PROCEEDINGS OF THE INTERNATIONAL CONFERENCE
ON FIFTH GENERATION COMPUTER SYSTEMS 1984,
edited by ICOT.

DAL - A LOGIC FOR DATA ANALYSIS

Luis FARINAS DEL CERRO

LSI Université Paul SABATIER
31062 TOULOUSE CEDEX

FRANCE

Ewa ORLOWSKA

Institute of Computer Science
Polish Academy of Sciences
00-901 WARSZAWA
POLAND

ABSTRACT

The aim of this paper is given some ideas how to analyze data using logic tools.

1 INTRODUCTION

The aim of this paper is to define a logic which enables us reasoning about data. Data are considered to be things or entities and properties meaningful for these entities. One of the main problems in data analysis is to induce patterns in a set of data items (Benzecri (1973)). This problem consists of two tasks :

(1) To agregate data into sets which can be adequately characterised by means of some of the given properties.

(2) For a set of data given a priori to choose those properties from a given set of properties which are adequate for defining this set.

That is the scheme of task (1) is: from properties to sets of data items and the scheme of task (2) is : from sets of data items to properties.

The similar problems are considered in the fields of information systems (Pawlak (1981)), knowledge representation (Pawlak (1983), Konrad et al (1981), Orlowska (1983)), and pattern recognition (FU (1974)).

In our approach the formal counterparts of data are a non-empty set of objects and a family of equivalence relations on this set. Objects will be interpreted as data items and relations correspond to properties of data items. Namely, each property induces an equivalence relation such that an equivalence class of the relation consists of those objects which are the same with respect to this property.

In section 2 we give a detailed explanation of (1) and (2) on the level of semantical structure. In section 2 we introduce a syntactic structure to be used to represent the given semantical requirements. We present a language in which facts concerning tasks (1) and (2) can be formulated in a formal way. Such approach is necessary if reasoning is to be carried out by a computer. Next, we provide a deduction method for the language which enables us to prove facts concerning tasks (1) and (2).

Our approach follows the ideas and methods developed in Pawlak (1982), Orlowska and Pawlak (1981), Orlawska (1983), Harel (1978) and Mirkowska (1981).

2 DEFINABILITY OF DATA

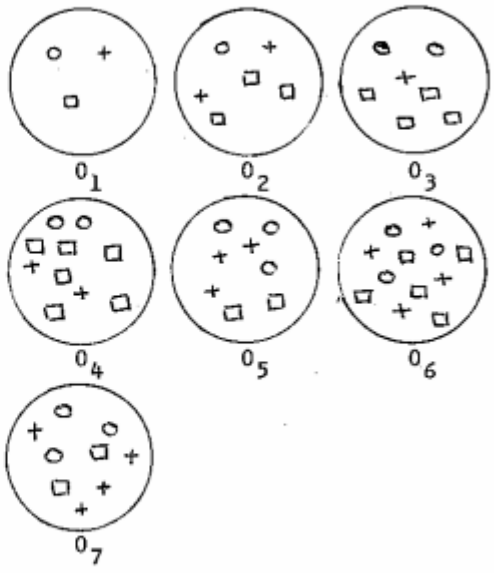
In this section we present a formal framework which enables us to express facts concerning relationships

between data items and properties of data items. We use the basic notions introduced in Konrad, Orlowska, and Pawlak (1981), Pawlak (1982), Orlowska (1983). The new idea is to consider what is called strong definability of sets of data items. It enables us to reflect an adequacy of properties for characterization of sets of data.

We consider a non-empty set OB whose elements are interpreted as data items. The elements of this set are referred to as objects. Each property meaningful for objects from set OB induces an equivalence relation in set OB. Namely, two objects are in the relation corresponding to a property if they cannot be distinguished by means of this property. Equivalence relations in set OB are referred to as indiscernibility relations. For example, if a characteristics "colour" is meaningful for elements of a set OB then we consider indiscernibility relation "to be of the same colour" which enables us to agregate objects into classes according to their colour. These classes are equivalence classes of the indiscernibility relation in question.

