Issues in construction of linguistic summaries

Miroslav Hudec[1]

Abstract. *Linguistic summaries are convenient approach for revealing intensity of relational knowledge in the data. Two main parts of summaries are summarizers defined as predicates and quantifiers. The validity of a rule critically depends on constructed fuzzy sets for predicates and quantifiers. This paper deals with the construction of membership function for predicates from the current content of a data set and the construction of membership function for quantifiers in the [0, 1] interval. The second aim is building complex summaries. Moreover, linguistic summaries can be used as flexible queries for ranking entities on higher hierarchical level using data on lower hierarchical level.*

Keywords**:** *linguistic quantifiers, linguistic summarizers, construction of membership functions, fuzzy queries.*

1 Introduction

Linguistic summaries are able to express relational knowledge and its intensity about collected data. For people linguistic summarisation is a desirable way how to communicate in natural language and obtain validity of uncertain rules from a data set. Linguistic summaries are of well-known structure *Q entities in database are (have) S* where *S* is a summarizer defined as linguistic term on the domain of examined attribute and *Q* is a fuzzy quantifier in sense of Zadeh (1983). An example of simple linguistic summary is: *most customers are middle aged*. Linguistic summaries could be more complex e.g. *most highly situated (altitude above the sea level) and small municipalities have high unemployment and small migration*.

A linguistic summary is a short sentence that describes relational knowledge in large data sets. The concept of linguistic summaries has been initially introduced in (Yager, 1982) and further developed in (Rasmussen and Yager, 1997; Kacprzyk and Yager, 2001; Kacprzyk and Zadrozny, 2009). Truth value of summaries is usually called validity and gets values from the [0, 1] interval by agreement. Data summarization is one of basic capabilities needed to any “intelligent” system (Kacprzyk and Zadrozny, 2009). In order to use advantages of the Structured Query Language (SQL) and linguistic summaries Rasmussen and Yager (1997) have created the SummarySQL language. FQUERY for Access (Kacprzyk and Zadrozny, 2009) makes possible to use fuzzy terms in usual fuzzy queries and for summarisation.

Galindo (2008) concluded that when the system uses badly defined membership functions, it will not work properly. So, these functions have to be carefully defined. In the same way this holds for linguistic summaries, because it is required to calculate the proportion of entities that satisfies (fully or partially) the summarizer *S* and validity of a rule.

This paper is focused on developing linguistic summaries by dynamically constructing fuzzy sets for summarizers *S* from the current database content applying results of Tudorie (2008) and Hudec and Sudzina (2012) and defining quantifiers. Section 2 describes the concept of linguistic summaries. Section 3 is devoted to construction of membership functions for predicates and quantifiers. Short illustrative examples are provided in Section 4. Section 5 examines further development of summarizers by preferences. Finally, conclusions are drawn in Section 6.

[1] Faculty of Economic Informatics, University of Economics in Bratislava, Slovakia

2 Linguistic summaries by fuzzy queries

According to Zadrożny and Kacprzyk (2009) an imprecise (fuzzy) query is a query containing natural language expressions, referred to as linguistic terms, to specify:

a) imprecise values e.g. *low salary*;
b) imprecise comparison operators e.g. *salary much greater than 2 000*;
c) non-standard aggregation scheme of the fulfilment degrees to partial conditions e.g. *most of municipalities have small migration*.

In this paper we are focused on the third issue. Nevertheless, evaluation of imprecise values in query conditions is the basis for the linguistic summaries.

Because for the humans the usual means of communication is natural language, an uncertain proposition (linguistic summary) would be desirable way to express relational knowledge about the data (Kacprzyk and Zadrożny, 2009; Rasmussen and Yager, 1997).

2.1 Linguistic summaries for extracting relational knowledge

Examples of linguistic summaries are as follows:

(a) Few municipalities have high altitude;
(b) Most municipalities have high unemployment and small migration;
(c) Most low polluted municipalities have high altitude and small number of inhabitants.

Linguistically quantified propositions are written in a general form:

$$Qx(Px) \tag{1}$$

where Q is a linguistic quantifier, $X = \{x\}$ is a universe of disclosure (e.g. the set of all municipalities) and *P(x)* is a predicate depicting summariser *S* e.g. *small migration*. Predicate *P* is a fuzzy set $P \in \mathcal{F}(X)$. where $\mathcal{F}(X)$ is a family of fuzzy sets defined on the domain of an examined variable.

