Relational attribute systems

Ivo Düntsch[*] [†]
School of Information and Software Engineering
University of Ulster
Newtownabbey, BT 37 0QB, N.Ireland
I.Duentsch@ulst.ac.uk

Günther Gediga[*]
FB Psychologie / Methodenlehre
Universität Osnabrück
49069 Osnabrück, Germany
Guenther@Gediga.de

Ewa Orłowska[*] [†]
Institute of Telecommunications
Szachowa 1
04–894, Warszawa, Poland
orlowska@itl.waw.pl

"Das Merkwürdigste an einem Loch ist der Rand."[1] [24]

Abstract

We introduce a relational operationalisation of data which generalises, among others, the deterministic information systems of [22], the indeterministic systems of [15] and [20], and the context relation of [26]; it can also be used for fuzzy data modelling. Using an example from the area of psychometrics, we show how our operationalisation can lead to an improved understanding of agreements and disagreements by experts in classification tasks.

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 given domain, which extends the widely used data table operationalisation [1]. It turns out that relations between objects and features are a suitable tool to achieve our aim. Using the 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,

[*] The order of authors is alphabetical, and equal authorship is implied.

[†] The author gratefully acknowledges support by the KBN/British Council Grant No WAR/992/174.

[1] "The most peculiar thing about a hole is its boundary."

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

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 setup. For example, the relation

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

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 organised as follows: In the first section we recall several modes of operationalisation which have appeared in the literature and their model assumptions. Section 3 introduces our relational operationalisation 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 operationalisation

When a domain of interest is investigated, one needs to introduce a language which possibly includes relation and/or operator symbols with which the properties of the domain can be described. This process is called "operationalisation" in the Social Sciences, and "knowledge representation" in Artificial Intelligence; its result is sometimes called an "empirical model" [12].

One of the oldest operationalisations of data is the

(2.1) $$\text{OBJECT} \mapsto \text{ATTRIBUTES}$$

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 [12]. A data array as shown in Table 1 is an example of such an operationalisation: The leftmost column denotes different specimen of Iris flowers, while each of the other columns describe one property (attribute) of each specimen.

A formal version of this type of operationalisation is the following [21]: A *single valued information system* is a structure

(2.2) $$\mathcal{I} = \langle U, \Omega, \{V_a : a \in \Omega\}\rangle,$$

where

- U is a finite set of objects.

Table 1: Fisher's Iris data [10]

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>					

- 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.

Any such operationalisation puts semantic constraints on the data set. A simple and widely used assumption 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. It follows from the assumption that each attribute is a function.

Given an information system $\mathcal{I}$ as above, Iwinski [13] calls an information system $\mathcal{I}' = \langle U', \Omega', \{V_v : v \in \Omega'\}\rangle$ a *decomposition of* $\mathcal{I}$, if

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 $f_a : \Omega_a \to \{a(x) : x \in U\}$.
   (b) For all $v \in \Omega_a, x \in U$,

$$v(x) = 1 \iff a(x) = f(v). \tag{2.3}$$

This procedure is called *binarisation* in [4] and [25].

Consider, for example the information system of Table 2, which, for simplicity, has only one attribute "Size" [excerpt from 26]; the decomposition of $\mathcal{I}$ 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 $a(x) =$ 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

Table 2: Planet system

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

hand, the attribute a is "colour", then being not red does usually not imply the presence of a particular colour. This type of binary attribute is called *asymmetric* [see 14]. This semantic information needs to be present in any data operationalisation.

Wille [26] operationalises a single-valued information system by taking its binarisation, 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 operationalisation of the data is given by the context relation $I \subseteq U \times \Omega'$ containing the pairs

$$\langle\text{Mercury,small}\rangle, \langle\text{Venus,small}\rangle, \langle\text{Earth,small}\rangle, \langle\text{Pluto,small}\rangle,$$
$$\langle\text{Uranus,medium}\rangle, \langle\text{Neptune,medium}\rangle,$$
$$\langle\text{Jupiter,large}\rangle, \langle\text{Saturn,large}\rangle.$$

A generalisation of single-valued information systems which could indicate incompleteness was introduced by Lipski [16, 17]:

> "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" [17].

