

Linguistic data summarization: a high scalability through the use of natural language?

Janusz Kacprzyk and Sławomir Zadrozny

Systems Research Institute

Polish Academy of Sciences

ul. Newelska 6

01-447 Warsaw, Poland

Email: {kacprzyk,zadrozny}_at_ibspan_dot_waw_dot_pl

ABSTRACT

We discuss aspects related to the scalability of data mining tools meant in a different way than whether a data mining tool retains its intended functionality as the problem size increases. We introduce a new concept of a cognitive (perceptual) scalability meant as whether as the problem size increases the method remains fully functional in the sense of being able to provide intuitively appealing and comprehensible results to the human user. We argue that the use of natural language in the linguistic data summaries provides a high cognitive (perceptual) scalability because natural language is the only fully natural means of human communication and provides a common language for individuals and groups of different backgrounds, skills, knowledge. We show that the use of Zadeh's protoform as general representations of linguistic data summaries, proposed by Kacprzyk and Zadrozny (2002; 2005a; 2005b), amplify this advantage leading to an ultimate cognitive (perceptual) scalability.

Keywords: Data Mining, Scalability, Fuzzy Logic, Natural Language Generation, Protoform, Flexible Query

INTRODUCTION

The purpose of this paper is to present a novel, different argument for the usefulness and power of linguistic data(base) summarization the essence of which was proposed by Yager (1982), and an extended, implementable version was shown by Kacprzyk & Yager (2001) and Kacprzyk, Yager & Zadrozny (2000).

We consider our further developments of the basic solutions presented in those papers which are relevant for our discussion, notably:

- a close relation between the linguistic data summarization and fuzzy database querying, to be more specific using fuzzy queries with linguistic quantifiers proposed by us (Kacprzyk & Ziółkowski, 1986) and in a much more extended form in (Kacprzyk, Zadrozny & Ziółkowski, 1989), and even more so in FQUERY for Access (Kacprzyk & Zadrozny, 2001b),

- our general approach to linguistic data summarization viewed as an interactive process in which fuzzy querying makes possible the articulation of the user's intentions, interests and information needs proposed by Kacprzyk & Zadrozny (1998; 2001a), and
- our formulation of linguistic data summarization in terms not only of the calculus of linguistically quantified proposition but in terms of Zadeh's protoforms (cf. (Kacprzyk & Zadrozny, 2002; 2005a; 2005b)) which can provide an extraordinary transparency, versatility and generality.

Our purpose in this paper will not be, however, a traditional exposition of the essence of those ideas which have been presented in our papers as referred to above, and which have proved to be very effective and efficient. We will discuss these tools and techniques from the perspective of this volume, that is, from the perspective of *scalability* of data mining (knowledge discovery) tools and techniques. In the case of linguistic data(base) summarization this will have a couple of aspects exemplified by both more technical computation time and memory related aspects of the scalability of databases and querying, and more conceptual aspects of what might be called a *cognitive* or *perceptual scalability* of tools from the point of view of human facilities and capabilities. Ultimately, we will argue that linguistic data summarization may be viewed from some points of view, notably with respect to the cognitive and perceptual scalability, as an ultimately scalable (in the cognitive or perceptual sense) tool for data mining and knowledge discovery.

BACKGROUND

The first question we should ask is: What is actually meant by *scalability*, in particular in the context of broadly perceived information technology? Usually, *scalability* is meant in two basic ways. First, it is understood as the ability of a computer application or system (i.e. hardware and/or software) to continue to function when the size of the problem in question (e.g. the size of a computer network, number of clients, size of data sets, etc.) changes, usually grows up. In our context of a broadly perceived data analysis, in this paper the scalability will be meant in the upward sense. Second, in a modern view, scalability is meant as the ability of a computer application and/or system not only to function as the size of the problem and/or context increases (or decreases but this case will not be considered) but to even take advantage of that increase in size and volume, for instance to provide more adequate results because of a larger basic data set, or an ability to more adequately grasp the very essence of a larger data set. Needless to say that scalability is a desirable property of any application or system, and virtually all nontrivial applications and systems are designed and implemented with scalability in mind.

As one can expect, though scalability is easily intuitively comprehensible, it is difficult to define, and may mean different things to different people, in particular when they come from different areas. What is relevant to us, a scalable online transaction processing system or database management system is the one that can be upgraded to process more transactions by adding new processors, devices and storage, and which can be upgraded easily and transparently. This is one of the reasons that we concern the scalability in the sense of what happens when the size and volume of data increase.

Scalability is a multidimensional concept. For instance, people often confuse *performance* and *scalability*. As pointed out by Haines (2006): "The terms "performance" and "scalability" are commonly used interchangeably but the two are distinct: performance measures the speed with which a single request can be executed, while scalability measures the ability of a request to maintain its performance under increasing size and volume. For example, the performance of a request may be said to generate a valid response within three seconds, but the scalability of the request concerns the ability to maintain that three-second response time as the user load increases" (p. 224). This distinction has a great impact for our discussion, and will be dealt with later.

Viewed simplistically, scalability is about “doing more of something” like responding to more user requests, executing more work or handling more data. Traditionally, this is done by either increasing the sheer computing power and/or data handling power exemplified by using parallel computation, grid computing, etc.

In this context a popular belief is that databases do not scale up well, i.e. that it is difficult to keep growing the size of a database, or too hard to handle the load of an increasing number of concurrent users. In other words, it is often believed that systems that are database centric are fundamentally incapable of efficiently coping with the (growing) demands of high performance distributed computing. This may be true to some extent even in view of a growing storage capacity at a diminishing cost, parallelization of processing, new software developments, etc. One can easily reach limits of the same inherent nature as those characteristic for even the best, most advances and densely packed traditional silicon integrated circuits: sooner or later, a new type of processors (biological?) will be needed.

This example of an unavoidable necessity of a technological change in processors can be rephrased in the context of the scaling up of database centric systems and applications which is what our work is concerned with.

Now, let us present the basic context we will be operating in, and issues related to scalability. We are concerned with data summarization which is one of basic capabilities of any "intelligent" system, and since for the human being the only fully natural means of communication is natural language, a linguistic summarization would be very desirable, exemplified by, for a data set on employees, a statement (linguistic summary) “almost all young and well qualified employees are well paid”.

Unfortunately, data summarization is still in general unsolved a problem. Very many techniques are available but they are not “intelligent enough”, and not human-consistent, partly due to a limited use of natural language (cf. Lesh & Mitzenmacher, 2004).

We deal with a conceptually simple approach to the linguistic database summaries introduced by Yager (1982; 1991; 1996), and then considerably advanced by Kacprzyk (2000), Kacprzyk & Yager (2001), and Kacprzyk, Yager & Zadrożny (2000; 2001), Zadrożny and Kacprzyk (1999), and implemented in Kacprzyk and Zadrożny (2000a-d; 2001a-e; 2002; 2003; 2005b). In this approach linguistic data summaries are derived as linguistically quantified propositions as, e.g., “most of the employees are young and well paid”, with a degree of truth (validity), possibly extended with other measures.

For an effective and efficient derivation of linguistic summaries, we employ Kacprzyk and Zadrożny's (1998; 2000a-d; 2001a) interactive approach to linguistic summaries in which the determination of a class of summaries of interest is done via Kacprzyk and Zadrożny's (1994; 1995a-b; 2001b) FQUERY for Access, a fuzzy querying add-in to Microsoft Access, extended to the querying over the Internet in Kacprzyk and Zadrożny (2000b). Since a fully automatic generation of linguistic summaries is not feasible at present, mainly because it is difficult if not impossible at all to automatically reveal the user's real intentions, interests and information needs, an interaction with the user is assumed for the determination of a class of summaries of interest, and this is done via the above fuzzy querying add-in.

Extending Kacprzyk & Zadrożny (2002; 2005a; 2005b), we show that by relating various types of linguistic summaries to fuzzy queries, with various known and sought elements, we can arrive at a hierarchy of prototypical forms, or – in Zadeh's (2002) terminology – protoforms, of linguistic data summaries. This seems to be a very powerful conceptual idea because it provides a simple structural expression, with a comprehensible semantics, of even the most complicated linguistic summaries.