The truth value of a statement (rule) is computed by the following equation (Zadrożny and Kacprzyk, 2009):

$$T(Qx(Px)) = \mu_Q(\frac{1}{\mathrm{n}}\sum_{i=1}^{n}\mu_{\mathrm{P}}(x_i)) \tag{2}$$

where n is the cardinality of a data set (number of entities), $\frac{1}{\mathrm{n}}\sum_{i=1}^{n}\mu_{\mathrm{P}}(x_i)$ is the proportion of objects in a data set that satisfy *P(x)* and μ_Q is the membership function of a quantifier.

Measure of validity can be calculated by quantifiers defined in Zadeh (1983) or using the OWA (Ordered Weighted Averaging) operator (Yager and Kacprzyk, 1997; Zadrożny and Kacprzyk, 2009). In this paper the former approach is used.

First type of summary is of the form:

Q entities in database are (have) S

Examples are rules (a) and (b) where summary (b) consist of two elementary conditions connected with the *and* aggregation operator. If summarizer consists of several atomic predicates $\mu_{\mathrm{P}}(x_i)$ is calculated in the following way:

$$\mu_{\mathrm{P}}(x_i) = f(\mu_{Pj}(x_i)) \tag{3}$$

where P_j is the *j*-th atomic predicate and *f* is either t-norm or t-conorm. The truth value of a statement (rule) is computed by the eq. (2).

Second type of summary is of the form:

Q R entities in database are (have) S

The example is the rule (c). The procedure for calculating truth value has the following form (Rassmusen and Yager, 1997):

$$T(Qx(Px)) = \mu_Q\left(\frac{\sum_{i=1}^{n} t(\mu_{\mathrm{P}}(x_i), \mu_{\mathrm{R}}(x_i))}{\sum_{i=1}^{n} \mu_{\mathrm{R}}(x_i)}\right) \qquad (4)$$

where $\frac{\sum_{i=1}^{n} t(\mu_{\mathrm{P}}(x_i), \mu_{\mathrm{R}}(x_i))}{\sum_{i=1}^{n} \mu_{\mathrm{R}}(x_i)}$ is the proportion of the *R* objects in a database that satisfy *S*, *t* is a t-norm, μ_Q is the membership function of a quantifier. The same discussion as for $\mu_{\mathrm{P}}(x_i)$ (eq. 3) applies for $\mu_{\mathrm{R}}(x_i)$.

2.2 Linguistic summaries for flexible queries

An example of query is the following *select regions where most of municipalities have small unemployment and low migration*. In the first step, validity of summaries is calculated for each region. In the second step regions are ranked downwards starting with region having the highest value of the rule validity.

The procedure for calculating validity of summary *Q entities in database are (have) S* for each data cluster (group) is created as the extension of (2):

$$T_i(Qx(Px)) = \mu_Q\left(\frac{1}{\mathrm{N_i}}\sum_{j=1}^{N_i} \mu_{\mathrm{P}}(x_{ji})\right), \quad i = 1,\ldots,R, \quad \sum_{i=1}^{R} \mathrm{N_i} = n \qquad (5)$$

where *n* is the number of entities in whole database, N_i is the number of entities in cluster *i* (municipalities in region *i*), *R* is the number of clusters in a database (e.g. regions), T_i is validity of rule for *i*-th cluster, and $\mu_p(x_{ji})$ is the proportion of objects in *i*-th cluster that satisfy summarizer *S*.

The procedure for calculating validity of summary *Q R entities in database are (have) S* for each cluster is created using the extension of (4):

$$T_i(Qx(Px)) = \mu_Q\left(\frac{\sum_{j=1}^{N_i} t(\mu_{\mathrm{S}}(x_{ji}), \mu_{\mathrm{R}}(x_{ji}))}{\sum_{j=1}^{N_i} \mu_{\mathrm{R}}(x_{ji})}\right) \quad , i = 1\ldots R, \quad \sum_{i=1}^{R} \mathrm{N_i} = n \ . \qquad (6)$$

The meaning of variables is the same as in (4, 5).

3 Construction of membership functions for predicates and quantifiers

The matching degree of each database record to query condition critically depends on constructed membership functions of predicates. Therefore, these functions have to be carefully constructed. The same holds for quantifiers. In the paper we are focused on relative quantifiers *most*, *about half* and *few*.