These considerations lead to the following definition: A *multi-valued information system* is a structure

$$\mathcal{I} = \langle U, \Omega, \{V_a : a \in \Omega\}\rangle, \tag{2.4}$$

where

- 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 [20], 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

   (2.5) x speaks German, Polish, and French and no other languages.

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

   (2.6) a speaks German, Polish, and French and possibly other languages.

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

   (2.7) The car that went too fast was either a Mercedes or a Ford.

   Here, exactly one of the statements

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

   is true, but it is not known which one.

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

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

   means that Ivo cooperates with Günther, or Ewa, or both.

3 Relational attribute systems

In this section we shall unify the operationalisations 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

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

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 V_A.

The attributes of a single-valued information system are functions $U :\to V_a$, while the attributes of a multivalued 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

$$x R_a v \iff v \in a(x). \tag{3.2}$$

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

While operationalisations such as those of [21] or [20] are not (openly) concerned with semantic constraints as part of the design process of an information system and only learn the given data, we will need to 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 operationalisation; in particular, we need 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

$$\mathcal{I} = \langle U, \{\Omega_a : a \in T\}, \mathcal{R}, \Delta\rangle, \tag{3.3}$$

where

1. U is a non-empty set of objects.
2. Each Ω_a is a non-empty set of attribute values, and the sets Ω_a are pairwise disjoint; we set $\Omega = \bigcup_{a \in T} \Omega_a$.
3. $\mathcal{R}$ is a set of relations such that for each $R \in \mathcal{R}$ there is some $a \in T$ with $R \subseteq U \times \Omega_a$.
4. Δ is a set of semantic constraints.

Each $a \in T$ is an attribute, and each $v \in \Omega_a$ a value which an object $x \in U$ can take under a, which is the a-property of x. The relations in $\mathcal{R}$ express our knowledge about the connection of $x \in U$ with the properties of a, and the constraints describe the type of operationalisation, 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 operationalisations from above can be found in our systems. Suppose that $\mathcal{I}$ is a single-valued information system. For each $a \in \Omega$ we let $\Omega_a = \{a(x) : x \in U\}$; we assume without loss of generality that the sets Ω_a are pairwise disjoint. $I_a \subseteq U \times \Omega_a$ is now defined by

$$x I_a v \iff a(x) = v. \tag{3.4}$$

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 $x R_a v$. It is not hard to see that this condition is equivalent to the relational equations

(3.5) $I_a \circ V_{\Omega_a} = U \times \Omega_a$ I_a is total.

(3.6) ${I_a}^{\smile} \circ I_a \subseteq 1'_{\Omega_a}$ I_a is functional.

For the situation of (2.5), we let a be the property "speaking a language" and

(3.7) $$x\, I_a\, v \iff x \text{ speaks language } v.$$

There are no constraints; however, if we want to prescribe that each person speaks at least one language, then we will have constraint (3.5). We notice how our relational notation allows us to generalise 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 $x\, B_a\, v$ as "x has possibly the a-property v". The constraints arising from Lipski's systems are

(3.8) ${B_a}^{\smile} \circ I_a = \emptyset$ "Certainly" and "Possibly" are not compatible.

(3.9) ${I_a}^{\smile} \circ I_a \subseteq 1'_{\Omega_a}$ I_a is functional.

This is reminiscent of fuzzy sets in that we do not necessarily have crisp attribute assignments, and also of rough sets, since we have only one relation per attribute for uncertainty. Note that condition (3.8) implies that $I_a \cap B_a = \emptyset$, and that $x\, I_a\, v_1$ and , $x\, B_a\, v_2$ are together impossible. For (2.8) we note that there are no semantic constraints.

