabulary) of terms, each term representing some concept that may be applied for the description of the attribute. For instance the, attribute "kind of location" of a house, may apply terms such as "city", "suburb", and "rural". Terms are semantically or pragmatically associated to other terms through various kinds of relationships. The relationships are often vague, *i.e.* only applying to some degree, and are therefore modeled by fuzzy relations. The key issue here is that the information base may contain objects that are of interest to the user but apply other terms than those used by the user. Such objects can only be related to the user query, if the system has the relevant knowledge on the relationships, and can apply it in searching the information base. For our purpose, the relevant knowledge is the knowledge on, for all pair of terms $(t_i, t_j) \in D \times D$, the degree to which t_i may replace t_j in objects of interest to the user. This knowledge is naturally represented by a fuzzy relation $S(D, D)$ that is reflexive and transitive, and, in general, asymmetric. The basic scheme for applying such relations in similarity based fuzzy query evaluation was introduced in [9].

Consider a criterion V *is* C with $C = t$, $t \in D$. Then, the fuzzy set $\tilde{C}$ of terms that are similar to (or may replace) t is defined by $\tilde{C} = \cup_{x \in D}\{S(x, t)/x\}$. Let $D = \{t_1, \ldots, t_m\}$ and $s_{ij} = S(t_i, t_j)$, then $\tilde{C}$ is described by $\tilde{C} = \cup_{i \in \mathrm{N}_m}\{s_{ij}/t_i\} = s_{1j}/t_1 + \cdots + s_{mj}/t_m$, where $\mathrm{N}_m = \{1, \ldots, m\}$.

Applying in this case the function g_2 (4), we have $g_2(w, s_{ij}) = s_{ij}^w$. Hence, $\tilde{C}_w = \cup_{i \in \mathrm{N}_m}\{s_{ij}^w/t_i\}$. The δ-cut of $\tilde{C}_w$, $\tilde{C}_w^{(\delta)} = \cup\{1/t_i \mid i \in \mathrm{N}_m, s_{ij}^w \geq \delta\}$, is seen to be the $\delta^{\frac{1}{w}}$-cut of $\tilde{C}$, *i.e.* :

$$\tilde{C}_w^{(\delta)} = \tilde{C}^{(\delta^{\frac{1}{w}})} = \cup\{1/t_i \mid i \in \mathrm{N}_m, s_{ij}^w \geq \delta^{\frac{1}{w}}\}$$

which translates to the envelope :

$$\bigvee_{t \in \tilde{C}^{(\delta^{\frac{1}{w}})}} (V = t)$$

where $\vee$ is the Boolean OR. The satisfaction of a needs criterion $\tilde{C}_w$ by an object represented in the information base is defined as follows. Let us consider the general case where term attribute values may be represented by multiple terms each attached the degree to which the term apply. Consider for instance a house with the attribute 'kind of location' characterized by $(0.9/\text{rural}, 0.7/\text{mountain})$, telling that the location is rural to degree 0.9 and mountain to the degree 0.7. Let, in general, the object attribute value considered be characterized by $(d_1/u_1, \ldots, d_p/u_p)$, $p \leq m$, $m = \|D\|$. Then the satisfaction of $\tilde{C}_w$ by ω is defined by the product [8]:

$$\mu_{\tilde{C}_w}(\omega) = \begin{array}{l}(d_1/u_1 + \cdots + d_p/u_p) \\ \times(s^w_{1j}/t_1 + \cdots + s^w_{mj}/t_m)\end{array} \quad (10)$$

Let i' denote the index of the term u_i in the term domain D. Then the evaluation of (10) is defined by:

$$\mu_{\tilde{C}_w}(\omega) = \oplus_{i=1}^{p}(d_{i'} \otimes s^w_{i'j}) \quad (11)$$

where $\oplus$ is a t-conorm, like the max operator, and $\otimes$ is t-norm, like the min operator, or, better for this purpose, some Archimedean t-conorm and t-norm, such as the following (in the Dombi family [3], for the parameter value 1): $a \oplus b = 1$, if $a = b = 1$, and $(a+b-2ab)/(1-ab)$ otherwise, and $a \otimes b = 0$, if $a = b = 0$, and $ab/(a+b-ab)$ otherwise.