Notice that, first, through the use of natural language to present (verbalize) the very essence and contents of data with respect to an aspect in question we certainly attain a high, maybe even the best

scalability. First, natural language can express that information in a fully comprehensible way no matter how large the data set is. Second, such simple linguistically quantified propositions with which data summaries are equated may semantically be adequate as representations of data sets of any size as they represent some highly abstracted linguistic statements, of a simple syntax and of what might be described as a “commonsense based” semantics. Third, protoforms of linguistic summaries provide a uniform, easily comprehensible form of linguistic summaries for any size of data sets, and virtually all intentions and information needs of the user. Finally, natural language summaries are comprehensible to individuals, small and larger groups, people from different backgrounds, people coming from various geographic locations, sexes, age groups, etc. Clearly, an obvious condition of an agreed upon semantics of language used should be assumed but this is a prerequisite of any human communication, and any implementation of a computer system to be employed by various human users.

A natural question is: what is the relation of the approach and view presented in this paper to the problem of natural language generation (NLG), and in particular to the scalability of natural language generation. We will not deal in more detail with these important issues. For an analysis of relations between the linguistic data summaries used in this paper, and in all our previous works, and some extension of the template based approach to natural language generation we refer the reader to Kacprzyk & Zadrozny (2009). Moreover, for very interesting remarks and their justification that natural language generation itself can be viewed as a very effective and efficient, yet conceptually simple and natural, and extremely human consistent way to improve the scalability of a dialog system, we refer the reader to Reiter (1995).

For more detail on the issue of scalable natural language generation we refer the reader to, for instance, Klarner (2004). Basically, in those works scalability of the natural language generation is also considered in the context of dialog systems, i.e. slightly more general than in our context of just the linguistic summarization of numerical sets of data, but concerns many aspects that are relevant for us too. Basically, scalability for (spoken) dialog systems is meant as the ability to:

- enlarge the domain content by modifying and extending its thematic orientation,
- refine the domain language to extend the linguistic coverage and expressibility of the domain,
- change the application domain which usually concerns the two above ones and can lead to completely new requirements for a dialog system and its parts,
- change the discourse domain which may alter the discourse type within the same domain.

As it can clearly be seen there are strong, intrinsic relations between our concept of a linguistic data summary, and its protoform based representation, and various concepts of scalability both in a general context of systems and applications in information technology, database related technology, and – finally – natural language generation (NLG).

It should be noted that our approach to scalability is different than that of most researchers who practically equate the property of scalability with whether, and how well, a given approach, tool, technique, ... can retain its functionality, effectiveness and efficiency when the size of the problem is growing, i.e. in our case the size of a data set is growing. This is upward scalability. Sometimes very relevant is downward scalability when the size of the problem is diminishing. A trivial example is that (many if not all) statistical methods are not downward scalable in this sense because they do not work properly for small problems (samples). The downward scalability is in general difficult to deal with.

Most works on the (upward) scalability concern the efficiency of search for a solution, here for a best linguistic summary, which may be called a *technical scalability*. In this work we are basically concerned with a much more general and foundational type of scalability, which might be called a *conceptual* or *perceptual scalability* which has to do with a fundamental question: will our tools remain conceptually or perceptually appropriate (human consistent) when our problem will greatly increase? We will advocate that due to the use of natural language we obtain an ultimate conceptual or perceptual scalability because a natural language statement will always be comprehensible to the

human being(s) no matter what size of the data set it is meant to represent. We will also give some remarks on technical scalability by, first, reviewing some approaches that make possible the generation of linguistic summaries for large data sets. We will not, however, mention our approach based on a relation between the generation of linguistic data summaries and association rules which was originally proposed by Kacprzyk & Zadrozny (2001d; 2003). This approach shows a different perspective and its role in the context of scalability, both technical and cognitive (perceptual), of linguistic data summaries needs a different exposition which will be presented in a next paper.

We will present now in more detail an implementation of our interactive approach to the derivation of linguistic summaries, and while discussing particular elements we will indicate relations to those scalability issues and aspects mentioned above. We hope that this will provide another justification to the power of both linguistic data summaries in the simple sense assumed here, and the power of Zadeh's protoforms, and maybe even – more generally – the power of Zadeh's computing with words and perceptions paradigm (cf. Zadeh & Kacprzyk, 1999). All this will be presented in a novel, not yet explored perspective of a conceptual (perceptual) scalability.

LINGUISTIC DATA(BASE) SUMMARIES

Data summarization is one of basic capabilities now needed by any "intelligent" system that is meant to operate in real life situations. Basically, due to the availability of relatively cheap and efficient hardware and software tools, we usually face an abundance of data that is beyond human cognitive, perceptual and comprehension skills.

Since for the human being the only fully natural means of communication is natural language, a linguistic (say, by a sentence or a small number of sentences in a natural language) summarization of a set of data would be very desirable and human consistent. For instance, having a data set on employees, a statement (linguistic summary) like “almost all younger and well qualified employees are well paid” would be useful and human consistent in many cases.

Unfortunately, data summarization is still in general unsolved a problem in spite of vast research efforts. Very many techniques are available but they are not “intelligent enough”, and not human consistent, partly due to a little use of natural language. This concerns, e.g., summarizing statistics, exemplified by the average, median, minimum, maximum, α -percentile, etc. which – in spite of recent efforts to soften them – are still far from being able to reflect a real human perception of their essence.

Linguistic Data Summarization – The Basic Case

In this paper we will use a simple yet effective and efficient approach to the linguistic summarization of data sets (databases) proposed by Yager (1982), and then presented in a more advanced, and implementable form by Kacprzyk & Yager (2001), and Kacprzyk, Yager & Zadrozny (2000). This will provide a point of departure for our further analysis of more complicated and realistic summaries.

In Yager's (1982) approach, we have (we use here the author's terminology):

- V is a quality (attribute) of interest, e.g. salary in a database of workers,
- $Y = \{y_1, \dots, y_n\}$ is a set of objects (records) that manifest quality V , e.g. the set of workers; hence $V(y_i)$ are values of quality V for object $y_i \in Y$;
- $D = \{V(y_1), \dots, V(y_n)\}$ is a set of data (the “database” on question)

A *linguistic summary* of a data set consists of:

- a summarizer S (e.g. young),
- a quantity in agreement Q (e.g. most),

- truth T - e.g. 0.7,

as, e.g., " $T(\text{most of employees are young})=0.7$ ". The truth T may be meant in a more general sense, e.g. as validity or, even more generally, as some quality or goodness of a linguistic summary.

Basically, given a set of data D , we can hypothesize any appropriate summarizer S and any quantity in agreement Q , and the assumed measure of truth will indicate the truth of the statement that Q data items satisfy the statement (summarizer) S .

Notice that we consider here some specific, basic form of a linguistic summary. We do not consider other forms of summaries exemplified by "over 70% of employees are under 35 years of age" that may be viewed to provide similar information as "most of employees are young" because the latter are clearly outside of the class of linguistic summaries considered. Notice also that we discuss here the linguistic summarization of sets of numeric values only. One can clearly imagine the linguistic summarization of symbolic attributes but this relevant problem is outside of the scope of this paper. We do not consider here the linguistic summarization of textual information.

We should also note that we do not consider in this paper some other approaches to the linguistic summarization of databases (data sets) that are based on a different philosophy, exemplified by works by Bosc et al. (2002), Dubois & Prade (1992), Raschia & Mouaddib (2002) or Rasmussen & Yager (1996; 1997a; 1997b; 1999). Basically, one can very briefly summarize the approaches employed as follows. First, Bosc et al. (1992) use a gradual rule view of linguistic summaries, which has been proposed by Dubois & Prade (1992) and use linguistic quantifiers as tools for the aggregation. Rasmussen & Yager (1999) consider both the traditional Yager summaries and a type of Dubois & Prade's gradual rules showing that they can be obtained (or, more precisely, verified) via some extension of SQL. Raschia & Mouaddib (2002) propose, and develop in a series of papers, a different approach based on hierarchical summaries, their tree representations, and relations to OLAP based techniques. Summaries are here meant as aggregated ("generalized") tuples which cover parts of the database at different levels of abstraction.