Even though we will concentrate in the sequel on the relations I and B, these are by no means the only conceivable ones. Another frequently used relation is the one which signals 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 $\mathcal{R}$; this generalises the dependencies of rough set theory, and the information relations of [18]. More concretely, we shall consider the case of the relations I_a and B_a as described in the previous section, i.e.

- $x\, I_a\, v$ means that x certainly has the a-property v.
- $x\, B_a\, v$ 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, and Ω_a.

Since I and B signal the (possible) presence of a property, all attributes are seen to be asymmetric. 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 the Ω. If $x\,I\,v$, we place ♣ into the cell $\langle x\,,v\rangle$, and for $x\,B\,v$ we write ♢. Using this notation, Lipski's conditions (3.8) and (3.9) can be stated equivalently as

(4.1) ♣ and ♢ cannot appear in the same row.

(4.2) There is at most one ♣ in every row.

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 [22], or to the egg-yolk model of [3], where

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

Our overall constraint is

$$I \cap B = \emptyset. \tag{4.3}$$

In rough set theory, two objects in a single-valued information system are called *indiscernible*, if they have the same feature vector. In a multivalued system there are other possibilities which use set theoretic relations on the sets $a(x)$. This leads to the *information relations* first studied in [18]. Our relational setting extends these relations in the following way: We will consider the relations

$$=,\ \subsetneq,\ \supsetneq,\ O,\ D, \tag{4.4}$$

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

$$\begin{aligned} &tO\,u \iff t \cap u \neq \emptyset, \text{ and } t \text{ and } u \text{ are incomparable with respect to } \subseteq,\\ &tD\,u \iff t \cap u = \emptyset. \end{aligned}$$

Then, the relations of (4.4) partition $M \times M$. Such "intersection tables" have been considered in qualitative spatial reasoning, for example, in [8, 9] for the interior I and boundary B of sets in a topological space. In Tucholsky's terms, the interior corresponds to the hole, and the boundary is the uncertainty, the investigation of which is 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

$$II,\ IB,\ IH,\ BI,\ BB,\ BH,\ HI,\ HB,\ HH. \tag{4.5}$$

We can now construct a relational table by indicating below each heading which of the relations of (4.4) holds. Of course, not all configurations are possible, since we have to observe the conditions

$$H = I \cup B \text{ and } I \cap B = \emptyset. \tag{4.6}$$

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	$=$	$\subsetneq$
$H(x) = H(y)$			$\subsetneq$			$\subsetneq$	$\supsetneq$	$\supsetneq$	$=$

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 78 arrangements, which are possible when we disregard the columns which contain H are shown in Table 5 on the following page. The EY column gives the number(s) of the corresponding egg-yolk configuration(s) as listed in [3, Figure 4]. Since several egg-yolk pairs can belong to the same I, B – configuration, and not every I, B – configuration is associated with an I, H – configuration, we see that the expressive powers of I, B – configurations and I, H – configurations are incomparable.

Suppose that $R, S \in \{I, B, H\}$, and that Q is one of the relations of (4.4). A relation T on U is called an *elementary information relation* if it has the form

$$xTy \iff \langle R(x), S(y)\rangle \in Q. \tag{4.7}$$

Any $\cup, \cap$ – combination of elementary information relations is called an *information relation*. This generalises the information relations of [18].

5 Example: Interrater reliability

A procedure often employed in psychological research is *expert-based categorisation*: A collection of N items – such as statements, behaviour sequences etc – are presented to an expert, who is asked to assign each one 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	$\dots$	C_n
No. of agreements:	k_1	k_2	$\dots$	k_n

A measurement which is frequently used to express the agreement is

$$\kappa = \frac{\sum_{i=1}^{n} k_i - \sum_{i=1}^{n} E[k_i]}{N - \sum_{i=1}^{n} E[k_i]}, \tag{5.1}$$

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	$\subsetneq$	
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	$\subsetneq$	$\subsetneq$	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		
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							

introduced by Cohen [2]. Here, $E[k_i]$ is the expectation of agreement under the hypothesis that the codings used by the two experts are independent.

