PRACE IPI PAN • ICS PAS REPORTS

Zdzisław Pawlak

Toward the theory of information systems

I. The notion of an information system

419
JUNE 1980
WARSZAWA

INSTYTUT PODSTAW INFORMATYKI POLSKIEJ AKADEMII NAUK
INSTITUTE OF COMPUTER SCIENCE POLISH ACADEMY OF SCIENCES
00-901 WARSAW, P.O. Box 22, POLAND

Zdzisław Pawlak

TOWARD THE THEORY OF INFORMATION SYSTEMS

I. The notion of an information system

419

Warsaw, July 1980

Pracę zgłosił Andrzej Blikle

Mailing address: Prof. dr Z. Pawlak
Institute of Computer Science PAS
P.O. Box 22
00-901 Warszawa PKiN

P r i n t e d a s a m a n u s c r i p t
N a p r a w a c h r ę k o p i s u

Nakład 700 egz. Ark. wyd. 1,35; ark. druk. 2,25.
Papier offset. kl. III, 70 g, 70 x 100. Oddano do
druku w lipcu 1980 r. W. D. N. Zam. nr 522/80 0-123

Abstract • Содержание • Streszczenie

In this paper the mathematical model of information system is investigated. The notion of an attribute is a starting point of our investigations. The dependency of attributes is discussed and some basic properties of attributes are stated. Then the notion of a subsystem is defined and some operations on information systems are discussed.

О теории систем информатики. Понятие информационной системы

В работе исследуется понятие математической модели информационной системы. Как исходное понятие для дальнейших рассуждений используется понятие атрибута. Затем исследуется понятие зависимости атрибутов и их другие свойства. В заключении вводится понятие подсистемы информационной системы, определяются некоторые операции на информационных системах и рассматриваются свойства этих операций.

O teorii systemów informatycznych

I. Pojęcie systemu informacyjnego

W pracy bada się pojęcie matematycznego modelu systemu informacyjnego. Punktem wyjścia do rozważań jest pojęcie atrybutu. Następnie bada się pojęcie zależności atrybutów oraz inne ich własności. Na koniec wprowadzone jest pojęcie podsystemu informacyjnego oraz określone i badane są pewne działania na systemach informacyjnych.

INTRODUCTION

This paper, first in the series, reports part of the activities of the Information Storage and Retrieval Systems Group at Warsaw.

We proposed and investigated in this Group certain mathematical model for attribute based information retrieval systems.

This model was first published by Pawlak [13] and extended by Marek and Pawlak [8]. In this report we use somewhat new formulation of the discussed model after Pawlak [14] and state some new problems.

The idea of an information system investigated in this report is slightly related to that of Salton [15] and Wang, Chiang [19], however there is an essential difference between them. In our approach in contrary to [15] and [19] the information language is formally introduced and intensively investigated, giving a powerful tool very useful in information retrieval system design and offering deeper insight and understanding of phenomena involved in information processing.

The proposed model of information system, have been implemented (see Marganski [9]) in an information retrieval system for agriculture library. The implementation shows many superior features of this method to other known methods of information retrieval, (for example inverted files method).

Further theoretical and practical investigations in this field are under way (see for example Jaegerman [3], Konikowska, Traczyk [4], Lipski [5], Lipski, Marek [6], E. Orłowska [10], [11], M. Orłowska [12], Traczyk [16], Wakulicz [18]).

1. INFORMATION SYSTEMS

In this paragraph we give the basic notions of the paper, which will be discussed in details in the rest of the paper.

The main notion is that of information system. The basic component of an information system is a finite set of objects X, for example human beings, books, etc. The objects are classified by means of a finite set A of attributes. With every attribute $a \in A$, there is associated a non-empty set V_a of values of an attribute a; V_a will be also referred to as domain of attribute a. For instance if a is "sex" then $V_a = \{male, female\}$, if a is color, then $V_a = \{red, green, blue\}$. Naturally some attributes can share the set of values, for example domain of attribute "length" and "height" is the same and it is the set of nonnegative reals.

In order to "describe" some properties of objects we introduce a function ϱ from $X \times A$ into $V (V = \bigcup_{a \in A} V_a)$, such that $\varrho(x,a) \in V_a$ for every $x \in X$ and $a \in A$.

This is to mean that by means of the function ϱ we associate with each object its description - a set of attribute values.

Now we can give formal definition of an information system (see Pawlak [14]).

By an information system we shall mean a 4-tuple

$$S = \langle X, A, V, \varrho \rangle ,$$

where

X - is a finite set of objects,

A - is a finite set of attributes,

$V = \bigcup_{a \in A} V_a$, where V_a is the set of values of attribute a, and $\mathrm{card}(V_a) \geq 2$.

ϱ - is a function from $X \times A$ into V.

