Int. J. Human-Computer Studies (2001) **55,** 293–309
doi:10.1006/ijhc.2001.0468
Available online at http://www.idealibrary.com on IDEAL®

Relational attribute systems

IVO DÜNTSCH

School of Information and Software Engineering, University of Ulster, Newtownabbey, BT37 0QB, N. Ireland. email: i.duentsch@ulst.ac.uk

GÜNTHER GEDIGA

Institut für Evaluation und Marktanalysen, Brinkstr. 19, 49143 Jeggen, Germany. email: gediga@eval-institut.de

EWA ORLOWSKA†

Institute of Telecommunications, Szachowa 1, 04–894, Warszawa, Poland. email: orlowska@itl.waw.pl

(Received 5 December 2000, and accepted in revised form 27 March 2001)

We introduce a relational operationalization of data which generalizes, among others, the deterministic information systems of Pawlak (1982), the indeterministic systems of Lipski (1976) and Orłowska and Pawlak (1987), and the context relations of Wille (1982); it can also be used for fuzzy data modelling. Using an example from the area of psychometrics, we show how our operationalization can lead to an improved understanding of agreements and disagreements by experts in classification tasks.

"Das Merkwürdigste an einem Loch ist der Rand".† (Tucholsky, 1975)

1. Introduction

In this paper we are concerned with developing a formal mechanism for describing the state of a researcher's knowledge about objects in a chosen domain, which extends the widely used data table operationalization in such a way that one or more boundary values are incorporated between "strictly yes" and "strictly no". It turns out that relations between objects and features are a suitable tool to achieve our aim. Using set theoretical properties and common relational operators, we are able to express not only the classical cases, but also semantical constraints such as single-valued, multiple-valued, deterministic or indeterministic attributes. We can introduce different relations for different states of knowledge. For example, consider the following.

- xIv if and only if x certainly has property v.
- xBv if and only if x possibly has property v.

† "The most peculiar thing about a hole is its boundary".

1071-5819/01/090293 + 17 $35.00/0

Relations of this kind induce binary relations on the object set in various ways. While in the classical case we have either equality or diversity, we can consider more differentiated cases in our set-up. For example, the relation

$$xTy \Leftrightarrow (\forall v)[xIv \text{ implies } yIv \text{ or } yBv]$$

allows us to compare the certain features of x with the certain or possible features of y. In this spirit, we can also find (possible) compatibility between object descriptions.

The paper is organized as follows: In the first section we recall several modes of operationalization which have appeared in the literature and their model assumptions. Section 3 introduces our relational operationalization of data domains, and Section 4 explores the relations among objects induced by the object–attribute relations. Finally, we present an example for our approach, which shows how expert ratings can be better understood and how possible reconciliation strategies can be found.

2. Domain operationalization

According to Gigerenzer (1981), a data modelling process consists of six parts (Figure 1).

(1) A *researcher* who decides what is to be investigated and how this is to be done.
(2) A *domain of interest* which is to be investigated.
(3) When a domain of interest is investigated, one needs to introduce a system with a language which includes relation and/or operator symbols as necessary with which the properties of the domain can be described. This system is called an *empirical model*.
(4) A mapping which assigns the observed properties and relations on the domain to the empirical system. This mapping is called *operationalization* in the Social Sciences, and *knowledge representation* in Artificial Intelligence. The result of the mapping is often called *(raw) data*.
(5) A mapping called *scaling* or *representation mapping* which assigns the elements of the empirical system to some kind of structure.
(6) A *representation system* or *numerical system* which is the result of the scaling process. This can be a linear order of numbers, a graphical structure, or similar systems.

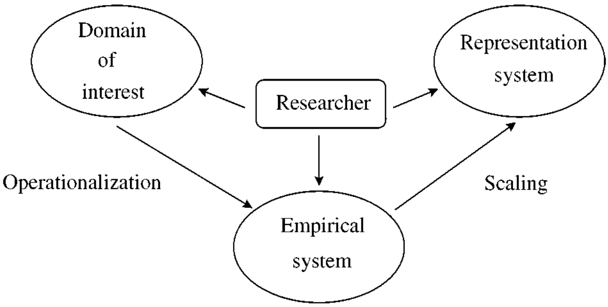

FIGURE 1. Data modelling.

As an example, let the domain of interest be the skills which a student has. An empirical system is a set of test questions, and an operationalization is a mapping which assigns to a set M of skills the set of those problems that can be solved with the skills in M (Düntsch & Gediga, 1995). A representation system can be a set of percentages, a set of grades or a knowledge structure in the sense of Doignon and Falmagne (1999), with the scaling being a mapping from the test results to the chosen scale or system.

We would like to emphasize that in this paper we shall be mainly concerned with the operationalization of a domain of interest and the corresponding empirical model. This step is (or should be) independent of further steps of analysis which construct numerical models, such as the currently in vogue probabilistic techniques. The further analysis, which is always possible with our operationalization, will be of less concern to us in the present context. This is not such a gap as it may seem at first glance. One can argue that the use of a numerical model with all its additional assumptions and their subsequent influences on the result is often neither necessary nor, indeed, desirable, and we invite the reader to consult Düntsch and Gediga (2000) for details.

One of the oldest operationalizations of data is the

$$\text{Object} \mapsto \text{Attributes} \tag{1}$$

assignment, i.e. in terms of *extension* ("Umfang") and *intension* ("Inhalt") of Leibniz and Kant: A researcher chooses a domain of interest, the attributes describing (parts of) the domain, and studies the objects in the domain which fall under the description. The data array as shown in Table 1 is an example of such an operationalization: The leftmost column denotes different specimen of Iris flowers, while each of the other columns describes one property (attribute) of each specimen.