3.1 Construction of membership functions for predicates

Let D_{min} and D_{max} be the lowest and the highest domain values of attribute A i.e. Dom(A) = [D_{min}, D_{max}] and L and H be the lowest and the highest values in the current content of a database respectively (Hudec and Sudzina, 2012). Usually attribute's domain is defined in a way that all theoretically possible values could be stored. In practice, collected data are often far from the values of D_{min} and D_{max}; that is, $[L, H] \subset [D_{min}, D_{max}]$ (either $[D_{min}, L]$ or $[H, D_{max}]$ are empty or even both of them are empty). This fact should be considered in linguistic summaries.

The uniform domain covering method (Tudorie, 2008) is an appropriate method for construction of membership functions for these tasks. At the beginning, values of L and H are obtained from the current database content. The length of fuzzy set core β and the slope α (Figure 1) are calculated using the following equations (Tudorie, 2008):

$$\alpha = \frac{1}{8}(H - L)\,, \tag{7}$$

$$\beta = \frac{1}{4}(H - L)\,. \tag{8}$$

Required parameters A, B C and D (Figure 1) are calculated using (7, 8):
$A = L + \beta$; $B = L + \beta + \alpha$; $C = H - \beta - \alpha$; $D = H - \beta$.

The uniform domain covering method is adequate because the main goal is to reveal relational dependencies among data where distribution of stored data should be reflected in the membership functions.

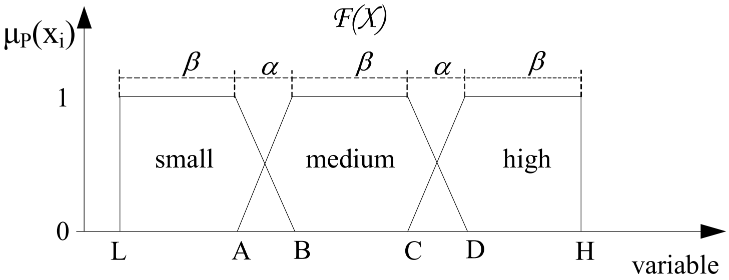

Figure 1: Linguistic and crisp domain of an attribute

3.2 Construction of membership functions for quantifiers

The validity of summaries examined in the paper is computed by relative quantifiers. A quantifier is constructed by a fuzzy set on the [0, 1] interval (Zadrożny and Kacprzyk, 2009). For compatibility with the construction of predicates, explained in Section 3.1, the [0, 1] interval is marked as the domain of a family of quantifiers.

For a regular non-decreasing quantifier (e.g. *most*) its membership function should meet the following property:

$$x \le y \Rightarrow \mu_Q(x) \le \mu_Q(y)\,; \mu_Q(0) = 0; \quad \mu_Q(1) = 1 \ . \tag{9}$$

The quantifier might be given as (Kacprzyk and Zadrożny, 2009):

$$\mu_Q(y) = \begin{cases} 1, & for \ \ y > 0.8 \\ 2y - 0.6 & for \ \ 0.3 \le y \le 0.8 \\ 0, & for \ \ y < 0.3 \end{cases} \quad . \tag{10}$$

The second way for modelling a linguistic quantifier is realised by the OWA operator. If quantifier is a regular non-decreasing (9) then the weight vector of an OWA operator is defined in the following way (Yager, 1988):

$$w_i = \mu_Q(\frac{i}{m}) - \mu_Q(\frac{i-1}{m}), \ \ i = 1,...,m \ . \tag{11}$$

The first approach is appropriate for ordinal summaries e.g. *most of municipalities have small gas consumption*. Number of municipalities that meet the predicate to some extent could be high (value of *m* in eq. 11) and it is time consuming to calculate all values of w_i for such a long vector. In this case (10) is a rational option.

Having an “aggregated” linguistic quantifier e.g.: *most of the predicates {P$_i$} are satisfied (i=1… n)* then the quantifier could be represented by the OWA operator using (11). Number of predicates is significantly smaller than number of entities in a database.

Equivalently, non-increasing quantifier e.g. *few* could be created as a “mirror picture” of (10) in the following way:

$$\mu_Q(y) = \begin{cases} 1, & for \ \ y < 0.2 \\ 1.4 - 2y & for \ \ 0.2 \le y \le 0.7 \\ 0, & for \ \ y > 0.7 \end{cases} \quad . \tag{12}$$

Having in mind the uniform domain method for construction of family of membership functions $\mathcal{X}$*(X)* on domain of attribute for summarizers (Section 3.1) we can create the family of membership functions $\mathcal{Q}$ for quantifiers in the same way depicted in Figure 2.