We notice that the cases where no degrees are applied for terms in object attribute values, or object attribute values are singletons, are obtained as special cases of (11).

5 Aggregation of criterion scores

The the overall satisfaction of a needs concept derived from a multicriteria query is defined by:

$$\mu_{\tilde{Q}_w}(\omega) = \otimes_{i=1}^{n}\mu_{(\tilde{C}_w)_i}(\omega) \quad (12)$$

where $\mu_{(\tilde{C}_w)_i}(\omega)$ is the satisfaction of the i'th needs criterion in the n-criteria needs concept, and $\otimes$ is a t-norm operator.

Though a t-norm aggregation in a sense is correct for a multicriteria problem, it is often appears too "restrictive" for query answering systems; thus, if just one of the arguments is zero, then the aggregated value is zero, regardless of the values of the other arguments. Hence, we would like to relax the aggregation more than the least restrictive t-norm, namely the min operator, representing a "pure" AND. A family of aggregation operators that offers such a property is the Ordered Weighting Averaging (OWA) operators, introduced by Yager [12]. A member of the OWA family is characterized by a tuple of positional weights, called the OWA weights, $\vec{v} = (v_1, \ldots, v_n)$, $v_i \in [0,1]$, $\sum_{i=1}^{n} v_i = 1$. Two properties the OWA operator are defined as measures on $\vec{v}$, namely the orness (the degree to which the operator represents a pure OR), $\text{orness}(\vec{v}) = \frac{1}{n-1}\sum_{i=1}^{n}((n-i)v_i)$, and the dispersion (measuring the degree to which all arguments are utilized in the aggregation), $\text{disp}(\vec{v}) = -\sum_{i=1}^{n}((v_i \ln v_i)$. The andness is defined by $\text{andness}(\vec{v}) = 1 - \text{orness}(\vec{v})$. The OWA aggregation of n arguments $(a_1, \ldots, a_n)$, $a_i \in [0,1]$ is defined as the OWA weighted sum of the arguments in non-increasing value order.

An importance weighted OWA aggregation is obtained as follows [7]. Let a_i be the satisfaction of the i'th criterion $\tilde{C}_i$ *without* importance weighting, *i.e.* $a_i = \mu_{\tilde{C}_i}(\omega)$. Let $\vec{w} = (w_1, \ldots, w_n)$, $w_i \in [0,1]$, $\max_{i=1}^{n} w_i = 1$, be the importance weights, such that w_i is the importance of satisfying the i'th criterion. Then, the overall satisfaction of $\tilde{C}_w$ by an object ω is defined by the importance weighted OWA aggregation as follows:

$$\mu_{\tilde{Q}_w}(\omega) = \sum_{i=1}^{n}(v_i b_{P(i)}) \quad (13)$$

where b_i is the importance weighted satisfaction of i'th criterion, defined by $b_i = \rho + w(b_i - \rho)$, $\rho = \text{andness}(\vec{w})$, and P is a permutation of $(1, \ldots, n)$ such that $b_{P(1)} \geq b_{P(2)} \geq \ldots b_{P(n)}$, *i.e.* the b_i's in non-increasing value order. For a given andness ρ, an OWA weighting tuple $\vec{v}$ with $\text{andness}(\vec{v})$ close to ρ, and with a good dispersion, can be shown to be obtained by $v_i = (\frac{i}{n})^\gamma - (\frac{i-1}{n})^\gamma$, $i = 1, \ldots, n$, with $\gamma = \frac{\rho}{1-\rho}$.

The user answer, *i.e.* the answer presented to the user, is obtained as a cut of the fuzzy set $\tilde{Q}_w$ defined by (13), characterized by a threshold $\zeta \in [0,1]$ and a size (crisp cardinality) $\kappa \in$ N. We define the (ζ, κ) level user answer as the θ-cut of $\tilde{Q}_w$, $\tilde{Q}_w^{(\theta)}$, with $\theta = \max(\zeta, \sigma_\kappa)$ where σ_κ is the satisfaction of the κ'th most satisfying object ω. The objects ω in $\tilde{Q}_w^{(\theta)}$ are for the user answer represented by a list of $(\mu_{\tilde{Q}_w}(\omega), \omega)$ pairs, ordered ordered in non-increasing order after the value of $\mu_{\tilde{Q}_w}(\omega)$.