A formal definition of this type of operationalization is the following: A *single-valued information system* is a structure

$$\mathscr{I} = \langle U, \Omega, \{V_a : a \in \Omega\}\rangle, \tag{2}$$

where

- U is a finite set of objects.
- Ω is a finite set of mappings $a : U \to V_a$; each $a \in \Omega$ is called an *attribute*.
- V_a is the set of *attribute values* of attribute a.

TABLE 1
Iris data (Fisher, 1936)

Specimen	Sepal length	Sepal width	Petal length	Petal width	Species
1	50	33	14	2	1
2	46	34	14	3	1
3	65	28	46	15	2
4	62	22	45	15	2
6	67	30	50	17	3
7	64	28	56	22	3
		⟨143 other values⟩			

TABLE 2
Planet system (Wille, 1982)

Planet	Size
Mercury	Small
Venus	Small
Earth	Small
Mars	Small
Jupiter	Large
Saturn	Large
Uranus	Medium
Neptune	Medium
Pluto	Small

TABLE 3
Decomposed planet system

Planet	Small	Medium	Large
Mercury	1	0	0
Venus	1	0	0
Earth	1	0	0
Mars	1	0	0
Jupiter	0	0	1
Saturn	0	0	1
Uranus	0	1	0
Neptune	0	1	0
Pluto	1	0	0

Any operationalization puts semantic constraints on the data set. A simple assumption which is used (but not always consciously) by practically all information systems is the "nominal scale restriction" which postulates that each object has exactly one value of each attribute at a given time, and that the observation of this value is without error. One of the aims of this paper is to relax these strict conditions.

Given an information system $\mathscr{I}$ as above, Iwinski (1988) calls an information system $\mathscr{I}' = \langle U', \Omega' \{V_v : v \in \Omega'\} \rangle$ a *decomposition of* $\mathscr{I}$, if the following hold.

(1) $U = U'$.
(2) $V_v = \{0, 1\}$ for all $v \in \Omega'$.
(3) For each $a \in \Omega$ there is some $\Omega_a \subseteq \Omega'$ such that (a) there is a bijection $f_a : \Omega_a \rightarrow \{a(x) : x \in U\}$ and (b) for all $v \in \Omega_a$, $x \in U$,

$$v(x) = 1 \Leftrightarrow a(x) = f(v). \tag{3}$$

Consider, for example, the information system of Table 2, which, for simplicity, has only one attribute "Size"; the decomposition of $\mathscr{I}$ is shown in Table 3.

Binary decomposition of attributes in this way faces the problem, that there are various forms of such attributes: Consider, for example, the attribute a "being alive" with the set of attribute values {yes, no}. If $a(x) = \text{no}$, then we can infer that x is dead. Thus, the absence of the property signals the presence of one other and vice versa. Binary attributes with this property are called *symmetric*. If, on the other hand, the attribute a is "colour", then being not red does not usually imply the presence of a particular colour. Such type of binary attribute is called *asymmetric* (see Jaccard, 1908). Clearly, the information value of symmetric and asymmetric binary attributes is different, and care should be taken not to confuse the two. It is straightforward to transform a symmetric binary attribute into two asymmetric ones.

Wille (1982) operationalizes a single-valued information system by taking its binarization, and then interpreting the occurrence of 1 in row x at (binary) attribute v as the presence of the pair $\langle x, v\rangle$ in a *context relation I*. In the planet example, the operationalization of the data is given by the context relation $I \subseteq U \times \Omega'$ containing the pairs

$$\langle \text{Mercury,small}\rangle, \langle \text{Venus,small}\rangle, \langle \text{Earth,small}\rangle, \langle \text{Mars,small}\rangle, \langle \text{Pluto,small}\rangle,$$

$$\langle \text{Uranus,medium}\rangle, \langle \text{Neptune,medium}\rangle,$$

$$\langle \text{Jupiter,large}\rangle, \langle \text{Saturn,large}\rangle.$$

A generalization of single-valued information systems which could indicate incompleteness was discussed by Lipski (1981).

> Information incompleteness means that instead of having a single value of an attribute, we have a subset of the attribute domain, which represents our knowledge that the actual value is one of the values in this subset, though we do not know which one.

A *multi-valued information system* is a structure

$$\mathscr{I} = \langle U, \Omega, \{V_a : a \in \Omega\}\rangle, \tag{4}$$

where the following hold.

- U is a finite set of objects.
- Ω is a finite set of mappings $a: U \to 2^{V_a}$; each $a \in \Omega$ is called an *attribute*.
- V_a is the set of *attribute values* of attribute a.

While Lipski indicates a semantic constraint, namely, that $a(x)$ is a set of possible values for $x \in U$, exactly one of which applies, the *indeterministic information systems* of Orłowska and Pawlak (1987), while formally the same as Lipski's system, do not put any semantic constraint on $a(x)$.

There are many other ways to give a semantic interpretation of a multi-valued information system; here are a few examples.

(1) $a(x)$ is interpreted conjunctively and exhaustively. For example, if a is the attribute "speaking a language", then, $a(x) = \{\text{German, Polish, French}\}$ can be interpreted as

$$x \text{ speaks German, Polish and French and no other languages.} \tag{5}$$

(2) $a(x)$ can also be interpreted conjunctively and non-exhaustively as in

$$a \text{ speaks German, Polish and French and possibly other languages.} \tag{6}$$

(3) $a(x)$ is interpreted disjunctively and exclusively. For example, a witness states that

$$\text{The car that went too fast was either a Mercedes or a Ford.} \tag{7}$$

Here, exactly one of the following statements is true, but it is not known which one.