We will not consider some other related techniques exemplified by the mining of fuzzy association rules (cf. (Chen, Liu & Li, 2001; Chen & Wei, 2002; Chen, Wei & Kerre; 2000; Hu, Chen & Tzeng, 2002; Lee & Lee-Kwang, 1997)), even in the context of linguistic summaries (cf. (Kacprzyk and Zadrozny, 2001d; 2003)). These approaches reflect a different perspective and, as already mentioned, will be a subject of a next paper which will consider scalability of linguistic data summaries in a comprehensive way, as a confluence of the technical and conceptual (perceptual) scalability.

First, we should consider the forms of the particular elements of a linguistic summary in our sense. Since we use natural language throughout our analysis, as it is the only fully natural and human consistent means of communication for the humans, we assume the summarizer S to be a linguistic expression semantically represented by a fuzzy set like, for instance "young" would be represented as a fuzzy set in the universe of discourse as, e.g., $\{1, 2, \dots, 90\}$, i.e. containing possible values of the human age, and "young" could be given as, e.g., a fuzzy set with a non-increasing membership function in that universe such that, in a simple case of a piecewise linear membership function, the age up to 35 years is for sure "young", i.e. the grade of membership is equal to 1, the age over 50 years is for sure "not young", i.e. the grade of membership is equal to 0, and for the ages between 35 and 50 years the grades of membership are between 1 and 0, the higher the age the lower its corresponding grade of membership. A simple one-attribute-related summarizer exemplified by "young" can clearly be extended to some confluence of attribute values as, e.g., "*young and well paid*".

Clearly, in the context of linguistic summarization of data, the most interesting are more sophisticated, *human-consistent* summarizers (concepts) as, e.g.:

- productive workers,
- stimulating work environment,

- difficult orders, etc.

whose definition involves complicated *combinations of attributes*, e.g.: a hierarchy (not all attributes are of the same importance), the attribute values are ANDed and/or ORed, k out of n , *most*, etc. of them should be accounted for, etc. The definition, processing and generation of such non-trivial summarizers needs some specific tools and techniques to be discussed later.

The quantity in agreement, Q , is an indication of the extent to which the data satisfy the summary. Once again, a precise indication is not human consistent, and a linguistic term represented by a fuzzy set is employed. Basically, two types of such a linguistic quantity in agreement can be used:

- absolute as, e.g., "about 5", "more or less 100", "several", and
- relative as, e.g., "a few", "more or less a half", "most", "almost all" etc.

Notice that the above linguistic expressions are the so-called fuzzy linguistic quantifiers (cf. Zadeh, 1983) that can be handled by fuzzy logic.

Similarly as for the fuzzy summarizer, the form (basically, the definition of a fuzzy linguistic quantifier) of a fuzzy quantity in agreement is also subjective, and can be either predefined or elicited from the user.

The calculation of the truth (or, more generally, validity) of the linguistic summary considered above is equivalent to the calculation of the truth value (from the unit interval) of a linguistically quantified statement (e.g., "*most* of the employees are *young*"). This can be calculated by using two most relevant techniques: Zadeh's (1983) calculus of linguistically quantified statements (cf. (Zadeh & Kacprzyk, 1999) or Yager's (1988) OWA operators (cf. (Yager & Kacprzyk, 1997)). Since these calculi are well known and are widely used in many works involving linguistic quantifier based aggregation of partial scores, we will discuss them only briefly in what follows and will refer the reader to, for instance, Zadeh's (1983; 1985) or Yager's (1988) source papers for more details.

A linguistically quantified proposition, exemplified by "most experts are convinced", is written as " Qy 's are F " where Q is a linguistic quantifier (e.g., *most*), $Y = \{y\}$ is a set of objects (e.g., experts), and F is a property (e.g., convinced). Importance B may be added yielding " QBy 's are F ", e.g., "most (Q) of the important (B) experts (y 's) are convinced (F)". The problem is to find $\text{truth}(Qy's \text{ are } F)$ or $\text{truth}(QBy's \text{ are } F)$, respectively, knowing $\text{truth}(y \text{ is } F), \forall y \in Y$ which is done here using Zadeh's (1983; 1985) fuzzy logic based calculus of linguistically quantified propositions.

Property F and importance B are fuzzy sets in Y , and a (proportional, nondecreasing) linguistic quantifier Q is assumed to be a fuzzy set in $[0,1]$ as, e.g.

$$\mu_Q(x) = \begin{cases} 1 & \text{for } x \geq 0.8 \\ 2x - 0.6 & \text{for } 0.3 < x < 0.8 \\ 0 & \text{for } x \leq 0.3 \end{cases} \quad (1)$$

Then, due to Zadeh (1983)

$$\text{truth}(Qy's \text{ are } F) = \mu_Q\left[\frac{1}{n} \sum_{i=1}^n \mu_F(y_i)\right] \quad (2)$$

$$\text{truth}(QBy's \text{ are } F) = \mu_Q\left[\frac{\sum_{i=1}^n (\mu_B(y_i) \wedge \mu_F(y_i))}{\sum_{i=1}^n \mu_B(y_i)}\right] \quad (3)$$

These formulas are clearly based on the non-fuzzy cardinalities of the respective fuzzy sets, the so-called Σ -Counts (cf. Zadeh, 1983).

An OWA operator (Yager, 1988; Yager & Kacprzyk, 1997) of dimension p is a mapping $F: [0,1]^p \rightarrow [0,1]$ if associated with F is a weighting vector $W = [w_1, \dots, w_p]^T$, $w_i \in [0,1]$, $w_1 + \dots + w_p = 1$, and

$$F(x_1, \dots, x_p) = w_1 b_1 + \dots + w_p b_p = W^T B \quad (4)$$

where b_i is the i -th largest element among x_1, \dots, x_p , $B = [b_1, \dots, b_p]$.

The OWA weights may be found from the membership function of Q due to (cf. Yager, 1988):

$$w_i = \begin{cases} \mu_Q(i) - \mu_Q(i-1) & \text{for } i = 1, \dots, p \\ \mu_Q(0) & \text{for } i = 0 \end{cases} \quad (5)$$

The OWA operators can model a wide array of aggregation operators (including linguistic quantifiers), from $w_1 = \dots = w_{p-1} = 0$ and $w_p = 1$ which corresponds to "all", to $w_1 = 1$ and $w_2 = \dots = w_p = 0$ which corresponds to "at least one", through all intermediate situations, and that is why they are widely employed.

An important case is when with the OWA operator importance qualification of the particular pieces of data is associated. Suppose that with the data $A = [a_1, \dots, a_p]$, a vector of importances $V = [v_1, \dots, v_p]$, such that $v_i \in [0,1]$ is the importance of a_i , $i = 1, \dots, p$, $v_1 + \dots + v_p = 1$, is associated. Then, for an *ordered weighted averaging operator with importance qualification* based on a linguistic quantifier Q , denoted OWA_Q , Yager (1988) proposed that, first, some redefinition of the OWA's weights w_i 's into \bar{w}_i 's is performed, and (4) becomes

$$w_1 = \dots = w_{p-1} = 0 \\ F_I(x_1, \dots, x_p) = \bar{w}_1 b_1 + \dots + \bar{w}_p b_p = \bar{W}^T B \quad (6)$$

where

$$\bar{w}_j = \mu_Q \left(\frac{\sum_{k=1}^j u_k}{\sum_{k=1}^p u_k} \right) - \mu_Q \left(\frac{\sum_{k=1}^{j-1} u_k}{\sum_{k=1}^p u_k} \right) \quad (7)$$

where u_k is the importance of b_k , i.e. the k -largest element of A .

Some Other Validity Measures of Linguistic Summaries

The basic validity criterion, i.e. the truth of a linguistically quantified statement given by (2) and (3), is certainly the most natural and important but it does not grasp all aspects of a linguistic summary. We will present here some other quality (validity) criteria, notably those proposed by Kacprzyk & Yager (2001), and Kacprzyk, Yager & Zadrożny (2000).