Given a pair of indiscernibility relations, say R and P, we consider the intersection $R \cap P$ and the transitive closure $R \circledast P$ of these relations. Clearly $R \cap P$ and $R \circledast P$ are equivalence relations. Equivalence classes of $R \cap P$ and $R \circledast P$ are obtained from equivalence classes of R and P by making their intersections and unions, respectively. Relations $R \cap P$ and $R \circledast P$ can be considered to be indiscernibility relations corresponding to properties which are composed in some way from properties related to R and P. Let us consider a simple example.

Example 2.1. Assume that we are given seven objects :



The natural properties meaningful for these objects are : number of circles, number of crosses, number of squares. Indiscernibility relations R_o R_+ and $R_\square$ corresponding to these properties provide the following classes :

R_o $\{0_1,0_2\}$ $\{0_3,0_4\}$ $\{0_5,0_6,0_7\}$

R_+ $\{0_1,0_3\}$ $\{0_2,0_4\}$ $\{0_6,0_7\}$ $\{0_5\}$

$R_\square$ $\{0_1\}\{0_2\}\{0_3\}\{0_4\}\{0_5,0_7\}$ $\{0_6\}$

The classes determined by relation $R_o \cap R_+$ are as follows :

$R_o \cap R_+ \{0_1\}\{0_2\}\{0_3\}\{0_4\}\{0_5\}\{0_6,0_7\}$

The classes reflect the pattern corresponding to property "to have the same number of circles and the same number of crosses".

The classes corresponding to relation $R_o \circledast R_+$ are as follows :

$R_o \circledast R_+ \{0_1,0_2,0_3,0_4\}\{0_5,0_6,0_7\}$

The property corresponding to this pattern can be expressed as "number of crosses and number of circles less than 3".

In general, a property corresponding to the intersection of indiscerni-

bility relations R and P can be defined as a conjunction of properties corresponding to R and P. In the case of $R \circledcirc P$ there is no such straightfoward relationship between the underlying properties. However the property corresponding to $R \circledcirc P$ can be defined as a certain relation between properties corresponding to R and P.

Given an indiscernibility relation in a set OB of data items, we consider the problem of definability of subsets of set OB in terms of the property corresponding to this relation. Let X and R be a subset of set OB and an indiscernibility relation in set OB, respectively.

A lower approximation $\underline{R}X$ of set X with respect to relation R is a union of those equivalence classes of R which are included in X.

An upper approximation $\overline{R}X$ of set X with respect to relation R is a union of those equivalence classes of R which have an element in common with X.

In terms of approximations we can define positive, negative, and borderline instances of a set, namely :

$\underline{R}X$ set of positive instances of X with respect to R

$OB-\overline{R}X$ set of negative instances of X with respect to R

$\overline{R}X-\underline{R}X$ set of borderline instances of X with respect to R.

By using the notions given above we can express facts concerning the kind of data analysis in which we are interested in the structure within a given a priori set of data. In this kind of task we assume that we are given a subset X of data items determined by means of a certain external condition and we are interested in establishing a relationship between X and the structuring of data provided by indiscernibility relations.

We say that set X is definable with respect to indiscernibility R iff $\underline{R}X = X = \overline{R}X$.

In other words a set is definable with respect to indiscernibility relation iff it can be covered by equivalence classes of this relation. This means that a pattern provided by X can be expressed by means of a property corresponding to this indiscernibility. If a set X is not definable with respect to an indiscernibility R then we can express the pattern given by X with a certain degree of inexactness. Set of positive instances of X with respect to R is the greatest definable set included in X which represents the pattern corresponding to X, set of negative instances of X is the greatest definable set whose elements do not obey this pattern, set of borderline instances consists of elements for which we cannot decide whether they obey the pattern or not.

We say that set X is strongly definable with respect to indiscernibility R iff $\underline{R}X = X = \overline{R}X$ and X is an equivalence class of R.

We make distinction between a possibility of covering a set by one equivalence class and by more than one class to reflect an adequacy or properties for characterization of sets of data. If a set is strongly definable with respect to a relation then we will consider a property corresponding to this relation to be adequate for describing the set.