- The car that went too fast was a Mercedes.
- The car that went too fast was a Ford.

(4) $a(x)$ is interpreted disjunctively and non-exclusively. For example, if x is "cooperates with", then

$$a(\text{Ivo}) = \{\text{Günther, Ewa}\} \tag{8}$$

means that Ivo cooperates with Günther or Ewa or both.

3. Relational attribute systems

In this section we shall unify the operationalizations described above and, in addition, make semantic constraints explicit.

We shall need some notation and definitions: Suppose that $R \subseteq A \times B$ is a binary relation. If $x \in A$, we let $R(x) = \{v : xRv\}$; furthermore, $R^{\smile}$ is the relation $\{\langle v, x\rangle : xRv\}$, called the *converse of* R. If $R \subseteq A \times B$ and $S \subseteq B \times C$ then the *composition of* R *and* S, written as $R \circ S$, is the relation

$$x(R \circ S)y \Leftrightarrow (\exists z \in B)[xRz \text{ and } zSy]. \tag{9}$$

Note that $R \circ S \subseteq A \times C$. The *identity relation on* A is $1'_A = \{\langle x, x\rangle : x \in A\}$, and the *universal relation* $A \times A$ on A is denoted by 1_A.

The attributes of a single-valued information system are functions $a : U \to V_a$, while the attributes of a multi-valued system assign to each $x \in U$ a set of (possible) values. Such a function $a : U \to 2^{V_a}$ corresponds to a relation $R_a \subseteq U \times V_a$ by setting

$$xR_a v \Leftrightarrow v \in a(x). \tag{10}$$

This generalizes the binary information systems and the context relations described above.

While operationalizations such as those of Pawlak (1982) or Orłowska and Pawlak (1987) are not (openly) concerned with semantic constraints as part of the design process of an information system and only consider the given data, we will take into account those constraints which occur among the attributes regardless of the extension given by a specific data set. This is a common procedure in the theory of relational data bases, in which constraints are specified *ab initio*. Thus, in order to be consistent, we need to specify these semantic constraints as part of the operationalization; for example, we have to state whether $a(x)$ is to be interpreted conjunctively or disjunctively.

We are now ready for our main definition: A *relational attribute system* (RAS) is a structure

$$\langle U, \Omega, \langle \mathscr{R}_a \rangle_{a \in \Omega}, \langle V_a \rangle_{a \in \Omega}, \Delta \rangle$$

such that the following statements hold.

(1) U is a finite set of objects.
(2) Ω is a finite set of attribute names.

(3) $R \subseteq U \times V_a$ for each $a \in \Omega$ and each $R \in \mathscr{R}_a$.
(4) Δ is a set of constraints.

The relations in $\mathscr{R}_a$ are the non-numerical counterparts of degrees of (un)certainty. A relation $R \in \mathscr{R}_a$ connects an object x with a property v from V_a, whenever v is a property of x to a degree represented by R. In this way, the relations in $\mathscr{R}_a$ express our knowledge about the certainty of possession of properties from V_a by object x. The constraints describe the type of operationalization, such as single-valued, multiple-valued, deterministic or indeterministic. We do not want to prescribe the (logical) form of the constraints. It will turn out, that for the simple (and most important) cases equations between relations are sufficient.

In what follows, we shall exhibit how the operationalizations from above can be found in our systems. Suppose that $\mathscr{I}$ is a single-valued information system. For each $a \in \Omega$ we let $I_a \subseteq U \times V_a$ be defined by

$$xI_a v \Leftrightarrow a(x) = v. \tag{11}$$

This is just the context definition from above. The constraint for this type of system translates into the fact that for each $x \in U$ there is exactly one $v \in \Omega_a$ such that $xI_a v$. It is not hard to see that this condition is equivalent to the relational equations

$$I_a \circ 1_{V_a} = U \times 1_{V_a} \quad I_a \text{ is total.} \tag{12}$$

$$I_a{}^{\smile} \circ I_a \subseteq 1'_{V_a} \quad I_a \text{ is functional.} \tag{13}$$

For the situation of equation (5), we let a be the attribute "language" and consider the property "speaking a language" described by the relation

$$xI_a v \Leftrightarrow x \text{ speaks language } v. \tag{14}$$

There are no constraints; however, if we want to prescribe that each person speaks at least one language, then we will have constraint (12). We notice how our relational notation allows us to generalize the one-valued deterministic information systems to many-valued deterministic systems.

To be able to express incomplete information we introduce another relation B_a, and we interpret $xB_a v$ as "x has possibly the a-property v". The constraints arising from Lipski's systems are

$$B_a{}^{\smile} \circ I_a = \emptyset \qquad \text{"Certainly" and "Possibly" are not compatible.} \tag{15}$$

$$I_a{}^{\smile} \circ I_a \subseteq 1'_{\Omega_a} \quad I_a \text{ is functional.} \tag{16}$$

This is reminiscent of fuzzy sets in that we do not necessarily have crisp attribute assignments, and also of rough sets (Pawlak, 1982), since we have only one relation per attribute for uncertainty. If we want to express a more finely grained "vagueness", we can express a fuzzy membership function in the obvious way. In this connection we should like to mention the possibilistic approach of Prade (1984) and Bosc and Prade (1994), in which possibility degrees are used to describe the strength of membership in a category. It may be worth pointing out that by modelling degrees of certainty by relations, we shift model assumptions—and thus, subjectivity—from the construction of a numerical model to the operationalization, which is a subjective decision by the researcher for every type of operationalization.