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:

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

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

(5.4) Each item is assigned to one or more categories “aequo loco”.

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

- $\langle E, C\rangle \in I_{a_i}$ means that expert E classifies item a_i as certainly belonging to category C.
- $\langle E, C\rangle \in B_{a_i}$ means that expert E classifies item a_i as possibly belonging to category C.

The conditions (5.2) – (5.4) 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

(5.5) $$I_{a_i} \cap B_{a_i} = \emptyset,$$

(5.6) $$I_{a_i}{}^{\smile} \circ I_{a_i} \subseteq 1_U,$$

(5.7) H_{a^i} is total.

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

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

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

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

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

(5.10) $$\kappa_1 = \frac{ID_1 - E[ID_1]}{N^* - E[ID_1]}$$

defines a value in analogy to κ of (5.1), which is equal to κ, if each expert makes exactly one choice for each a_j, and this choice is a $\clubsuit$. A sensible interpretation of κ_1 can be given only if $\frac{N^*}{N}$ 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 the I-values but on the B-values as well. Thus, we let

(5.11) $$ID_2 = \sum_{j=1}^{N} \mathrm{ind}(I_{a_j}(E) = I_{a_j}(E') \text{ and } B_{a_j}(E) = B_{a_j}(E')),$$

(5.12) $$\kappa_2 = \frac{ID_2 - E[ID_2]}{N^* - E[ID_2]}.$$

T_3**:** A softer requirement than 2. is that the experts agree on H:

$$ID_3 = \sum_{j=1}^{N} \mathrm{ind}(H_{a_j}(E) = H_{a_j}(E')), \tag{5.13}$$

$$\kappa_3 = \frac{ID_3 - E[ID_3]}{N - E[ID_3]}. \tag{5.14}$$

Note that T_3 is incomparable to T_1.

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

$$ID_4 = \sum_{j=1}^{N} \mathrm{ind}(H_{a_j}(E) \subseteq H_{a_j}(E') \text{ or } H_{a_j}(E') \subseteq H_{a_j}(E)), \tag{5.15}$$

$$\kappa_4 = \frac{ID_4 - E[ID_4]}{N - E[ID_4]}. \tag{5.16}$$

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

$$ID_5 = \sum_{j=1}^{N} \mathrm{ind}(I_{a_j}(E) \subseteq H_{a_j}(E') \text{ or } I_{a_j}(E') \subseteq H_{a_j}(E)), \tag{5.17}$$

$$\kappa_5 = \frac{ID_5 - E[ID_5]}{N^* - E[ID_5]}. \tag{5.18}$$

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

$$NID = \sum_{j=1}^{N} \mathrm{ind}(H_{a_j}(E) \cap H_{a_j}(E') = \emptyset), \tag{5.19}$$

we define

$$\kappa_6 = 1 - \frac{NID}{E[NID]}. \tag{5.20}$$

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 ... T_6 will occur, up to the point that the objects which fulfil T_6 are in the set complement of the set built by one of the relations $T_1, \ldots, T_5$.

Gediga et al. [11] present an instrumentarium for the evaluation of software usability which contains 75 questions rating the seven usability categories of ISO 9241-10. These are

Table 6: κ-relations

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

1. Suitability for the task,
2. Selfdescriptiveness,
3. Controllability,
4. Conformity with user expectations,
5. Error tolerance,
6. Suitability for individualisation,
7. Suitability for learning.

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

We have asked two experts to assign categories to each of these questions, using the semantic constraints (5.5) – (5.7). It turns out that $N^* = 73$, which is sufficiently close to $N = 75$. The values for the various IDs, expectations (E) after 1000 simulations, and κ corresponding to T_1 to T_6 are shown in Table 7 on the next page. Note that the column headed "6" lists the results for NID. We also give the significance α after 1000 simulations, and 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" [23]. 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

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

and κ_3 show that the difference of the resulting equalities to 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 5 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 multivalued information systems, and have proposed a relational operationalisation 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 [5], and 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 uncertainty in spatial reasoning.