Example 2.2. Let us consider the set $OB = \{O_1, O_2, \ldots, O_7\}$ given in example 2.1., and indiscernibility relations R_0, R_+, and $R_\square$. Set $X = \{O_1, O_2, O_3, O_4\}$ is definable with respect to R_0, R_+, and

$R_\Box$, but it is not strongly definable with respect to these relations. However X is strongly definable with respect to relation $R_0 \circledast R_+$. Set $Y = \{0_1, \ldots, 0_2, 0_7\}$ is not definable neither by R_0, R_+, and $R_\Box$, nor by $R_0 \cap R_+$, and $R_0 \cap R_\Box$. Set Y is definable with respect to $R_+ \cap R_\Box$, but none of the relations obtained from R_0, R_+, and $R_\Box$ by using operations $\circledast$ and $\cap$ is sufficient to provide the strong definability of Y. The approximations of set Y with respect to some of the indiscernibility relations in question are given below.

- $\underline{R}_0 Y = \{0_1, 0_2\}$
- $\overline{R}_0 Y = \{0_1, 0_2, 0_5, 0_6, 0_7\}$ positive instances of Y with respect to R_0 : $0_1, 0_2$. Negative instances of Y with respect to R_0 : $0_3, 0_4$. Borderline instances of Y with respect to R_0 : $0_5, 0_6, 0_7$
- $\underline{R_0 \cap R_+} Y = \{0_1, 0_2\}$
- $\overline{R_0 \cap R_+} Y = \{0_1, 0_2, 0_6, 0_7\}$

In terms of the notions defined above we can discuss the ability or properties to define sets of data. We can consider a certain property P to be better than a property P' for characterization of a set X of data iff the approximations of X with respect to the indiscernibility determined by P are closer to X (with respect to inclusion) than the approximations with respect to the indiscernibility corresponding to P'.

In the following we list some properties of approximations.

$\underline{R}X \subseteq X$

$\underline{R}\underline{R}X = \underline{R}X$

$\underline{R}(X \cap Y) = \underline{R}X \cap \underline{R}Y$

$\underline{R}X \cup \underline{R}Y \subseteq \underline{R}(X \cup Y)$

If $X \subseteq Y$ then $\underline{R}X \subseteq \underline{R}Y$

$\underline{R}X \cup \underline{S}X \subseteq \underline{R \cap S}X$

$\underline{X} \subseteq \overline{R}X$

$\overline{R}\overline{R}X = \overline{R}X$

$\overline{R}(X \cup Y) = \overline{R}X \cup \overline{R}Y$

$\overline{R}(X \cap Y) = \overline{R}X \cap \overline{R}Y$

If $X \subseteq Y$ then $\overline{R}X \subseteq \overline{R}Y$

$\underline{R \circledast S}X \subseteq \underline{R}X \cap \underline{S}X$

In the next section we present a formal language in which we can express facts concerning definability of data discussed in the present section. The set-theoretical notions introduced here provide a basis for semantics of the language.

3 THE LANGUAGE OF LOGIC DAL

Expressions of the language of logic DAL are built from the symbols of the following pairwise disjoint sets :

VAR propositional variables

VARREL relational variables

$\neg, \wedge, \vee, \leftrightarrow, \rightarrow$ classical propositional operations of negation, conjunction, implication and equivalence, respectively.

$\circledast, \cap$ binary operations on relations

$[\,], \langle\,\rangle$ unary modal propositional operations

( ) brackets

We assume that sets VAR and VARREL are non-empty, at most denumerable sets.

Set EREL of relational expressions is the least set satisfying the following conditions :

VARREL $\subseteq$ EREL

R,S $\in$ EREL implies $R \circledast S$, $R \cap S \in$ EREL

Relational variables are intended to represent indiscernibility relations, and operations $\cap$ and $\circledast$ will be interpreted as the intersection and the transitive closure of the union of relations.