Note that condition (15) implies that $I_a \cap B_a = \emptyset$, and that xI_av_1 and xB_av_2 are together impossible. For equation (8) we note that there are no semantic constraints.

Even though we will concentrate in the sequel only on the relations I and B, these are by no means the only conceivable ones. Another frequently used relation is the one which signals the absence of a property such as "not red".

4. Relational properties

In this section we shall look at relations between objects, which are induced by the relations in $\mathscr{R}$; this generalizes the dependencies of rough set theory, and the information relations of Orłowska (1995). More concretely, we shall consider the case of the relations I_a and B_a as described in the previous section, i.e.

- xI_av means that x certainly has the a-property v.
- xB_av means that x possibly has the a-property v.

In the following considerations we will concentrate on the case of a single attribute a, and consequently drop the subscripts from I_a, B_a.

In order to picture the relations I and B, we agree on the following convention: U and Ω are finite, and we write the system as a data matrix with rows labelled by the elements of U, and columns labelled by the elements of Ω. If xIv, we place $\clubsuit$ into the cell $\langle x, v\rangle$, and for xBv we write $\diamond$. Using this notation, Lipski's conditions (15) and (16) can be stated equivalently as

$$\clubsuit \text{ and } \diamond \text{ cannot appear in the same row.} \tag{17}$$

$$\text{There is at most one } \clubsuit \text{ in every row.} \tag{18}$$

We also set $H = I \cup B$; then, $H(x)$ is the set of those attribute values which x certainly possesses and those which it possibly possesses. This is similar to the lower and upper approximation of rough set analysis (Pawlak, 1982), or to the egg-yolk model of Cohn and Gotts (1996), where

$$\underbrace{H(x)}_{\text{egg}} = \underbrace{I(x)}_{\text{yolk}} \cup \underbrace{B(x)}_{\text{white}}.$$

Our overall constraint is

$$I \cap B = \emptyset. \tag{19}$$

In rough set theory, two objects in a single-valued information system are called *indiscernible*, if they have the same feature vector. In a multi-valued system there are other possibilities which use set theoretic relations on the sets $a(x)$. This leads to the *information relations* first studied by Orłowska (1995). Our relational setting extends these relations in the following way: We will consider the relations

$$=,\ \subsetneq,\ \supsetneq,\ O,\ D, \tag{20}$$

where for a set M and subsets t, u of M,

$$tOu \Leftrightarrow t \cap u \neq \emptyset, \text{ and } t \text{ and } u \text{ are incomparable with respect to } \subseteq,$$

$$tDu \Leftrightarrow t \cap u = \emptyset.$$

Then, the relations of equation (20) partition $M \times M$. Such "intersection tables" have been considered in qualitative spatial reasoning for the interior I and boundary B of sets in a topological space (Egenhofer & Franzosa, 1991; Egenhofer, 1994). In Tucholsky's terms, the interior corresponds to the hole, and the boundary is the uncertainty, the investigation of which is often much more interesting than studying I.

Given x, y in U, there are nine ways of relating an element of $\{I(x), B(x), H(x)\}$ with an element of $\{I(y), B(y), H(y)\}$, and we denote these possibilities by row headings

$$II,\ IB,\ IH,\ BI,\ BB,\ BH,\ HI,\ HB,\ HH. \tag{21}$$

We can now construct a relational table by indicating below each heading which of the relations of equation (20) holds. Of course, not all configurations are possible, since we have to observe the conditions

$$H = I \cup B \quad \text{and} \quad I \cap B = \emptyset. \tag{22}$$

If one of the entries is $=$, then additional constraints occur which are listed in Table 4. There, for example, the entry D in the cell $\langle I(x) = I(y),\ BI \rangle$ means that $I(x) = I(y)$ implies $B(x) \cap I(y) = \emptyset$.

The 80 arrangements, which are possible when we disregard the columns which contain H, are shown in Table 5. The EY column gives the number(s) of the corresponding I, H ("egg-yolk") configuration(s) as listed in Cohn and Gotts (1996), Figure 4. Figure 2 shows that several egg-yolk configurations can belong to the same I, B configuration: Case 52 of the I, B Table 5 corresponds to the egg-yolk configurations 10 and 12 of Cohn and Gotts (1996). Furthermore, since not every I, B configuration is associated with an I, H configuration, we see that the expressive powers of I, B and I, H configurations are incomparable.

TABLE 4
Equality constraints

	II	IB	IH	BI	BB	BH	HI	HB	HH
$I(x) = I(y)$	$=$	D	$\subsetneq$	D			$\supsetneq$		
$I(x) = B(y)$	D	$=$	$\subsetneq$		D			$\supsetneq$	
$I(x) = H(y)$	$\supsetneq$	$\supsetneq$	$=$	D	D	D	$\supsetneq$	$\supsetneq$	$\supsetneq$
$B(x) = I(y)$	D			$=$	D	$\subsetneq$	$\supsetneq$		
$B(x) = B(y)$		D		D	$=$	$\subsetneq$		$\supsetneq$	
$B(x) = H(y)$	D	D	D	$\supsetneq$	$\supsetneq$	$=$	$\supsetneq$	$\supsetneq$	$\supsetneq$
$H(x) = I(y)$	$\subsetneq$	D	$\subsetneq$	$\subsetneq$	D	$\subsetneq$	$=$	D	$\subsetneq$
$H(x) = B(y)$	D	$\subsetneq$	$\subsetneq$	D	$\subsetneq$	$\subsetneq$	D	$=$	$\subsetneq$
$H(x) = H(y)$			$\subsetneq$			$\subsetneq$	$\supsetneq$	$\supsetneq$	$=$