First, Yager (1982) proposed a measure of informativeness whose essence is: suppose that we have a data set, whose elements are from a space X . One can view the data set itself as its own most informative description, and any other summary implies a loss of information, and therefore informativeness comes into play

The degree of truth is unfortunately not a good measure of informativeness (cf. Yager, 1982; 1991). Let the summary be characterized by the triple (S, Q, T) , and let a related summary be

characterized by the triple (S^c, Q^c, T) such that S^c is the negation of S , i.e. $\mu_{S^c}(\cdot) = 1 - \mu_S(\cdot)$, and similarly $\mu_{Q^c}(\cdot) = 1 - \mu_Q(\cdot)$. Then, Yager (1982; 1991) proposed the following measure of informativeness of a summary

$$I = [T \cdot SP(Q) \cdot SP(S)] \vee [(1 - T) \cdot Sp(Q^c) Sp(S^c)] \quad (8)$$

where $SP(Q)$ is the specificity of Q given as

$$SP(Q) = \int_0^1 \frac{1}{\text{card}Q_\alpha} d\alpha \quad (9)$$

where Q_α is the α -cut of Q and $\text{card}(\cdot)$ is the ‘‘cardinality’’ (in fact, the area) of the respective set; and similarly for Q^c, S, S^c . Notice that in (8) we also have the specificity of $S/S^c, SP(S/S^c)$, which is meant similarly.

The rationale behind this measure of informativeness differs from that of, e.g., Chen, Liu & Li (2001). Unfortunately, this measure of informativeness is by no means a definite solution. First, let us briefly mention George and Srikanth’s (1996a; 1996b) proposal. Suppose that a linguistic summary of interest involves more than 1 attribute (e.g., ‘‘age’’, ‘‘salary’’ and ‘‘seniority’’ in the case of employees). Basically, for the same set of data, two summaries are generated:

- a constraint descriptor which is the most specific description (summary) that fits the largest number of tuples in the relation (database) involving the attributes in question,
- a constituent descriptor which is the description (summary) that fits the largest subset of tuples with the condition that each tuple attribute value takes on at least a threshold value of membership.

George and Srikanth (1996a; 1996b) use these two summaries to derive a fitness function (goodness of a summary) that is later used for deriving a solution (a best summary) via a genetic algorithm they employ. This fitness function represents a compromise between the most specific summary (corresponding to the constraint descriptor) and the most general summary (corresponding to the constituent descriptor).

Then, some additional measures have been developed by Kacprzyk & Yager (2001) and Kacprzyk, Yager & Zadrozny (2000). Let us briefly repeat some basic notation. We have a data set (database) D that concerns some objects (e.g. employees) $Y = \{y_1, \dots, y_n\}$ described by some attribute V (e.g. age) taking on values in a set $X = \{x_1, x_2, \dots\}$ exemplified by $\{20, 21, \dots, 100\}$ or even $\{\text{very young, young, } \dots, \text{old, very old}\}$ though this case will not be considered here. Let $d_i = V(y_i)$ denote the value of attribute V for object y_i . Therefore, the data set to be summarized is given as a table

$$D = [d_1, \dots, d_n] = [V(y_1), V(y_2), \dots, V(y_n)] \quad (10)$$

In a more realistic case the data set is described by more than one attribute. Let $V = \{V_1, V_2, \dots, V_m\}$ be a set of such attributes taking values in $X_i, i = 1, \dots, m$; $V_j(y_i)$ denotes the value of attribute V_j for object y_i , and attribute V_j takes on its values from a set X_j .

The data set to be summarized is therefore:

$$D = \{[V_1(y_1), V_2(y_1), \dots, V_m(y_1)], [V_1(y_2), V_2(y_2), \dots, V_m(y_2)], \dots, [V_1(y_n), V_2(y_n), \dots, V_m(y_n)]\} \quad (11)$$

In case of multiple (m) attributes the description (summarizer) S is assumed as a family of fuzzy sets $S = \{S_1, S_2, \dots, S_m\}$ where $S_i \in S$ is a fuzzy set in $X_i, i = 1, \dots, m$. Then, $\mu_S(y_i), i = 1, 2, \dots, n$, may be defined as:

$$\mu_S(y_i) = \min_{j \in \{1, 2, \dots, m\}} [\mu_{S_j}(V_j(y_i))] \quad (12)$$

and

$$r = \frac{\sum_{i=1}^n \mu_S(y_i)}{n} \quad (13)$$

and $T = \mu_Q(r)$.

So, having S , we can calculate the truth value T of a summary for any quantity in agreement. To find a best (optimal) summary, we should calculate T for each possible summarizer, and for each record in the database in question which may be computationally prohibitive for virtually all non-trivial databases and number of attributes. Therefore, from the point of view of scalability, this suggests that the process of finding an optimal linguistic summary is not *technically scalable*.

A natural line of reasoning would be to either limit the number of attributes of interest or to limit the class of possible summaries by setting a more specific description (e.g. very young, young and well paid, etc. employees). This will limit the search space, and may help attain an acceptable technical scalability.

We will deal now with the second option. The user can limit the scope of a linguistic summary to, for instance, those for which the "age" takes on the value "young" only, i.e. to fix the summarizer related to that attribute. This would correspond to the searching of the database using the query w_g equated with the fuzzy set in X_g corresponding to "young" related to attribute V_g (i.e. age), i.e. characterized by $\mu_{w_g}(\cdot)$. In such a case, $\mu_S(y_i)$ given by (12) becomes

$$\mu_S(y_i) = \min_{j \in \{1, 2, \dots, m\}} [\mu_{S_j}(V_j(y_i)) \wedge \mu_{w_g}(V_g(y_i))], \quad i=1, \dots, n \quad (14)$$

where " \wedge " is the minimum (or, more generally, a t -norm), and then

$$r = \frac{\sum_{i=1}^n \mu_S(y_i)}{\sum_{i=1}^n \mu_{w_g}(V_g(y_i))} \quad (15)$$

and $T = \mu_Q(r)$. This is clearly related to how Zadeh's calculus of linguistically quantified propositions works.

Now, we will briefly mention the 5 quality measures of linguistic database summaries, in particular four additional ones as introduced in Kacprzyk & Yager (2001), and Kacprzyk, Yager & Zadrozny (2000):

- a truth value [which basically corresponds to the degree of truth of a linguistically quantified proposition representing the summary given by, say, (2) or (3)],
- a degree of imprecision,
- a degree of covering,
- a degree of appropriateness,
- a length of a summary.

For notational simplicity later on, let us rewrite (12) and (1) as:

$$\mu_S(d_i) = \min_{j \in \{1, 2, \dots, m\}} [\mu_{S_j}(V_j(y_i))], \quad i=1, \dots, n \quad (16)$$

and

$$r = \frac{\sum_{i=1}^n [\mu_S(V_g(y_i)) \wedge \mu_{w_g}(V_g(y_i))]}{\sum_{i=1}^n \mu_{w_g}(V_g(y_i))} \quad (17)$$

where, clearly, (16) and (17) are equivalent to (12) and (15) though rewritten in the form more suitable for our present discussion.

The **degree of truth**, T_1 , is the basic validity criterion introduced in the source Yager's (1982) work and commonly employed. It is clearly equal to

$$T_1 = \mu_Q(r) \quad (18)$$

which results directly from Zadeh's (1983; 1985) calculus of linguistically quantified propositions.

The **degree of imprecision** is an obvious and important validity criterion. Basically, a very imprecise linguistic summary (e.g. on almost all winter days the temperature is rather cold) has a very high degree of truth yet it is not useful.

Suppose that description (summarizer) S is given as a family of fuzzy sets $S = \{S_1, S_2, \dots, S_m\}$. For a fuzzy set $S_j, j=1, \dots, m$, we can define its degree of fuzziness as, e.g.:

$$\text{in}(S_j) = \frac{\text{card} \{x \in X_j : \mu_{S_j}(x) > 0\}}{\text{card } X_j} \quad (19)$$

where card denotes the cardinality of the corresponding (nonfuzzy) set and the domains X_j are all assumed to be finite (what is reasonable from the practical point of view). That is, the "flatter" the fuzzy set S_j the higher the value of $\text{in}(S_j)$.

The degree of imprecision, T_2 , of the summary – or, in fact, of S – is then defined as:

$$T_2 = 1 - \sqrt[m]{\prod_{j=1, \dots, m} \text{in}(S_j)} \quad (20)$$

Notice that the degree of imprecision T_2 depends on the form of the summary only and not on the database, that is its calculation does not require the searching of the database (all its records) which is very important.