Set FOR of all formulas of the language is the least set satisfying the following conditions :

VAR $\subseteq$ FOR

A,B $\in$ FOR implies $\neg A$, $A \vee B$, $A \wedge B$,

$A \to B$, $A \leftrightarrow B \in FOR$

$A \in FOR$ and $R \in EREL$ imply $[R]A$, $<R>A \in FOR$.

Formulas are intended to represent sets of data items. In particular formulas built by using modal operations correspond to approximations of sets. Since in the language of DAL we allow compound relational expressions, we can express relationships between approximations with respect to various properties and we can explicitly describe indiscernibility relations corresponding to these properties.

4 SEMANTICS OF THE LANGUAGE OF LOGIC DAL

To define meaning of formulas of logic DAL we should fixe a set OB of data items and a family of equivalence relations in set OB corresponding to properties of these data. To be more formal, we define the notions of model and satisfiability of the formulas in a model. By a model we mean a triple : $M = (OB, \{\rho_R\}_{R \in EREL}, m)$, where OB is a non-empty set; for any $R \in EREL$ ρ_R is an equivalence relation in set OB such that $\rho_{R \cap S}$ is the greatest equivalence relation in set OB included both in ρ_R and ρ_S; $\rho_{R \cup S}$ is the least equivalence relation including both ρ_R and ρ_S ; $m : VAR \cup VARREL \to \mathcal{P}(OB) \cup \{\rho_R\}_{R \in EREL}$ is a meaning function such that :

. $m(p) \subseteq OB$ for $p \in VAR$

.. $m(R) = \rho_R$ for $R \in EREL$

Given a model M we say that a formula A is satisfied by an object o in model M(M,o sat A) iff the following conditions are satisfied :

M, o sat P iff $o \in m(p)$ for $p \in VAR$

M, o sat $\neg A$ iff not M, o sat A

M, o sat A v B iff M, o sat A or M, o sat B

M, o sat $A \wedge B$ iff M, o sat A and M, o sat B

M, o sat $A \to B$ iff M, o sat $(\neg A \vee B)$

M, o sat $A \leftrightarrow B$ iff M, o sat $(A \to A) \wedge (B \to A)$

M, o sat $[R]$ A iff for all $o' \in OB$ if $(o,o') \in \rho_R$ then M, o' sat A

M, o sat $<R>A$ iff there is an $o' \in OB$ and that $(o,o') \in \rho_R$ and M, o' sat A.

Given a model M, to each formula A of the language we assign a set of objects called an extension of A in model M($ext_M A$) :

$ext_M A = \{o \in OB : M, o \text{ sat } A\}$

The immediate consequences of this definition are the following :

Fact 4.1.

(a) $ext_M p = m(p)$ for $p \in VAR$

(b) $ext_M A = -ext_M A$

(c) $ext_M (A \vee B) = ext_M A \cup ext_M B$

(d) $ext_M (A \wedge B) = ext_M A \cap ext_M B$

(e) $ext_M (A \to B) = -ext_M A \cup ext_M B$

(f) $ext_M (A \leftrightarrow B) = (ext_M A \cap ext_M B) \cup (-ext_M A \cap -ext_M B)$

(g) $ext_M [R]A = \rho_R\ ext_M A$

(h) $ext_M <R>A = \overline{\rho_R}\ ext_M A$

Hence the classical propositional operations are interpreted as set-theoretical operations and the modal operations correspond to the operations of lower and upper approximation.

We admit the usual notions of truth and validity of formulas. A formula A is true in a model M ($\models_M A$) iff $ext_M A = OB$. A formula A is valid in logic DAL ($\models A$) iff A is true in every model for DAL. A formula A is a semantical consequence of a set Γ of formulas ($\Gamma \models A$) iff for any model M formula A is true in M whenever all formulas from Γ are true in M. A formula A is satisfiable iff M, o sat A for some model M and object o. A set Γ of formulas is satisfied in a model M by an object

o (M, o sat Γ) iff M, o sat A for all A∈Γ. A set Γ is satisfiable iff M, o sat Γ for some model M and object o.