Table 5
Set configurations without H

No.	*II*	*IB*	*BI*	*BB*	EY	No.	*II*	*IB*	*BI*	*BB*	EY	No.	*ii*	*ib*	*bi*	*bb*	EY
1.	$=$	D	D	$=$	46	2.	D	$=$	$=$	D		3.	$=$	D	D	D	
4.	$=$	D	D	$\subsetneq$	41	5.	$=$	D	D	$\supsetneq$	40	6.	$=$	D	D	O	39
7.	D	$=$	$\subsetneq$	D		8.	D	$=$	$\supsetneq$	D		9.	D	$=$	O	D	
10.	D	$=$	D	D		11.	D	$\subsetneq$	$=$	D		12.	D	$\supsetneq$	$=$	D	
13.	D	O	$=$	D		14.	D	D	$=$	D		15.	$\subsetneq$	D	D	$=$	
16.	$\supsetneq$	D	D	$=$		17.	O	D	D	$=$		18.	D	D	D	$=$	
19.	D	D	D	$\subsetneq$		20.	D	D	D	$\supsetneq$		21.	D	D	D	O	2
22.	D	D	D	D	1	23.	$\subsetneq$	D	D	D		24.	$\supsetneq$	D	D	D	
25.	O	D	D	D		26.	D	$\subsetneq$	D	D		27.	D	$\supsetneq$	D	D	
28.	D	O	D	D		29.	D	D	$\subsetneq$	D		30.	D	D	$\supsetneq$	D	
31.	D	D	O	D		32.	D	$\subsetneq$	$\subsetneq$	D		33.	D	$\subsetneq$	$\supsetneq$	D	
34.	D	$\subsetneq$	O	D		35.	D	$\supsetneq$	$\subsetneq$	D		36.	D	$\supsetneq$	$\supsetneq$	D	
37.	D	$\supsetneq$	O	D		38.	D	O	$\subsetneq$	D		39.	D	O	$\supsetneq$	D	
40.	D	O	O	D		41.	$\subsetneq$	D	D	$\subsetneq$		42.	$\subsetneq$	D	D	$\supsetneq$	
43.	$\subsetneq$	D	D	O		44.	$\supsetneq$	D	D	$\subsetneq$		45.	$\supsetneq$	D	D	$\supsetneq$	
46.	$\supsetneq$	D	D	O		47.	O	D	D	$\subsetneq$		48.	O	D	D	$\supsetneq$	
49.	O	D	D	O		50.	D	$\subsetneq$	$\supsetneq$	O	19, 28, 34, 42	51.	D	$\subsetneq$	O	O	11, 13
52.	D	O	$\supsetneq$	O	10, 12	53.	O	$\supsetneq$	$\subsetneq$	D		54.	O	O	$\subsetneq$	D	
55.	O	$\supsetneq$	O	D		56.	$\subsetneq$	D	O	$\supsetneq$	33, 45	57.	$\subsetneq$	D	O	O	18, 26, 32, 38
58.	O	D	O	$\supsetneq$		59.	$\supsetneq$	O	D	$\subsetneq$	36, 44	60.	O	O	D	$\subsetneq$	
61.	$\supsetneq$	O	D	O	17, 25, 27, 61	62.	O	O	O	O	14, 15, 16, 20, 21, 22, 35, 29, 43	63.	D	D	$\supsetneq$	$\supsetneq$	7
64.	D	D	$\supsetneq$	O	64	65.	D	D	O	$\supsetneq$		66.	D	D	O	O	3
67.	$\supsetneq$	$\supsetneq$	D	D	23, 30	68.	O	$\supsetneq$	D	D		69.	$\supsetneq$	O	D	D	
70.	O	O	D	D		71.	D	$\subsetneq$	D	$\subsetneq$	8	72.	D	$\subsetneq$	D	O	6
73.	D	O	D	$\subsetneq$		74.	D	O	D	O	4	75.	$\subsetneq$	D	$\subsetneq$	D	24, 37
76.	O	D	$\subsetneq$	D		77.	$\subsetneq$	D	O	D		78.	O	D	O	D	
79.	D	O	O	O	9	80.	O	O	O	D							

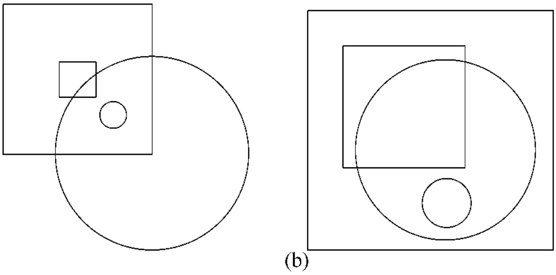

FIGURE 2. Two I, B equivalent egg-yolk pairs: (a) egg-yolk 10; (b) egg-yolk 12.

Suppose that $R, S \in \{I, B, H\}$, and that Q is one of the relations of equation (20). A relation T on U is called an *elementary information relation* if it has the form

$$xTy \Leftrightarrow \langle R(x), S(y) \rangle \in Q. \tag{23}$$

Any $\cup$, $\cap$–combination of elementary information relations is called an *information relation*. This generalizes the information relations of Orłowska (1995).