The **degree of covering**, T_3 , is defined as

$$T_3 = \frac{\sum_{i=1}^n t_i}{\sum_{i=1}^n h_i} \quad (21)$$

where:

$$t_i = \begin{cases} 1 & \text{if } \mu_S(y_i) > 0 \quad \text{and} \quad \mu_{w_g}(V_g(y_i)) > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$h_i = \begin{cases} 1 & \text{if } \mu_{w_g}(V_g(y_i)) > 0 \\ 0 & \text{otherwise} \end{cases}$$

and the denominator of (21) is assumed to be different from 0 - otherwise T_3 is defined to be equal 0.

The degree of covering says how many objects in the data set corresponding to the query w_g are "covered" by the particular summary. Its interpretation is simple as, e.g., if it is equal to 0.15, then this means that 15% of the objects are consistent with the summary in question. The value of this degree depends clearly on the contents of the database.

The **degree of appropriateness** is probably the most relevant measure. Suppose that the summary containing the description (fuzzy sets) $S = (S_1, S_2, \dots, S_m)$ is partitioned into m partial summaries each of which encompasses the particular attributes V_1, V_2, \dots, V_m , such that each partial summary corresponds to one fuzzy set only, then if we denote:

$$S_j(y_i) = \mu_{S_j}(V_j(y_i)) \quad (23)$$

then

$$r_j = \frac{\sum_{i=1}^n h_i}{n}, \quad j = 1, \dots, m \quad (23)$$

where, $h_i = \begin{cases} 1 & \text{if } S_j(y_i) > 0 \\ 0 & \text{otherwise} \end{cases}$, and the degree of appropriateness, T_4 , is defined as:

$$T_4 = \text{abs} \left(\prod_{j=1, \dots, m} r_j - T_3 \right) \quad (24)$$

The degree of appropriateness means that, for a database concerning the employees, if – for instance – 50% of them are less than 25 years old and 50% are highly qualified, then we may expect that 25% of the employees would be less than 25 years old and highly qualified; this would correspond to a typical, fully expected situation. However, if the degree of appropriateness is, e.g., 0.39 (i.e. 39% are less than 25 years old and highly qualified), then the summary found reflects an interesting, not fully expected relation in our data. This degree describes therefore how characteristic for the particular database the summary found is. T_4 is very important because a trivial summary like, for instance, "100 % of employees is of some age" has truth equal 1 but its degree of appropriateness is clearly equal 0.

The **length** of a summary is relevant because a long summary is not easily comprehensible by the human user. This length, T_5 , may be defined in various ways, and the below form has proven to be useful:

$$T_5 = 2 (0.5^{\text{card } S}) \quad (25)$$

Now, the (total) degree of validity, T , of a particular linguistic summary is defined as the weighted average of the above 5 degrees of validity, i.e.:

$$T = T(T_1, T_2, T_3, T_4, T_5; w_1, w_2, w_3, w_4, w_5) = \sum_{i=1,2, \dots, 5} w_i T_i \quad (26)$$

and the problem is to find an optimal summary, $S^* \in \{S\}$, such that

$$S^* = \arg \max_S \sum_{i=1,2, \dots, 5} w_i T_i \quad (27)$$

where: w_1, \dots, w_5 are weights assigned to the particular degrees of validity, with values from the unit interval, the higher, the more important such that $\sum_{i=1,2, \dots, 5} w_i = 1$.

The definition of weights, w_1, \dots, w_5 , is a problem in itself, and will not be dealt with in more detail. The weights can be predefined or elicited from the user.

As we have already mentioned, the linguistic summarization meant in terms of (27) is clearly not technically scalable, even if some more sophisticated search techniques are used which limit the size of the problem as exemplified by George & Srikanth's (1996a; 1996b) use of a genetic algorithm. However, let us notice that the situation is completely different when cognitive (perceptual)

scalability is accounted for. It is clear that the very concept of linguistic data summary as presented above is what might be said totally cognitively (perceptionally) scalable because it is comprehensible to a human being, either an individual or a group of individuals, no matter what size of the data set is, and also to a large extent independently of the background, sex, age, etc. of the individuals. This is a direct result of, on the one hand, the use of natural language, which is the only fully natural means of articulation and communication of a human being, and – on the other hand – of a simple and intuitively appealing form of a linguistic summary which basically says what most of the data exhibit, i.e. what *usually happens* (holds). This is in fact what is looked for and found by all data analysis tools and techniques.

PRACTICAL DETERMINATION OF LINGUISTIC DATA SUMMARIES

One can clearly notice that a fully automatic determination of a best linguistic summary, i.e. the solution of (26) may be infeasible in practice due to a high number of possible summaries. In (Kacprzyk & Zadrozny, 1998; 2001a) an *interactive approach* was proposed with a *user assistance* in the definition of summarizers, by the indication of attributes and their combinations of interest. This proceeds via a user interface of a fuzzy querying add-on. Basically, the queries (referring to summarizers) allowed are:

- *simple* as, e.g., "salary is *high*"
- *compound* as, e.g., "salary is *low* AND age is *old*"
- *compound (with quantifier)*, as, e.g., "*most* of {salary is *high*, age is *young*, ..., training is *well above average*}."

We will also use "natural" linguistic terms, i.e. (7±2!) exemplified by: *very low*, *low*, *medium*, *high*, *very high*, and also "comprehensible" fuzzy linguistic quantifiers as: *most*, *almost all*, ..., etc.

In (Kacprzyk & Zadrozny, 1994; 1995a; 1995b; 2001b), a conventional DBMS is used, and a fuzzy querying tool is developed to allow for queries with fuzzy (linguistic) elements of the "simple", "compound" and "compound with quantifier" types. This fuzzy querying system (add in) has been developed for Microsoft Access® but its concept is clearly applicable to any DBMS. The main problems to be solved are here: (1) how to extend the syntax and semantics of the query, and (2) how to provide an easy way of eliciting and manipulating those terms by the user.

We will now briefly describe the very essence of FQUERY for Access, emphasizing those aspects which are relevant for the purposes of this paper. One should notice that we will use here terms, exemplified by "attributes", "fields", etc. as used in our source papers on FQUERY for Access, which should help the interested readers follow more specialized discussions concerning FQUERY for Access given in these papers. These insignificant terminological differences should not lead to any confusion or misunderstanding. It should be noted that a slightly different approach to the use of linguistic quantifiers in fuzzy queries has been proposed – cf. Bosc, Lietard & Pivert (1995) – but it will not be used here.

FQUERY for Access is embedded in the native Access's environment as an add in. All the code and data is put into a database file, a *library*, installed by the user. Definitions of attributes, fuzzy values etc. are maintained in a dictionary (a set of regular tables), and a mechanism for putting them into the Query-By-Example (QBE) sheet (grid) is provided. Linguistic terms are represented within a query as parameters, and a query transformation is performed to provide for their proper interpretation during the query execution.

FQUERY for Access is an add-in that makes it possible to use fuzzy terms in queries. Briefly speaking, the following types of fuzzy terms are available:

- fuzzy values, exemplified by *low* in "profitability is *low*",
- fuzzy relations, exemplified by *much greater than* in "income is *much greater than* spending", and
- linguistic quantifiers, exemplified by *most* in "*most* conditions have to be met".

The elements of the first two types are elementary building blocks of fuzzy queries in FQUERY for Access. They are meaningful in the context of numerical fields only. There are also other fuzzy constructs allowed which may be used with scalar fields.

If a field is to be used in a query in connection with a fuzzy value, it has to be defined as an *attribute*. The definition of an attribute consists of two numbers: the attribute's values lower (LL) and upper (UL) limit. They set the interval which the field's values are assumed to belong to, according to the user. This interval depends on the meaning of the given field. For example, for *age* (of a person), the reasonable interval would be, e.g., [18,65], in a particular context, i.e. for a specific group. Such a concept of an attribute makes it possible to universally define fuzzy values.

Fuzzy values are defined as fuzzy sets on [-10, +10]. Then, *the matching degree* $md(\cdot, \cdot)$ of a simple condition referring to attribute AT and fuzzy value FV against a record R is calculated by:

$$md(AT = FV, R) = \mu_{FV}(\tau(R(AT)))$$

where: $R(AT)$ is the value of attribute AT in record R, μ_{FV} is the membership function of fuzzy value FV, $\tau: [LL_{AT}, UL_{AT}] \rightarrow [-10, 10]$ is the mapping from the interval defining AT onto [-10,10] so that we may use the same fuzzy values for different fields. A meaningful interpretation is secured by τ which makes it possible to treat all fields domains as ranging over the unified interval [-10,10]. For simplicity, it is assumed that the membership functions of fuzzy values are trapezoidal.