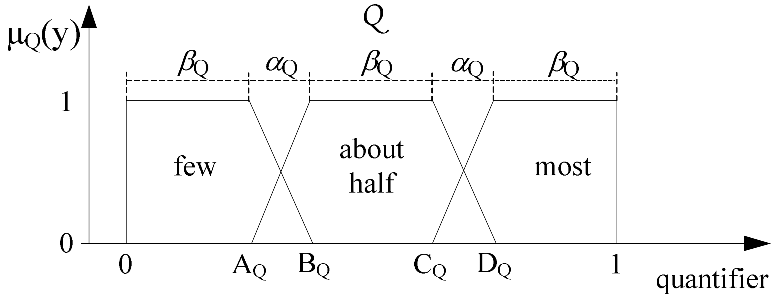

Figure 2: The domain for quantifiers

The length of fuzzy set core *β* and the slope *α* are calculated using (7) and (8). In this case the values are following:

$$\alpha_Q = \frac{1}{8}, \ \beta_Q = \frac{1}{4}, \ A_Q = 0.25, \ B_Q = 0.375, \ C_Q = 0.625, \ D_Q = 0.75 \ . \tag{13}$$

Applying (13), parameters of the quantifier *most* are calculated in the following way:

$$\mu_Q(y) = \begin{cases} 1, & for \;\; y > 0.75 \\ \dfrac{y}{0.125} - 5 & for \;\; 0.625 \le y \le 0.75 \\ 0, & for \;\; y < 0.625 \end{cases} \quad . \tag{14}$$

In this way quantifier is more restrictive than quantifier defined in (10). On the other hand, quantifiers are uniformly distributed in their domain (Figure 2).

Comparison of both approaches is depicted in Figure 3. The increasing part of the quantifier *most* in (10) starts earlier and inclines slower in comparison with (14). The core of (10) is shorter than for (14). In addition, intersection of fuzzy sets most and few defined by (13) is empty because these quantifiers are opposite and there is also the space for the quantifier *about half* which has overlapping boundaries with quantifiers *most* and *few*.

Presumably, the following question might appear: which approach for the quantifiers' construction is more appropriate? The discussion is provided in Section 4.

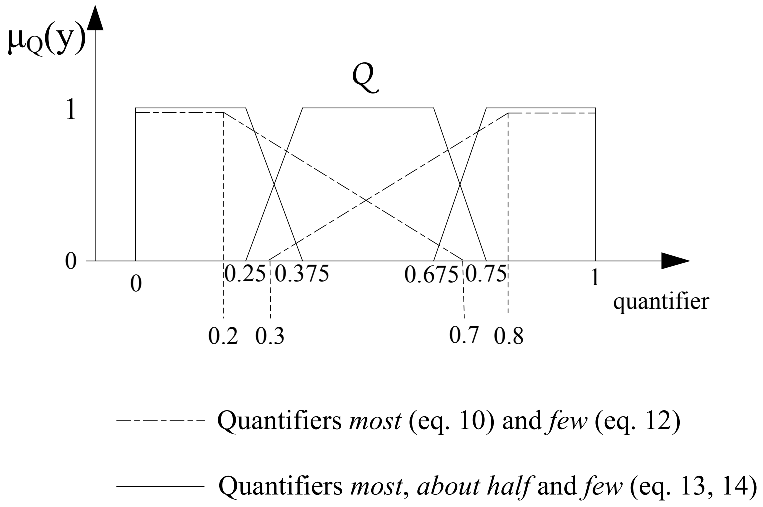

Figure 3: Comparison of definitions of quantifiers.

Moreover, if we want to extend family of fuzzy sets $\mathcal{X}(X)$ to five: *very small*, *small*, *medium*, *high* and *very high* we need only to divide the domain to five linguistic terms uniformly distributed (Tudorie, 2009). The same could hold for the family of quantifiers $\mathcal{Q}$

4 Illustrative examples

This section is devoted to small examples of both kind of summaries examined in the Section 3.

4.1 Summaries for extraction relational knowledge in the data

In the following three short examples quantifiers defined in (10) and (14) were evaluated.