5. Example: Interrater reliability

A procedure often employed in psychological research or as a data-generating process for supervised learning algorithms in Artificial Intelligence is *expert-based categorization*: A collection of N items—such as statements, behaviour sequences, etc.—is presented to an expert, who is asked to assign each to *exactly one* of n categories C_i. If two experts solve this task, then these categories can be cross-classified in a table as follows:

Category	C_1 C_2 ... C_n
No. of agreements	k_1 k_2 ... k_n

One problem of this procedure is that experts often cannot or will not assign the items to a unique category, since statements or behavioural sequences can often be interpreted in more than one way, so that there could be more than one category to which they could be assigned. By having to assign an item to exactly one category, this information is suppressed, and, in case the experts ratings differ significantly, it cannot be said whether the experts strongly disagree, or whether the categories are not sufficiently discriminating.

In order to surmount this problem, one can offer the experts a choice among the following alternatives:

Each item is assigned to a unique category, as described above. (24)

Each item is assigned to a main category and zero or more lesser categories. (25)

Each item is assigned to one or more categories "aequo loco". (26)

We can express these situations with our RAS operationalization as follows: $U = \{E_1, \dots, E_t\}$ is the set of experts, and for each item a_i, $1 \leqslant i \leqslant N$, we let $V_{a_i} = \{C_1, \dots, C_n\}$ be the set of possible categories. The relations which we consider are I_{a_i} and B_{a_i}; their meaning is given by the following.

- $\langle E, C \rangle \in I_{a_i}$ means that expert E classifies item a_i as certainly belonging to category C (the "main" category).
- $\langle E, C \rangle \in B_{a_i}$ means that expert E classifies item a_i as possibly belonging to category C (the "lesser" categories).

The conditions (24)–(26) lead to the constraints that I_{a_i} and B_{a_i} are disjoint, that each a_i can be certainly assigned to only one category v_j, and that each expert makes at least one certain or possible assignment. In other words, we assume that

$$I_{a_i} \cap B_{a_i} = \emptyset, \tag{27}$$

$$I_{a_i}^{\smile} \circ I_{a_i} \subseteq 1_U, \tag{28}$$

$$H_{a^i} \text{ is total.} \tag{29}$$

Note that equation (27) is strictly weaker than equation (15), and that we do not prescribe incompatibility of I_a and B_a.

5.1. MEASURES OF AGREEMENT

Cohen (1960) argues that in evaluating agreement one has to take into account that agreement may be due to chance, and that one needs to remove the chance-agreement from consideration. He proposed the following measure, which we will use as a basis in the sequel:

$$\kappa = \frac{\sum_{i=1}^{n} k_i - \sum_{i=1}^{n} E[k_i]}{N - \sum_{i=1}^{n} E[k_i]}. \tag{30}$$

Here, $E[k_i]$ is the expectation of agreement under the hypothesis that the codings used by the two experts are independent. We denote by $\mathrm{ind}(A)$ the indicator function of an event A, i.e.

$$\mathrm{ind}(A) = \begin{cases} 1 & \text{if } A \text{ is true,} \\ 0 & \text{otherwise.} \end{cases}$$

Suppose we consider two experts E, E'. Different estimates of the reliability of the assignment of items to categories, i.e. their discriminating power, can be obtained by considering the following situations.

T_1: Our first situation considers only the agreement on I. Let

$$ID_1 = \sum_{j=1}^{N} \mathrm{ind}(I_{a_j}(E) = I_{a_j}(E')) \tag{31}$$

be the number of identical I_{a_j} assignments, and

$$N^* = \sum_{j=1}^{N} \mathrm{ind}(I_{a_j}(E) \neq \emptyset \quad \text{and} \quad I_{a_j}(E') \neq \emptyset) \tag{32}$$

be the number of instances where both experts make one definite choice (and possible additional ◇ entries), though not necessarily the same one. Then,

$$\kappa_1 = \frac{ID_1 - E[ID_1]}{N^* - E[ID_1]} \tag{33}$$

defines a value in analogy to κ of equation (30), which is equal to κ, if each expert makes exactly one choice for each a_j, and this choice is a ♣. A sensible interpretation of κ_1 can be given only if N^*/N is close to 1, since otherwise there are not enough non-empty I-sets.

T_2: One can sharpen the conditions by requiring that the experts agree not only on the I-values but on the B-values as well. Thus, we let

$$ID_2 = \sum_{j=1}^{N} \operatorname{ind}(I_{a_j}(E) = I_{a_j}(E') \quad \text{and} \quad B_{a_j}(E) = B_{a_j}(E')), \tag{34}$$

$$\kappa_2 = \frac{ID_2 - E[ID_2]}{N^* - E[ID_2]}. \tag{35}$$

T_3: A softer requirement than T_2, is that the experts agree on H:

$$ID_3 = \sum_{j=1}^{N} \operatorname{ind}(H_{a_j}(E) = H_{a_j}(E')), \tag{36}$$

$$\kappa_3 = \frac{ID_3 - E[ID_3]}{N - E[ID_3]}. \tag{37}$$

Note that T_3 is incomparable to T_1.