We define the domain of attribute a as

$$V_a = \{ v \in V : \underset{x \in X}{E} \varrho(x,a) = v \}.$$

Example 1

Let us consider very simple information system defined as follows:

$$X = \{x_1, x_2, x_3, x_4, x_5\},$$
$$A = \{sex, salary, age\},$$
$$V = \{V_{sex} \cup V_{sal} \cup V_{age}\},$$

where

$$V_{sex} = \{male, female\},$$
$$V_{sal} = \{low, medium, high\},$$
$$V_{age} = \{young, medium, old\}.$$

The salary "low" is less than $ 6000 a year; "medium" - between $ 6000 and $ 24,000; "high" - more than $ 24,000 a year.

The age "young" is to mean less than 21; "medium" - between 21 - 40, "old" more than 40.

The function ϱ in our example is defined by the following table:

X	SEX	SALARY	AGE
x_1	male	low	young
x_2	male	high	medium
x_3	female	low	young
x_4	male	medium	old
x_5	female	low	medium

We shall also use the notion of a descriptor of an attribute a.

By a descriptor we shall mean any element of the set $\{a\} \times V_a$. That is to mean that descriptors are pairs of the form (a,v), where $v \in V_a$. For instance in example 1 the following are descriptors: (AGE, young), (SALARY, low), (SEX, male).

For every $x \in X$ we define the function φ_x from A into V such that $\varphi_x(a) = \varphi(x,a)$. We shall call this function information (or data) about x in S.

For instance in example 1 information about x_2 is the following function:

$\varphi_{x_2} =$

SEX	SALARY	AGE
male	high	middle

In other words information about x in S is simply a set of descriptors corresponding to all attributes in the system. Thus we may write information about x_2 in the form:

$\{$(SEX,male), (SALARY, high), (AGE, middle)$\}$.

Let us notice that our information about objects is exhaustive and exclusive, i.e., values of each attribute exhaust all possibilities, and only one attribute value can be associated with each object.

Because we deal in this paper only with finite systems, that is systems having finite number of objects, finite number of attributes and finite domains of attributes we may identify the notion of an information systems with the finite table defining the function φ. The columns of the table, labelled with attributes, are composed of values of corresponding attributes and rows of the table, labelled with objects, are informations about corresponding objects. Of course we admit occurence of the seme rows in the table. Naturally the order of columns and rows in the table is insignificant.

2. PROPERTIES OF INFORMATION SYSTEMS

In the paragraph we shall give some more details about information systems which will give better insight in the considered notion.

Any function φ from A into V such that for every a, $\varphi(a) \in V_a$ will be called information in S. The set of all informations in S will be denoted by $Inf(S)$. There are evidently

$$\prod_{a \in A} card(V_a)$$

informations (different) in the system S.

For instance in the example given in the previous paragraph we have

$$card(V_{sex}) \cdot card(V_{sal}) \cdot card(V_{age}) = 2 \cdot 3 \cdot 3 \cdot = 18 \text{ informations.}$$

For every $\varphi \in Inf(S)$, we define $X_\varphi = \{x \in X: \varphi_x = \varphi\}$. We can interpret $X\varphi$ as a set of all objects $x \in X$ whose information in S is identical with φ. This is to mean that objects belonging to the set $X\varphi$ are undistinguishable in the system S.

An information φ is said to be empty iff $X_\varphi = \emptyset$. Otherwise it is said to be nonempty.

An information φ is said to be selective iff $card(X_\varphi) = 1$. System S is said to be selective iff every nonempty information in S is selective. A system S is said to be complete iff every information in S is non-empty.

Example 2

Let $S = \langle X, A, V, \varphi \rangle$ be an information system defined by the table

X	a	b	c
x_1	p_1	q_2	r_1
x_2	p_2	q_3	r_2
x_3	p_1	q_2	r_1
x_4	p_1	q_1	r_3

The function φ such that $\varphi(a) = p_1$, $\varphi(b) = q_2$, $\varphi(c) = r_1$ is an information in S and $X_\varphi = \{x_1, x_3\}$, because

$$X_\varphi = \{x \in X: \varrho_x = \varphi\} =$$
$$= \{x \in X: \bigwedge_{a \in A} \varrho_x(a) = \varphi(a)\} =$$
$$= \bigcap_{a \in A} \{x \in X: \varrho(x,a) = \varphi(a)\} =$$
$$= \{x \in X: \varrho(x,a) = p_1\} \cap$$
$$\cap \{x \in X: \varrho(x,b) = q_2\} \cap$$
$$\cap \{x \in X: \varrho(x,c) = r_1\} =$$
$$= \{x_1, x_3, x_4\} \cap \{x_1, x_3\} \cap \{x_1, x_3\} =$$
$$= \{x_1, x_3\}.$$

So the system is neither selective nor complete, because $\mathrm{card}(X_p) = 2$ and there are empty informations in the system, for example $\varphi'(a) = p_1$, $\varphi'(b) = q_1$, $\varphi'(c) = r_1$.