Example 1. Let's have 10 entities of which 6 fully meet the summarizer (value of 1), 3 partially do with values of 0.9, 0.7 and 0.2 respectively and one record does not meet the condition (value of 0) then the proportion of objects in a data set that satisfy *P(x)* obtains the value of 0.78. The validity of rule calculated by (10) is 0.96 and by (14) is 1.

Example 2. Let's have 10 entities with the following membership degrees to the summarizer 0; 0; 0; 0.4; 0.4; 0.4; 0.4; 0.5; 0.5; 1, then the proportion of objects in a data set that satisfy *P(x)* obtains the value of 0.36. The validity of rule calculated by (10) is 0.12 and by (14) is 0.

Example 3. Let's have again 10 entities with the following membership degrees to the summarizer 0; 0.3; 0.4; 0.6; 0.8; 0.9; 1; 1; 1; 1, then the proportion of objects in a data set that satisfy *P(x)* obtains the value of 0.7. The validity of rule calculated by (10) is 0.8 and by (14) is 0.6.

Results are more or less similar. However, partially belonging of value 0.36 to the quantifier *most* (10) even with small value is disputable.

Additional analysis of both approaches is required. Therefore, it is a topic for the further research. Anyway, user opinion of the strength of the quantifiers should be incorporated into the construction of quantifiers' membership functions.

4.2 Summaries as flexible queries

For example we want to know to which extent is the following rule (query) satisfied *most of municipalities has small attitude above sea level*. The result for all eight regions of the Slovak Republic is presented in Table 1 (Hudec, 2013). Table 1 shows that regions Bratislava, Trnava and Nitra are flat whereas regions Žilina and Prešov are hilly. Region Trenčín is more flat than hilly. The same holds for region Košice but it is a slightly hillier than region Trenčín. Data about municipalities were used for summaries but result is visible for regions only ranked according to value of rule validity.

Table 1: Linguistic summary for each region

Region	**Validity of the summary**
Bratislava	1
Trnava	1
Nitra	1
Trenčín	0.7719
Košice	0.6314
Bánska Bystrica	0.2116
Žilina	0
Prešov	0

5 Further improvements of linguistic summaries

In summarizers not all elementary predicates always have the same importance. The aim of preferences is to distinguish elementary conditions according to their importance inside the overall summarizer.

Applying preferences linguistic summaries become more sophisticated covering additional class of problems e.g. *most of municipalities have high altitude above sea level and low pollution where the second condition is more important than the first one*. In order to calculate validity of the rule weights should be associated with each elementary condition.

This issue could be solved by appropriate fuzzy implications (Zadrożny et al, 2008). The idea how to calculate the matching degree of an elementary predicate *Pi* according to an importance weight w_j and fuzzy implications has the following form (Zadrożny et al, 2008):

$$\mu(P_j^*, x_i) = (w_j \Rightarrow \mu(P_j, x_i)) \qquad (15)$$

where $\Rightarrow$ is a fuzzy implication, P_j is j-th elementary predicate and x_i, is i-th entity in database which meet the summarizer. In order to be meaningful, weights should satisfy several requirements (Dubois and Prade, 1997). One of them is the following:

if w_i=0 then result should be such as if P_i does not exist.

Applying this requirement shows us that Mamdani implication is not adequate whereas Kleene-Dienes, Godel and Gougen implications match this requirement. Examples of the first two implications are briefly outlined below.

The Kleene-Dienes implication has the following structure:

$$\mu(P_j^*, x_i) = \max(\mu(P_j, x_i), 1 - w_j) \ . \qquad (16)$$

Apparently, for small importance of P_j (w_j is close or equal to 0), the satisfaction of atomic predicate P_j has a very small influence moving to no influence on the query satisfaction ($w_j \to 0 \Rightarrow \mu(P_j^*, x_i) \to 1$). In another case when w_j is close to 1, the satisfaction of P_j is essential for the satisfaction of the overall condition ($w_j \to 1 \Rightarrow \mu(P_j^*, x_i) \to \mu(P_j, x_i)$).

Contrary, the Mamdani implication is not suitable for this approach. It can be shown on the following example:

$$\mu(P_j^*, x_i) = \min(\mu(P_j, x_i), w_j) \ . \qquad (17)$$

Because of the small importance of w_i the overall matching degree is close to 0. In case when w_i=0, the overall matching is 0 regardless of other elementary conditions. It implies that the requirement *if w_i=0 then result should be such as if Pi does not exist* is not satisfied for the implication (17).