Linguistic quantifiers provide for a flexible aggregation of simple conditions. In FQUERY for Access the fuzzy linguistic quantifiers are defined in Zadeh's (1983; 1985) sense, as fuzzy set on [0, 10] interval instead of the original [0, 1] – cf. *most* given as (1). They may be interpreted either using original Zadeh's (1983) approach or via the OWA operators (cf. (Yager, 1988) or (Yager & Kacprzyk, 1997)); Zadeh's interpretation will be considered in what follows. The membership functions of fuzzy linguistic quantifiers are assumed piece-wise linear, hence two numbers from [0,10] are needed. Again, a mapping from [0, N], where N is the number of conditions aggregated, to [0,10] is employed to calculate the matching degree of a query. More precisely, the matching degree, $md(\cdot, \cdot)$, for the *query* "*Q* of *N* conditions are satisfied" for record R is equal to

$$md(Q, condition_i, R) = \mu_Q[\tau(\sum_i md(condition_i, R))]$$

and we can also assign different importance degrees for particular conditions. Then, the aggregation formula is equivalent to (3). The importance is identified with a fuzzy set on [0,1], and then treated as property *B* in (3).

Before a fuzzy term may be used in a query, it has to be defined using the toolbar provided by FQUERY for Access and stored internally. This feature, i.e. maintenance of dictionaries of fuzzy terms defined by users, strongly supports our approach to data summarization discussed in this paper. In fact, the package comes with a set of predefined fuzzy terms but the user may enrich the dictionary too.

When the user initiates the execution of a query it is automatically transformed by appropriate FQUERY for Access's routines and then run as a native query of Access. The transformation consists primarily in the replacement of parameters referring to fuzzy terms by calls to functions implemented by the package which secure a proper interpretation of these fuzzy terms. Then, the query is run by Access as usually. Details can be found in Kacprzyk & Zadrozny (1994 – 1995b).

It is obvious that fuzzy queries directly correspond to summarizers in linguistic summaries. Thus, the derivation of a linguistic summary may proceed in an interactive (user assisted) way as follows:

- the user formulates a set of linguistic summaries of interest (relevance) using the fuzzy querying add in,
- the system retrieves records from the database and calculates the validity of each summary adopted, and
- a best (most appropriate) linguistic summary is chosen.

The use of fuzzy querying is very relevant because we can restate the summarization in the fuzzy querying context. First, (2) may be interpreted as:

"Most records match query S " (28)

where S replaces F in (2) since we refer here directly to the concept of a summarizer (of course, S is in fact the whole condition, e.g., price = *high*, while F is just the fuzzy value, i.e. *high* in this condition; this should not lead to confusion).

Similarly, (3) may be interpreted as:

"Most records meeting conditions B match query S " (29)

Thus, (29) says something only about a subset of records specified by (28). In database terminology, B corresponds to a *filter* and (29) claims that *most* records passing through B match query S . Moreover, since the filter may be fuzzy, a record may pass through it to a degree from $[0,1]$.

And, again, one can argue for a very high conceptual (perceptual) scalability of linguistic data summaries because their determination boils down to a well known process of database querying which virtually all users of computer systems, even novice users, are accustomed to.

Looking at the form of (28) and (29), which specify the user's interest and intent as to linguistic data summaries put in the context of database querying, it was proposed by Kacprzyk & Zadrozny (2002; 2005b) that the concept of a *protoform* in the sense of Zadeh (2002; 2006) is highly relevant. A protoform is defined as an abstract prototype, that is, in our context, for the query (summary) given by (28) and (29) as follows, respectively:

"*Most R's are S*" (30)

and

"*Most BR's are S*" (31)

where R means "records", B is a filter, and S is a query.

Since protoforms can obviously form a hierarchy, we can define higher level (more abstract) protoforms, for instance replacing *most* by a generic linguistic quantifier Q , we obtain, respectively:

$$"QR's \text{ are } S" \quad (32)$$

and

$$"QBR's \text{ are } S" \quad (33)$$

Obviously, the more abstract protoforms correspond to cases in which we assume less about summaries sought. There are two limit cases, where we: (1) assume totally abstract protoform or (2) assume all elements of a protoform are given as specific linguistic terms. In case 1 data summarization will be extremely time consuming, as the search space may be enormous, but may produce interesting, unexpected views on data. In case 2 the user has to guess a good candidate formula for summarization but the evaluation is fairly simple, just equivalent to the answering of a (fuzzy) query. Thus, the second case refers to the summarization known as *ad hoc queries*.

Then, going further along this line, we can show in Table 1 a classification of linguistic summaries into 5 basic types corresponding to protoforms of a more and more abstracted form.

Table 1: Classification of linguistic summaries

Type	Given	Sought	Remarks
1	S	Q	Simple summaries through ad-hoc queries
2	SB	Q	Conditional summaries through ad-hoc queries
3	$QS^{structure}$	S^{value}	Simple value oriented summaries
4	$QS^{structure}B$	S^{value}	Conditional value oriented summaries
5	Nothing	SBQ	General fuzzy rules

where $S^{structure}$ denotes that attributes and their connection in a summary are known, while S^{value} denotes a summarizer sought.

Type 1 may be easily produced by a simple extension of fuzzy querying as in Kacprzyk & Zadrozny's (2001b) FQUERY for Access. Basically, the user has to construct a query – a candidate summary, and it has to be determined what is the fraction of rows matching this query and what linguistic quantifier best denotes this fraction. A Type 2 summary is a straightforward extension of Type 1 by adding a fuzzy filter. Type 3 summaries require much more effort. Their primary goal is to determine typical (exceptional) values of an attribute. So, query S consists of only one simple condition built of the attribute whose typical (exceptional) value is sought, the "=" relational operator and a placeholder for the value sought. For example, using the following summary in the context of personal data: $Q = "most"$ and $S = "age=?"$ (here "?" denotes a placeholder mentioned above) we look for a typical value of age. A Type 4 summary may produce typical (exceptional) values for some, possibly fuzzy, subset of rows. From the computational point of view Type 5 summaries represent the most general form considered here: fuzzy rules describing dependencies between specific values of particular attributes. Here the use of B is essential, while previously it was optional. The summaries of Type 1 and 3 have been implemented as an extension to Kacprzyk & Zadrozny's (1994; 1995a-b; 2001b) FQUERY for Access. Two approaches to Type 5 summaries have been proposed. Firstly, a subset of such summaries may be produced by exploiting similarities with the *association rules* concept (Agrawal & Srikant, 1994) and employing their efficient algorithms. Second, genetic algorithm may be employed to search the summaries' space as initiated by George & Srikanth (1996a; 1996b). We will not consider these issues because they refer more to technical scalability and are dealt with in a different perspective than the one assumed in this paper.

Clearly, the protoforms are a powerful conceptual tool because we can formulate many different types of linguistic summaries in a uniform way, and devise a uniform and universal way to handle

different linguistic summaries. Therefore, Kacprzyk & Zadrozny (2002; 2005) have certainly confirmed frequent claims by Zadeh and other researchers that protoforms are powerful indeed.

Notice, that all our previous statements about a very high conceptual (perceptual) scalability of linguistic data summaries in the form considered here are valid to an even higher extent when protoforms are involved. Namely, the simplicity and intuitive appeal of the protoforms used in the context of linguistic data summaries make them applicable to data sets of any size. Even if the size of a data set increases, the very essence of a particular protoform just catches the contents of the data set in a user comprehensible form. And, by imposing a general template on the form of a summary, a protoform would presumably make the transition to the analysis of data sets of a larger size much smoother because no new general pattern of expected results would be necessary. That is why one can argue that our approach of using linguistic data summaries for data mining (knowledge discovery) can be viewed as a significant step towards the ultimate scalability of data mining (knowledge discovery) tools and techniques in all cases when the human user plays a significant role.