T_4: An even softer requirement is that $H(E)$ and $H(E')$ are comparable:

$$ID_4 = \sum_{j=1}^{N} \operatorname{ind}(H_{a_j}(E) \subseteq H_{a_j}(E') \text{ or } H_{a_j}(E') \subseteq H_{a_j}(E)), \tag{38}$$

$$\kappa_4 = \frac{ID_4 - E[ID_4]}{N - E[ID_4]}. \tag{39}$$

T_5: We can also only require that the certain assignments of one expert are contained in the H-set of the other:

$$ID_5 = \sum_{j=1}^{N} \operatorname{ind}(I_{a_j}(E) \subseteq H_{a_j}(E') \text{ or } I_{a_j}(E') \subseteq H_{a_j}(E)), \tag{40}$$

$$\kappa_5 = \frac{ID_5 - E[ID_5]}{N^* - E[ID_5]}. \tag{41}$$

T_6: If the assignment is reliable, we should not observe many instances of

$$H_{a_j}(E) \cap H_{a_j}(E') = \emptyset.$$

If

$$NID = \sum_{j=1}^{N} \operatorname{ind}(H_{a_j}(E) \cap H_{a_j}(E') = \emptyset), \tag{42}$$

Table 6
κ-relations

T	II	IB	IH	BI	BB	BH	HI	HB	HH
1	$=$								
2	$=$				$=$				
3									$=$
4					or				$\subseteq$ $\supseteq$
5			$\subseteq$		or		$\supseteq$		
6									D

we define

$$\kappa_6 = 1 - \frac{NID}{E[NID]}. \tag{43}$$

These situations correspond to the relations $T_1, \ldots, T_6$ depicted in Table 6. Observe that we have combined $\subsetneq$ ($\supsetneq$) and $=$ into $\subseteq$ ($\supseteq$). If both experts use only the first coding alternative (exact assignments – the "classical approach"), no differences among the 6 relations $T_1, \ldots, T_6$ will occur, up to the point where the objects which fulfil T_6 are in the set complement of the set built by one of the relations $T_1, \ldots, T_5$.

5.2. APPLICATION: SOFTWARE USABILITY

Gediga, Hamborg and Düntsch (1999) present an instrumentarium for the evaluation of software usability which contains 75 questions rating the seven usability categories of ISO 9241-10. These are as follows.

(1) Suitability for the task,
(2) Self-descriptiveness,
(3) Controllability,
(4) Conformity with user expectations,
(5) Error tolerance,
(6) Suitability for individualization,
(7) Suitability for learning.

In this case, we have 75 attributes a_i, each with the categories $C_1, \ldots, C_7$ which correspond to the seven usability criteria listed above.

TABLE 7
Experimental results

	T_1	T_2	T_3	T_4	T_5	T_6
$ID_{1..5}$ and NID_6	50	32	35	65	70	6
Expectation	9.927	3.724	4.320	18.395	33.219	47.772
κ_i	0.635	0.397	0.235	0.823	0.925	0.874
Significance	0.001	0.001	0.001	0.001	0.001	0.001
%	66.7	42.7	46.7	86.7	93.3	8.0

We have asked two experts to assign categories to each of these questions, using the semantic constraints (27)–(29). It turns out that $N^* = 73$, which is sufficiently close to $N = 75$.

In order to find out the expectations of the κ_i-statistics and their significance, 1000 simulated (pseudo-) random agreements using the random labelling technique were drawn. The expectation was estimated by the mean of the simulated $\kappa_i(\sigma)$ statistics. The significance of the empirical κ_i is estimated by

$$\text{significance} = \frac{1 + \text{number of } \kappa_i(\sigma) \geqslant \kappa_i}{1001}.$$

The values for the various IDs, significance and expectations after 1000 simulations, and κ corresponding to T_1–T_6 are shown in Table 7. Note that the column headed "6" lists the results for NID. We also give the percentages of (dis-) agreement.

The relation T_1 is fulfilled in 2/3 of all instances, which means that 50 items of the test are assigned to the same ♣-category by both raters. In analogy to the classical procedure, we can regard a value of $\kappa_1 = 0.635$ as "GOOD" (Robson, 1993). Whereas the analysis of T_1 is approximately the same as the classical procedure, the other types of relations offer different insights. The strong equality T_2 holds in 32 (42.7%) of the cases, and the hull-equality T_3 is given in 35 cases (46.7%). Both results tell us that the assignment of the ◇-value is by far less stable than the assignment of the ♣-value. The values of κ_2 and κ_3 show that the difference of the resulting equalities from those which can be achieved by random are much smaller than in case of T_1.

Looking at T_4 we observe that the "equality up to different strictness" describes the situation quite well, because the ratings of 65 items (86.7%) can be described in that way.

Relation T_5 holds for 70 cases (93.3%), which means that at least one ♣-category of one rater is at least mentioned by the other rater—the other five items (6.7%) are of interest, because of obvious disagreement.

Finally, T_6 holds for 6 items (8.0%), which means that the experts totally disagree on only a few items. Note that T_6 is stricter than T_5 if $N^* = N$; if this condition does not hold (as in our example), T_6 and T_5 address different relationships.

6. Summary and outlook

We have investigated semantic interpretations of multi-valued information systems, and have proposed a relational operationalization which enables the researcher to express a distinction between certain and (im-)possible facts or events. In terms of methodology, the proposed procedures are in the "non-invasive" spirit of data analysis (Düntsch & Gediga, 2000), and they integrate the characteristics of rough set and fuzzy set analysis in a straightforward manner. Our approach shows connections to ideas in spatial reasoning research; we have shown what kind of relations can be set up in this general framework and how these relations are related to the egg-yolk representation of boundary regions in spatial reasoning.

In an example of our approach, we have shown how to generalize traditional methods of expert-based classification, and that it is possible, without using many additional resources, to obtain a more detailed picture of the interplay of the raters' choices, and to explain previously hidden differences. The main advantage of the new classification scheme is that we have a better chance of understanding why experts disagree in categorization, and in which cases a compromise among experts may or may not be feasible. Furthermore, the application of this technique in Artificial Intelligence enables researchers to perform more refined comparisons of expert opinions with results of algorithms in the context of supervised learning.