The proportion of objects in a database that satisfy *P(x)* applying the Kleene-Dienes implication (16) is calculated in the following way:

$$\frac{1}{\mathrm{n}} \sum_{i=1}^{n} t(\max_{j=1,\dots,N}(\mu(P_j, x_i), 1 - w_j)) \ . \qquad (18)$$

Finally, validity of rule is expressed by the equation:

$$T(Qx(Px)) = \mu_Q(\frac{1}{\mathrm{n}} \sum_{i=1}^{n} t(\max_{j=1,\dots,N}(\mu(P_j, x_i), 1 - w_j))) \quad . \qquad (19)$$

6 Conclusion

The paper demonstrates how we can start with a simple linguistic summary and build more complex summaries. Although fuzzy set theory has been already established as an adequate framework to deal with linguistic summaries, there is still space for improvements. The critical parts are construction of membership functions for linguistic terms (summarizers) *small*, *medium*, *high* and construction of relative quantifiers *few*, *about half*, *most*. The former can be satisfactorily solved if we calculate parameters of membership functions directly from the current database content using the uniform domain covering method. The later can be satisfactorily solved if we calculate parameters of relative quantifiers in the [0, 1] interval by the same method as for summarizers. Finally, summarizers were extended by preferences described as fuzzy implications.

Relational knowledge about the data is valuable either for decision making or for broad audience. Both of them usually are not interested in data itself but in the relational knowledge that could support decision making or can satisfy their curiosity.

References

Dubois, D. and Prade, H. (1997). Using fuzzy sets in flexible querying: Why and how? In: Andreasen, T., Christiansen, H. and Larsen H.L. (Eds.) Flexible Query Answering Systems (pp. 45-60). Kluwer Academic Publishers, Dordrecht.

Galindo, J. (2008). Introduction and Trends to Fuzzy Logic and Fuzzy Databases. In: Galindo J. (Ed.) Handbook of Research on Fuzzy Information Processing in Databases (pp. 1-33). IGI Global, London.

Hudec, M. (2013). Applicability of Linguistic summaries. In: 11th Balkan conference on operational research. Belgrade. Accepted for publication.

Hudec, M. and Sudzina, F. (2012). Construction of fuzzy sets and applying aggregation operators for fuzzy queries. In: 14th International Conference on Enterprise Information Systems (ICEIS 2012). Wroclav.

Kacprzyk, J. and Zadrożny S. (2009). Protoforms of Linguistic Database Summaries as a Human Consistent Tool for Using Natural Language in Data Mining. *Internationbal Journal of Software Science and Computational Intelligence*, 1: 1-11.

Kacprzyk, J., and Yager, R. (2001). Linguistic summaries of data using fuzzy logic. *International Journal of General Systems*, 30: 33–154.

Rasmussen, D. and Yager, R. (1997). Summary SQL - A Fuzzy Tool for Data Mining. *Intelligent Data Analysis*, 1: 49-58.

Tudorie, C. (2008). Qualifying objects in classical relational database querying. In: Galindo J. (Ed.) Handbook of Research on Fuzzy Information Processing in Databases (pp. 218-245). IGI Global, London.

Tudorie, C. (2009). Intelligent interfaces for database fuzzy querying. The annals of “Dunarea de Jos” University of Galati, Fascicle III, 32(2).

Yager, R., and Kacprzyk, J. (Eds.). (1997). The Ordered Weighted Averaging Operators: Theory and Applications. Kluwer, Boston.

Yager, R. (1982). A new approach to the summarization of data. *Information Sciences*, 28: 69–86.

Yager, R. (1988). On ordered weighted avaraging operators in multicriteria decision making. *IEEE Transactions on Systems,Man and Cybernetics*, SMC-18: 183 - 190.

Zadeh, L.A. (1983). A computational approach to fuzzy quantifiers in natural languages. *Computers and Mathematics with Applications*, 9: 149 – 184.

Zadrożny, S., de Tré, G., de Caluwe, R. and Kacprzyk, J. (2008). An overview of fuzzy approaches to flexible database querying. In: J. Galindo (Ed.), Handbook of Research on Fuzzy Information Processing in Databases (pp 34-54). IGI Global, London.

Zadrożny, S. and Kacprzyk, J. (2009) Issues in the practical use of the OWA operators in fuzzy querying. *Journal of Intelligent Information Systems*, 33: 307-325.