SOME FUTURE RESEARCH DIRECTIONS

Among many possible future works related to the concept of a cognitive (perceptual) scalability of data mining tools and techniques via linguistic data summaries, the following ones seem important and viable. First, the issue of a “comprehensive” scalability of linguistic data summaries can be considered in the sense that both the traditionally meant scalability (i.e. the retaining of functionality as the problem size, for instance the size of a database, increases) and the cognitive (perceptual) scalability proposed are combined. This has to do with many aspects including the development of more effective and efficient fuzzy querying tools, and of generation methods of linguistic data summaries, for instance using some more advanced evolutionary tools than in George & Srikanth (1996b).

An interesting future research direction would be to extend the arguments of this paper to cover another relevant approach to linguistic data summaries, namely through the use of gradual rules introduced by Dubois & Prade (1992). Similarly, an interesting issue would be to analyze yet another, different approach to linguistic summarization by Raschia & Mouaddib (2002), maybe even more so by considering their later papers in which a relation to OLAP has been indicated. The use of another approach to the introduction of quantified statements into fuzzy queries due to Bosc, Lietard and Pivert (1995) and their later works can be interesting.

Finally, one can also consider in the perspective of cognitive (perceptual) scalability the use of various protoforms extending our works Kacprzyk & Zadrozny (2005a; 2005b), in which an approach has also been proposed relating the generation of linguistic data summaries to some ways of generating some fuzzy association rules so that quite effective and efficient (though maybe not fully scalable) algorithms for association rule mining can be employed.

CONCLUDING REMARKS

We have discussed some aspects related to a crucial issue of scalability of data mining (knowledge discovery) tools and techniques by considering some special modern approach in that area, the so called linguistic data summaries.

We have argued first that the scalability should be meant in a more sophisticated way than just in terms of whether a particular tool and/or technique can retain its intended functionality, effectiveness and efficiency as the size of the problem (here the size and volume of data) increases.

We have introduced a new concept of a cognitive (perceptual) scalability whose essence is whether as the size of the problem increases a particular method will be fully functional, effective and efficient, but in the sense of being able to provide intuitively appealing and comprehensible results.

We have argued that the use of natural language in the linguistic summaries provides a high cognitive (perceptual) scalability because natural language is the only fully natural means of articulation and communication of a human being, and also the use of natural language provides a common language for both the individuals and groups of different background, technical skills, knowledge, etc. No other communication means, as numbers or graphics, exhibit this property to the same extent.

Then, going even further in this direction, we have shown that Zadeh's protoform as general representations of linguistic data summaries, as proposed by Kacprzyk and Zadrozny (2002; 2005a; 2005b) amplify even more this advantage leading to what might be called an ultimate cognitive (perceptual) scalability.

REFERENCES

- Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, & C. Zaniolo (Eds.), *20th International Conference on Very Large Databases, Santiago de Chile, Chile* (pp. 487-499). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
- Bosc, P., Dubois, D., Pivert, O., Prade, H., & de Calmes, M. (2002). Fuzzy summarization of data using fuzzy cardinalities. In *9th International Conference Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2002)* (pp. 1553-1559) Annecy, France.
- Bosc, P., Lietard, L., & Pivert, O. (1995). Quantified statements and database fuzzy querying. In: P. Bosc, & J. Kacprzyk (Eds.), *Fuzziness in Database Management Systems* (pp. 275-308). Heidelberg: Physica-Verlag.
- Chen, G., Liu, D., & Li, J. (2001). Influence and conditional influence – new interestingness measures in association rule mining. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'2001)* (pp. 1440-1443). Vancouver, Canada.
- Chen, G., & Wei, Q. (2002). Fuzzy association rules and the extended mining algorithm. *Information Sciences*, 147, 201–228.
- Chen, G., Wei, Q., & Kerre E.E. (2000). Fuzzy data mining: discovery of fuzzy generalized association rules. In G. Bordogna, & G. Pasi (Eds.), *Recent Research Issues on Fuzzy Databases* (pp. 45-66). Heidelberg and New York: Springer-Verlag.
- Dubois, D., & Prade, H. (1992). Gradual rules in approximate reasoning. *Information Sciences*, 61, 103-122.
- George, R., & Srikanth, R. (1996a). A soft computing approach to intensional answering in databases. *Information Sciences*, 92, 313-328.
- George, R., & Srikanth, R. (1996b). Data summarization using genetic algorithms and fuzzy logic. In F. Herrera, & J.L. Verdegay (Eds.), *Genetic Algorithms and Soft Computing* (pp. 599-611). Heidelberg: Physica-Verlag.
- Haines, S. (2006). *Pro Java EE 5 Performance Management and Optimization*. Berkeley, CA: Apress.
- Hu, Y.-Ch., Chen, R.-Sh., & Tzeng G.-H. (2002). Mining fuzzy association rules for classification problems. *Computers and Industrial Engineering*, 43, 735-750.
- Kacprzyk, J. (2000). Intelligent data analysis via linguistic data summaries: a fuzzy logic approach. In R. Decker & W. Gaul (Eds.), *Classification and Information Processing at the Turn of the Millennium* (pp. 153-161). Berlin, Heidelberg and New York: Springer-Verlag.

- Kacprzyk, J., & Yager, R.R. (2001). Linguistic summaries of data using fuzzy logic. *International Journal of General Systems*, 30, 133-154.
- Kacprzyk, J., Yager, R.R., & Zadrożny, S. (2000). A fuzzy logic based approach to linguistic summaries of databases. *International Journal of Applied Mathematics and Computer Science*, 10, 813-834.
- Kacprzyk, J., Yager, R.R., & Zadrożny, S. (2001). Fuzzy linguistic summaries of databases for an efficient business data analysis and decision support. In W. Abramowicz, & J. Żurada (Eds.), *Knowledge Discovery for Business Information Systems* (pp. 129-152). Boston: Kluwer.
- Kacprzyk, J., & Zadrożny S. (1994). Fuzzy querying for Microsoft Access. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'94) vol. 1* (pp. 167-171). Orlando, USA.
- Kacprzyk, J., & Zadrożny, S. (1995a). Fuzzy queries in Microsoft Access v. 2. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'95), Workshop on Fuzzy Database Systems and Information Retrieval* (pp. 61-66). Yokohama, Japan.
- Kacprzyk, J., & Zadrożny, S. (1995b). FQUERY for Access: fuzzy querying for a Windows-based DBMS. In P. Bosc, & J. Kacprzyk (Eds.), *Fuzziness in Database Management Systems* (pp. 415-433). Heidelberg: Physica-Verlag.
- Kacprzyk, J., & Zadrożny, S. (1998). Data mining via linguistic summaries of data: an interactive approach. In T. Yamakawa, & G. Matsumoto (Eds.), *Methodologies for the Conception, Design and Application of Soft Computing - Proceedings of IIZUKA'98* (pp. 668-671). Iizuka, Japan.
- Kacprzyk, J., & Zadrożny, S. (2000a). On combining intelligent querying and data mining using fuzzy logic concepts. In G. Bordogna, & G. Pasi (Eds.), *Recent Research Issues on the Management of Fuzziness in Databases* (pp. 67 - 81). Heidelberg and New York: Physica-Verlag.
- Kacprzyk, J., & Zadrożny, S. (2000b). Data mining via fuzzy querying over the Internet. In O. Pons, M.A. Vila, & J. Kacprzyk (Eds.), *Knowledge Management in Fuzzy Databases* (pp. 211-233). Heidelberg and New York: Physica-Verlag.
- Kacprzyk, J., & Zadrożny, S. (2000c). On a fuzzy querying and data mining interface. *Kybernetika*, 36, 657-670.
- Kacprzyk, J., & Zadrożny, S. (2000d). Computing with words: towards a new generation of linguistic querying and summarization of databases. In P. Sinčák, & J. Vaščák (Eds.), *Quo Vadis Computational Intelligence?* (pp. 144–175). Physica-Verlag (Springer-Verlag): Heidelberg and New York.
- Kacprzyk, J., & Zadrożny, S. (2001a). Data mining via linguistic summaries of databases: an interactive approach. In L. Ding (Ed.), *A New Paradigm of Knowledge Engineering by Soft Computing* (pp. 325-345). Singapore: World Scientific.
- Kacprzyk, J., & Zadrożny, S. (2001b). Computing with words in intelligent database querying: standalone and Internet-based applications. *Information Sciences*, 34, 71–109.
- Kacprzyk, J., & Zadrożny, S. (2001c). On linguistic approaches in flexible querying and mining of association rules. In H.L. Larsen, J. Kacprzyk, S. Zadrożny, T. Andreassen, & H. Christiansen (Eds.), *Flexible Query Answering Systems. Recent Advances* (pp. 475 – 484), Heidelberg and New York: Springer-Verlag.
- Kacprzyk, J., & Zadrożny, S. (2001d). Fuzzy linguistic summaries via association rules. In A. Kandel, M. Last, & H. Bunke (Eds.), *Data Mining and Computational Intelligence* (pp. 115-139). Heidelberg and New York: Physica-Verlag (Springer-Verlag).