Given a model M = (OB, $\{\rho_R\}_{R \in \mathrm{REL}}$, m), meaning function m provides a family of sets of data items which we are interested in. Next, we consider compound sets, which are expressed by means of formulas obtained from propositional variables by performing classical propositional operations. We can also express approximations of these sets with respect to relations admitted in the model. As a consequence we can discuss on a formal level definability of data in terms of properties related to these relations.

In the following we show how we can express facts concerning sets of data by means of formulas of DAL. As usually, we can express inclusion and equality of sets.

Fact 4.2.

(a) $\models_M A \to B$ iff $ext_M A \subseteq ext_M B$

(b) $\models_M A \leftrightarrow B$ iff $ext_M A = ext_M B$

definability and strong definability of sets of data can be expressed as follows.

Fact 4.3.

(a) $\models_M \langle R\rangle A \to [R]A$ iff $ext_M A$ is definable with respect to relation ρ_R

(b) $\models_M (B \to [R]A) \to (([\bar{R}]B \to A \wedge \neg A) \vee ([R]B \to [\bar{R}]A))$ for every formula B iff $ext_M [R]A$ is an equivalence class of relation ρ_R or the empty set.

The formula in condition (a) assures that the upper approximation of set $ext_M A$ is included in the lower approximation and hence by the definition of definability set $ext_M A$ is definable with respect to ρ_R. The family of formulas in condition (b) assures that the lower approximation of any set of data properly included in the lower approximation of set $ext_M A$ equals the empty set. This means that the lower approximation of $ext_M A$ consists of exactly one equivalence class of relation ρ_R. Hence the formulas from condition (a) and the family of formulas from condition (b) assure the strong definability of set $ext_M A$.

Fact 4.4.

$\models_M [R]A \to [P]A$ and $\models_M \langle P\rangle A \to \langle R\rangle A$ iff $ext_M A$ is characterised better by ρ_P than by ρ_R.

The above formulas express inclusions of approximations of set $ext_M A$ with respect to relations ρ_P and ρ_R. The lower approximation with respect to ρ_P is greater than the lower approximation with respect to ρ_R, and the upper approximation with respect to ρ_P is smaller than the upper approximation with respect to ρ_R. This means that the approximations with respect to ρ_P are closer to set $ext_M A$ than the approximations with respect to ρ_R.

Fact 4.5.

$\models_M [R]A \to [P]A$ and $\models_M \langle P\rangle A \to \langle R\rangle A$ for every formula A iff $\rho_P \subseteq \rho_R$

The condition says that inclusion of indiscernibility relations can be expressed by a family of formulas which assure that for any set of data lower approximations with respect to ρ_P are greater than the lower approximations with respect to ρ_R, and the upper approximations with respect to ρ_P are smaller than the upper approximations with respect to ρ_R.

We conclude, that in the language of logic DAL we can express various kinds of information : facts concerning sets of data items, facts concerning indiscernibility relations corresponding

to properties of data items, and relationships between data items and their properties, especially those concerning definability.

5 AXIOMATIZATION

In this section we present a deductive system for the language of DAL. We admit the following schemes of axioms and inference rules. Let R, S, and P denote arbitrary relational expressions and let A, B denote formulas.

Axioms of DAL

A1. All formulas having the form of a tautology of the classical propositional logic

A2. $[R](A \to B) \to ([R]A \to [R]B)$

A3. $[R]A \to A$

A4. $\langle R\rangle A \to [R]\langle R\rangle A$

A5. $[R \circledcirc S]A \to [R]A \wedge [S]A$

A6. $(([P]A \to [R]A) \wedge ([P]A \to [S]A)) \to ([P]A \to [R \circledcirc S]A)$

A7. $[R]A \vee [S]A \to [R \cap S]A$

A8. $(([R]A \to [P]A) \wedge ([S]A \to [P]A)) \to ([R \cap S]A \to [P]A)$

Rules of inference

$$\frac{A,\ A \to B}{B}$$ modus ponens

$$\frac{A}{[R]A}$$ necessitation

For a fixed relation R axioms A1, ..., A4 and the rules of inference correspond to the axiomatization of modal logic S5. Axioms A5, A6 provide the definition of operation $\circledcirc$ and axioms A7, A8 give the definition of operation $\cap$.