Finding a good envelope for $\tilde{Q}_w$, that is an envelope that will retrieve $\tilde{Q}_w^{(\theta)}$ and not much more, is not a trivial problem. A simple practical solution is to apply the heuristic envelope defined as the ORing of the criterion envelopes obtained as the δ-cut of the $\tilde{C}_w$'s with $\delta = \zeta$.

6 Conclusion

We have presented a scheme for modeling and utilizing knowledge on user preferences and similarities in attribute domains for soft interpretation of multicriteria queries. The knowledge is applied to soften the user query into the needs concept, representing the user's information needs by fuzzy sets. The degree to which an object represented in the information base satisfies the needs concept is defined as the object's membership degree in the needs concept. Numerical and non-numerical (term) attributes are handled in a common framework. This allows us to apply the scheme for information bases with any mix of the two kinds of attributes, such as administrative database (mostly numerical attributes), document and text bases (mostly term attributes), and product and catalogue base (often with several attributes of each kind).

The application framework considered is information agents that learn user preferences from observing the user's interaction activity, and utilize this knowledge to achieve more accuracy in their interpretation of queries from the user.

References

[1] Bosc, P. and Pivert, O.: Fuzzy querying in conventional databases. In J. Kacprzyk and L. Zadeh, Eds., *Fuzzy Logic for the Management of Uncertainty,* John Wiley (1992), 645-671.

[2] Bosc, P. and Pivert, O.: Extending SQL Retrieval Features for Handling of Flexible Queries. In Dubois, D., Prade H., and Yager, R.R., Eds., *Fuzzy Set methods in Information Engineering: A Guided Tour of Applications,* John Wiley & Sons, pp. 233-251, 1997.

[3] Dombi, J.: A general class of fuzzy operators, the De Morgan class of fuzzy operators and fuzziness measures induced by fuzzy operators. *Fuzzy Sets and Systems* **8**(2): 149-163 (1982).

[4] D. Dubois, D., and Prade, H.: Weighted minimum and maximum operations in fuzzy set theory, *Information Sciences* **8**: 205-210, 1986.

[5] Yager, R.R.: A note on weighted queries in information retrieval systems, *Journal of the American Society of Information Sciences* **38**:23-24, 1987.

[6] Kacprzyk, J. and Zadrozny, S.: Fuzzy Querying in Microsoft Access V.2. In Dubois, D., Prade H., and Yager, R.R., Eds., *Fuzzy Set methods in Information Engineering: A Guided Tour of Applications,* John Wiley & Sons, pp. 223-232, 1997.

[7] Larsen, H.L.: Importance weighted OWA aggregation of multicriteria queries. *Proceeding of the North American Fuzzy Information Processing Society conference, New York, 10-12 June 1999 (NAFIPS'99)*, pp. 740-744.

[8] Larsen, H.L , Nilsson, J.F.: Fuzzy Querying in a Concept Object Algebraic Datamodel. In Christiansen, H., Andreasen, T., Larsen, H.L., Eds., *Flexible Query Answering Systems*, Kluwer Academic Publishers, 1997, pp. 123–139.

[9] Larsen, H.L. and Yager, R.R. The use of fuzzy relational thesauri for classificatory problem solving in information retrieval and expert systems. *IEEE J. on System, Man, and Cybernetics* **23**(1):31-41, 1993.

[10] Larsen, H.L. and Yager, R.R.: Query Fuzzification for internet Information retrieval. In Dubois, D., Prade H., and Yager, R.R., Eds., *Fuzzy Set methods in Information Engineering: A Guided Tour of Applications,* John Wiley & Sons, pp. 291-310, 1997.

[11] Yager, R.R.: Multiple objective decision making using fuzzy sets. *IEEE J. on System, Man, and Cybernetics* **9**:375-382, 1977.

[12] Yager, R.R.: On ordered Weighted aggregation operators in multi-criteria decision making. *IEEE J. on Systems, man, and Cybernetics* **18**: 183-190, 1988.

[13] Yager, R.R.:*On weighted media aggregation.* Technical report #MII-1317, Machine Intelligence Institute, Iona College, New Rochelle, NY (1993).