In an example of our approach, we have shown how to generalise 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. A main advantage of the new classification scheme is that we have a better chance of understanding why experts disagree in categorisation, and in which cases a compromise among experts is feasible or not.

We are currently undertaking an investigation of the logical background of the presented structures [6], based on the relational semantics of [19], and more detailed case studies to gauge the possibilities

and limits of the concepts [7].

References

[1] Codd, E. F. (1970). A relational model of data for large shared databanks. *Comm. of the ACM*, **13**, 377–387.

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

[3] Cohn, A. G. & Gotts, N. M. (1996). The 'egg-yolk' representation of regions with indeterminate boundaries. In P. Burrough & A. M. Frank (Eds.), *Proc. of the GISDATA Specialist Meeting on Geographical Objects with Undetermined Boundaries*, 171–187. Francis Taylor.

[4] Düntsch, I. & Gediga, G. (1998). Simple data filtering in rough set systems. *International Journal of Approximate Reasoning*, **18**, 93–106.

[5] Düntsch, I. & Gediga, G. (2000). ROUGHIAN – Rough Information Analysis. *International Journal of Intelligent Systems*. To appear.

[6] Düntsch, I., Gediga, G. & Orłowska, E. (1999a). A logic for relational attribute systems. In preparation.

[7] Düntsch, I., Gediga, G. & Orłowska, E. (1999b). Measuring the reliability of expert assignment. In preparation.

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

[9] Egenhofer, M. & Franzosa, R. (1991). Point–set topological spatial relations. *International Journal of Geographic Information Systems*, **5**, 161–174.

[10] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, **7**, 179–188.

[11] 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.

[12] Gigerenzer, G. (1981). Messung und Modellbildung in der Psychologie. Basel: Birkhäuser.

[13] Iwinski, T. B. (1988). Contraction of attributes. *Bull. Polish Acad. Sci. Math.*, **36**, 623–632.

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

[15] Lipski, W. (1976). Informational systems with incomplete information. In S. Michaelson & R. Milner (Eds.), *Third International Colloquium on Automata, Languages and Programming*, 120–130, University of Edinburgh. Edinburgh University Press.

[16] Lipski, W. (1979). On semantic issues connected with incomplete information data bases. *ACM Trans. Database Systems*, **4**, 262–296.

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

[18] Orłowska, E. (1995). Information algebras. In *Proceedings of AMAST 95*, vol. 639 of *Lecture Notes in Computer Science*. Springer–Verlag.

[19] Orłowska, E. (1996). Relational proof systems for modal logics. In H. Wansing (Ed.), *Proof theory of modal logic*, 55–78. Dordrecht: Kluwer.

[20] Orłowska, E. & Pawlak, Z. (1987). Representation of nondeterministic information. *Theoretical Computer Science*, **29**, 27–39.

[21] Pawlak, Z. (1973). Mathematical foundations of information retrieval. ICS Research Report 101, Polish Academy of Sciences.

[22] Pawlak, Z. (1982). Rough sets. *Internat. J. Comput. Inform. Sci.*, **11**, 341–356.

[23] Robson, C. (1993). Real World Research: A resource for social scientist and practioner researchers. Oxford: Blackwell.

[24] Tucholsky, K. (1975). Zur soziologischen Psychologie der Löcher. In M. Gerold-Tucholsky & F. J. Raddatz (Eds.), *Kurt Tucholsky, Gesammelte Werke*, vol. 9, 152–153. Hamburg: Rowohlt Taschenbuch Verlag.

[25] Wang, H., Düntsch, I. & Gediga, G. (2000). Classificatory filtering in decision systems. *International Journal of Approximate Reasoning*. To appear.

[26] Wille, R. (1982). Restructuring lattice theory: an approach based on hierarchies of concepts. In I. Rival (Ed.), *Ordered sets*, vol. 83 of *NATO Advanced Studies Institute*, 445–470. Dordrecht: Reidel.