In the usual way we define the notions of proof and theorem. A proof of a formula A from a set Γ of formulas is a finite sequence of formulas each of which is either an axiom or an element of set Γ or else is obtainable form earlier formulas by a rule of inference. A formula A is derivable from a set Γ ($\Gamma \vdash A$) iff it has a proof form set Γ. A formula A is a theorem of DAL ($\vdash A$) iff it is derivable merely from axioms. A set Γ of formulas is consistent if the formula of the form $A \wedge \neg A$ is not derivable from Γ.

It is easy to see that the axioms are valid and the rules preserve validity. Hence the following theorem holds.

Fact 5.1. (Soundness theorem)

(a) $\vdash A$ implies $\models A$

(b) $\Gamma \vdash A$ implies $\Gamma \models A$

(c) Γ satisfiable implies Γ consistent.

In the following we list some theorems of logic DAL.

Fact 5.2.

(a) $\vdash [R \circledcirc R]A \leftrightarrow [R]A$

(b) $\vdash [R \cap R]A \leftrightarrow [R]A$

(c) $\vdash [(R \circledcirc S) \circledcirc P]A \leftrightarrow [R \circledcirc (S \circledcirc P)]A$

(d) $\vdash [(R \cap S) \cap P]A \leftrightarrow [R \cap (S \cap P)]A$

(e) $\vdash [(R \circledcirc S) \cap P]A \to [(R \cap P) \circledcirc (S \cap P)]A$

(f) $\vdash (R \circledcirc P) \cap (S \circledcirc P)A \to [(R \cap S) \circledcirc P]A$

(g) $\vdash [R \circledcirc S]A \to [R][S]A$

(h) $\vdash [R][S]A \to [R \cap S]A$

Fact 5.3. (Completeness theorem)

(a) $\models A$ implies $\vdash A$

(b) $\Gamma \models A$ implies $\Gamma \vdash A$

(c) Γ consistent implies Γ satisfiable.

6 CONCLUSIONS

In this paper we have given some ideas how to analyse data using logic tools. We followed the classical framework of logic programming, namely we defined a logic language for expressing data analysis problems and we developed a deductive system for the language to use proof procedures to obtain solutions to these problems.

Data analysis has been understood

as a process of obtaining patterns in a set of data items. We considered two main tasks involved in data analysis :

- to agregate data into sets according to their properties
- to define properties adequate for characterisation of sets of data.

We defined formal counterparts of sets of data and properties. Namely, sets of data are defined by means of formulas of the language of logic DAL and properties are defined by means of relational expressions of the language. We presented the deductive system for logic DAL. We have given a particular interest to the notion of strong definability of sets of data which enables us to establish properties which adequately characterize these data.

The results of the present paper can be extended to languages with the other operations on relations e.g. with the composition of relations, or to relations which are not necessarily equivalence relations e.g. tolerance relations (reflexive and symmetric). We expect that a mechanical proof procedure can be defined for logic DAL by u-

Konrad, E., Orlowska, E., Pawlak, Z. Knowledge representation systems. ICSPAS Report 433, 1981.

Mirkowska, G. PAL - Propositional algorithmic logic, logic of programs. Lecture Notes in Computer Science, 31-101, Springer-Verlag, 1981.

Orlowska, E. Representation of vague information. ICSPAS Reports 503, 1983.

Orlowska, E., Pawlak, Z. Expressive power of knowledge representation systems. ICSPAS Reports 432, accepted for publication in the Journal of Man Machine Studies, 1981.

Pawlak, Z. Information systems-theoretical foundations. Information Systems 6, 205-218, 1981.

Pawlak, Z. Rough sets. International Journal of Computer and Information Sciences, 11, 341-356, 1982.

Pawlak, Z. Rough classification. ICSPAS Reports 506, accepted for publication in the Journal of Man-Machine Studies, 1983.