- Kacprzyk, J., & Zadrozny, S. (2001e). Using fuzzy querying over the Internet to browse through information resources. In B. Reusch, & K.-H. Temme (Eds.), *Computational Intelligence in Theory and Practice* (pp. 235-262). Heidelberg and New York: Physica-Verlag (Springer-Verlag).
- Kacprzyk, J., & Zadrozny, S. (2002). Protoforms of linguistic data summaries: towards more general natural-language-based data mining tools. In A. Abraham, J. Ruiz del Solar, & M. Koeppen (Eds.), *Soft Computing Systems* (pp. 417—425). Amsterdam: IOS Press.
- Kacprzyk, J., & Zadrozny, S. (2003). Linguistic summarization of data sets using association rules. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'03)* (pp. 702-707). St. Louis, USA.
- Kacprzyk, J., & Zadrozny, S. (2005a). Protoforms of linguistic database summaries as a tool for human-consistent data mining. In *14th Annual IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2005)* (pp. 591-596). Reno, NV, USA: IEEE.
- Kacprzyk, J., & Zadrozny, S. (2005b). Linguistic database summaries and their protoforms: towards natural language based knowledge discovery tools. *Information Sciences*, 173, 281-304.
- Kacprzyk, J., & Zadrozny, S. (2009). Protoforms of Linguistic Database Summaries as a Human Consistent Tool for Using Natural Language in Data Mining. *International Journal of Software Science and Computational Intelligence*, 1(1).
- Kacprzyk, J., Zadrozny, S., & Ziolkowski, A. (1989). FQUERY III+: a 'human-consistent' database querying system based on fuzzy logic with linguistic quantifiers. *Information Systems*, 14, 443-453.
- Kacprzyk, J., & Ziolkowski, A. (1986). Database queries with fuzzy linguistic quantifiers. *IEEE Transactions on Systems, Man and Cybernetics SMC*, 16, 474-479.
- Klarner, M. (2004). Hyperbug - A Scalable Natural Language Generation Approach. In R. Portzel (Ed.), *2nd International Workshop on Scalable Natural Language Understanding (ScaNaLu-2004)* (pp. 65-71). Boston, MA, USA: Association for Computational Linguistics.
- Lee, J.-H., & Lee-Kwang, H. (1997). An extension of association rules using fuzzy sets. In *7th IFSA World Congress* (pp. 399-402). Prague, Czech Republic.
- Lesh, N., & Mitzenmacher, M. (2004). Interactive data summarization: an example application. In *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI '04), Gallipoli, Italy* (pp.183-187). New York, NY: ACM.
- Raschia, G., & Mouaddib, N. (2002). SAINTETIQ: a fuzzy set-based approach to database summarization. *Fuzzy Sets and Systems*, 129, 137-162.
- Rasmussen, D. & Yager, R.R (1999) Finding fuzzy and gradual functional dependencies with summarySQL. *Fuzzy Sets and Systems*, 106, 131-142.
- Reiter, E. (1995). Building Natural Language Generation Systems. In A. Cawsey (Ed.), *AI and Patient Education Workshop*, Glasgow, UK: University of Glasgow.
- Yager, R.R. (1982). A new approach to the summarization of data. *Information Sciences*, 28, 69-86.
- Yager, R.R. (1988). On ordered weighted averaging operators in multicriteria decision making. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-18, 183-190.
- Yager, R.R. (1991). On linguistic summaries of data. In G. Piatetsky-Shapiro, & W.J. Frawley (Eds.), *Knowledge Discovery in Databases* (pp. 347-363). Menlo Park: AAAI Press/The MIT Press.
- Yager, R.R. (1996). Database discovery using fuzzy sets. *International Journal of Intelligent Systems*, 11, 691-712.

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

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

Zadeh, L.A. (1985). Syllogistic reasoning in fuzzy logic and its application to usuality and reasoning with dispositions. *IEEE Transaction on Systems, Man and Cybernetics*, SMC-15, 754-763.

Zadeh, L.A. (2002). A prototype-centered approach to adding deduction capabilities to search engines – the concept of a protoform. In *BISC Seminar, 2002*. Berkeley: University of California.

Zadeh, L.A. (2006). From search engines to question answering systems - the problems of world knowledge relevance deduction and precisiation. In E. Sanchez (Ed.), *Fuzzy Logic and the Semantic Web* (pp. 163-210). Amsterdam: Elsevier.

Zadeh, L., & Kacprzyk, J. (Eds.) (1999). *Computing with Words in Information/Intelligent Systems, 1. Foundations, 2. Applications*. Heidelberg and New York: Physica-Verlag.

Zadrożny, S., & Kacprzyk, J. (1999). On database summarization using a fuzzy querying interface. In *IFSA '99 World Congress* (pp. 39-43). Taipei, Taiwan R.O.C.

ADDITIONAL READING

Anwar, T.M., Beck, H.W., & Navathe, S.B. (1992). Knowledge mining by imprecise querying: A classification based system. In *International Conference on Data Engineering* (pp. 622-630), Tampa, USA.

Bosc, P., & Kacprzyk, J. (Eds.) (1995). *Fuzziness in Database Management Systems*. Heidelberg: Physica-Verlag.

Bosc, P., & Pivert, O. (1992). Fuzzy querying in conventional databases. In L.A. Zadeh, & J. Kacprzyk (Eds.), *Fuzzy Logic for the Management of Uncertainty* (pp. 645-671). New York: Wiley.

Kacprzyk, J., Pasi, G., Vojtaš, P., & Zadrożny, S. (2000), Fuzzy querying: issues and perspective. *Kybernetika*, 36, 605-616.

Kacprzyk, J., & Zadrożny, S. (1999). The paradigm of computing with words in intelligent database querying. In L.A. Zadeh & J. Kacprzyk (Eds.), *Computing with Words in Information/Intelligent Systems. Part 2. Foundations* (pp. 382-398). Heidelberg and New York: Springer-Verlag.

Petry, F.E. (1996). *Fuzzy Databases: Principles and Applications*. Boston: Kluwer.

Rasmussen, D., & Yager, R.R. (1996). Using SummarySQL as a tool for finding fuzzy and gradual functional dependencies. In *6th International Conference Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'96)* (pp. 275 - 280). Granada, Spain.

Rasmussen, D., & Yager, R.R. (1997a). Fuzzy query language for hypothesis evaluation. In T. Andreassen, H. Christiansen, & H. L. Larsen (Eds.), *Flexible Query Answering Systems* (pp. 23-43). Boston: Kluwer.

Rasmussen, D., & Yager, R.R. (1997b). A fuzzy SQL summary language for data discovery. In D. Dubois, H. Prade, & R.R. Yager (Eds.), *Fuzzy Information Engineering: A Guided Tour of Applications* (pp. 253-264). New York, NY: Wiley.

Rasmussen, D., & Yager, R.R. (1999). Finding fuzzy and gradual functional dependencies with SummarySQL. *Fuzzy Sets and Systems*, 106, 131-142.

Yager, R.R., & Kacprzyk, J. (1999). Linguistic Data Summaries: A Perspective. In *IFSA '99 Congress* (pp. 44-48). Taipei, Taiwan R.O.C.

Zadeh, L.A., & Kacprzyk, J. (Eds.) (1992). *Fuzzy Logic for the Management of Uncertainty*. New York, NY: Wiley.

Zadrozny, S., Kacprzyk, J., & Gola, M. (2005). Towards Human Friendly Data Mining: Linguistic Data Summaries and Their Protoforms. *Lecture Notes in Computer Science*, 3697, 697-702.