Even though there are other formalisms which deal with operationalizations of uncertainty caused by boundary regions such as various modal and default logics or possibilistic systems, we believe that the simplicity and transparency of our relational operationalization as well as its generality and applicability in many areas of research make it a novel and viable alternative to known procedures. Furthermore, relational methods such as those which we have used for semantic constraints are well understood and have a respected mathematical foundation; being equational in nature, they are also easily implemented. It is, of course, not our intention to dismiss all other approaches to the problem. Our system is geared towards the minimization of model assumptions, and can always be used at least as a first step in data analysis. If sufficient reliable domain knowledge is present, stronger methods may lead to better results; it should, however, not be forgotten, that the outcome must be interpreted in the context of the model assumptions.

We are currently undertaking an investigation of the logical background of the presented structures, based on the relational semantics of Orłowska (1996) and the extensions given by Düntsch, McCaull and Orłowska (2000). We are also carrying out more detailed case studies to explore the psychometrical implications of the many-valued expert agreement procedure, based on the description of the fine structure of the possibilities.

The authors gratefully acknowledges support by the KBN/British Council Grant No. WAR/992/174 and COST Action 274 "Theory and Applications of Relational Structures as Knowledge Instruments" (www.tarski.org). We should like to thank the anonymous referees for their constructive and helpful comments.

References

Bosc, P. & Prade, H. (1994). An introduction to fuzzy set and possibility theory-based approaches to the treatment of uncertainty and imprecision in data base management systems. In A. Motro & P. Smets, Eds. *Proceedings of Uncertainty Management in Information Systems: from Needs to Solutions Workshop (UMIS'94)*.

COHEN, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20,** 37–46.

COHN, A. G. & GOTTS, N. M. (1996). The ‘egg-yolk’ representation of regions with indeterminate boundaries. In P. BURROUGH & A. M. FRANK, Eds. *Proceedings of the GISDATA Specialist Meeting on Geographical Objects with Undetermined Boundaries*, pp. 171–187. London: Taylor & Francis.

DOIGNON, J. P. & FALMAGNE, J. C. (1999). *Knowledge Spaces*. Berlin: Springer-Verlag.

DÜNTSCH, I. & GEDIGA, G. (1995). Skills and knowledge structures. *British Journal of Mathematical and Statistical Psychology*, **48,** 9–27.

DÜNTSCH, I. & GEDIGA, G. (2000). *Rough set data analysis: A road to non-invasive knowledge discovery*. *Methoδos Primers*, Vol. 2. Bangor: UK. Methoδos Publishers.

DÜNTSCH, I., MCCAULL, W. & ORŁOWSKA, E. (2000). Structures with many-valued information and their relational proof theory. *Proceedings of the 30th IEEE International Symposium on Multiple-Valued Logic*, pp. 293–301.

EGENHOFER, M. (1994). Deriving the composition of binary topological relations. *Journal of Visual Languages and Computing*, **5,** 133–149.

EGENHOFER, M. & FRANZOSA, R. (1991). Point–set topological spatial relations. *International Journal of Geographic Information Systems*, **5,** 161–174.

FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7,** 179–188.

GEDIGA, G., HAMBORG, K.-C. & DÜNTSCH, I. (1999). The Isometrics usability inventory: An operationalisation of ISO 9241/10. *Behaviour and Information Technology*, **18,** 151–164.

GIGERENZER, G. (1981). *Messung und Modellbildung in der Psychologie*. Basel: Birkhäuser.

IWINSKI, T. B. (1988). Contraction of attributes. *Bulletin of the Polish Academy Sciences and Mathematics*, **36,** 623–632.

JACCARD, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin de la Society Vaudoise des Sciences Naturelles*, **44,** 223–270.

LIPSKI, W. (1976). Informational systems with incomplete information. In S. MICHAELSON & R. MILNER, Eds. *Third International Colloquium on Automata, Languages and Programming*, pp. 120–130. Edinburgh: University of Edinburgh, Edinburgh University Press.

LIPSKI, W. (1981). On databases with incomplete information. *Journal of the ACM*, **28,** 41–70.

ORŁOWSKA, E. (1995). Information algebras. In *Proceedings of AMAST 95*, *Lecture Notes in Computer Science*, Vol. 639. Springer-Verlag.

ORŁOWSKA, E. (1996). Relational proof systems for modal logics. In H. WANSING, Ed. *Proof Theory of Modal Logic*, pp. 55–78. Dordrecht, Kluwer.

ORŁOWSKA, E. & PAWLAK, Z. (1987). Representation of nondeterministic information. *Theoretical Computer Science*, **29,** 27–39.

PAWLAK, Z. (1982). Rough sets. *International Journal of Computing and Information Sciences*, **11,** 341–356.

PRADE, H. (1984). Lipski’s approach to incomplete information data bases restated and generalized in the setting of zadeh’s possibility theory. *Information Systems*, **9,** 27–42.

ROBSON, C. (1993). *Real World Research: A Resource For Social Scientist and Practioner Researchers*. Oxford: Blackwell.

TUCHOLSKY, K. (1975). Zur soziologischen Psychologie der Löcher. In M. GEROLD-TUCHOLSKY & F. J. RADDATZ, Eds. *Kurt Tucholsky, Gesammelte Werke*, Vol. 9, pp. 152–153. Hamburg: Rowohlt Taschenbuch Verlag.

WILLE, R. (1982). Restructuring lattice theory: an approach based on hierarchies of concepts. In I. RIVAL, Ed. *Ordered Sets*, *NATO Advanced Studies Institute*, Vol. 83, pp. 445–470. Dordrecht: Reidel.

Paper accepted for publication by Associate Editor, E. H. Mamdani