Let $S = \langle X, A, V, \varrho \rangle$ be an information system. We define two binary relations $\widetilde{a}$ ($a \in A$), and $\widetilde{S}$ on S in the following way:

$$x \mathrel{\widetilde{a}} y \text{ iff } \varrho(x,a) = \varrho(y,a),$$
$$x \mathrel{\widetilde{S}} y \text{ iff } \varrho_x = \varrho_y.$$

Two objects are in the relation $\widetilde{a}$ iff they are undistinguishable with respect to the attribute a; and similarly, two objects are in the relation $\widetilde{S}$ if they have the same information in S. (i.e., they are undistinguishable with respect to every attribute $a \in A$).

In the recent example $x_1 \mathrel{\widetilde{a}} x_4$ (x_1, x_4 are undistinguishable with respect to the attribute a because $\varrho_{x_1}(a) = \varrho_{x_4}(a)$ and objects x_1, x_3 are undistinguishable with respect to every attribute in A; i.e., $x_1 \mathrel{\widetilde{S}} x_3$, because $\varrho_{x_1} = \varrho_{x_3}$.

It is easy to check that:

For every information system $S = \langle X, A, V, \varrho \rangle$, $\widetilde{a}$, $\widetilde{S}$, are equivalence relations on X and.

$$\widetilde{S} = \bigcap_{a \in A} \widetilde{a}.$$

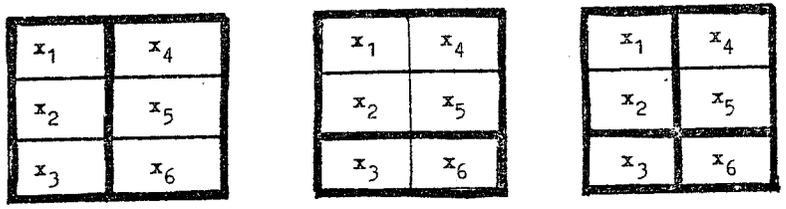
The equivalence classes of the relation $\widetilde{S}$ will be called <u>elementary (atomic) sets</u> in S or when X is fixed, elementary (atomic) sets. The family of all elementary sets in S will be denoted by E_S.

<u>Example 3</u>

Let $S = \langle X, A, V, \varrho \rangle$ be an information system defined as follows

X	a	b
x_1	p_1	q_1
x_2	p_1	q_1
x_3	p_1	q_2
x_4	p_2	q_1
x_5	p_2	q_1
x_6	p_2	q_2

The partitions generated by the system are depicted below



$\tilde{a}$ $\tilde{b}$ $\tilde{a} \cap \tilde{c}$

Thus the partition $\tilde{a}$ consists of two equivalence classes

$$\{x_1, x_2, x_3\}, \{x_4, x_5, x_6\};$$

Partition $\tilde{b}$ gives also two equivalence classes

$$\{x_1, x_2, x_4, x_5\}, \quad \{x_3, x_6\},$$

and the product partition $\tilde{S} = \tilde{a} \cap \tilde{b}$ consists of four elementary sets

$$\{x_1, x_2\}, \{x_4, x_5\}, \{x_3\}, \{x_6\}.$$

That is to say if we classified objects of a given set by means of all attributes and their values (descriptors) we automatically introduce a partition of the set of all objects. In each equivalence class (elementary set) of this partition there are objects which are undistinguishable in the system. In general each elementary set contains more than one element. (The system is not selective). That is to mean the "description power" of a chosen set of attributes and its values is not strong enough to describe every single member of the set X.

Let us observe that if φ, ψ are different informations in the system S, then

$$X_\varphi \cap X_\psi = \emptyset,$$
$$\bigcup X_\varphi = X,$$
$$\varphi \in \mathrm{Inf}(S)$$

and if φ is not empty information than X_φ is an elementary set in S. In other words informations generate a partition of the set X, and this is exactly the partition generated by the relation $\tilde{S}$.

Thus with every elementary set in S we can associate exactly one information in S, and conversely, with every information in S we can associate exactly one elementary set in S (possibly an empty set).

Let $S = \langle X, A, V, \varrho \rangle$ be an information system. We shall define a new information system $S^* = \langle E_S, A, V, \varrho^* \rangle$, called the <u>representation</u> of the system S, where

$$\varrho^* : E_S \times A \to V$$

and

$$\varrho^*(e,a) = v, \quad e \in E_S, \quad a \in A$$

if and only if

$$\varrho(x,a) = v$$

for all $x \in e$.

In other words if we remove all duplicate rows in the table S and replace objects by elementary sets containing this objects in the table S, so we obtain representation of the system S.

For example if the system S is given by the table

X	a	b	c
x_1	u_1	v_1	w_2
x_2	u_2	v_3	w_1
x_3	u_2	v_2	w_3
x_4	u_1	v_1	w_2
x_5	u_1	v_1	w_2
x_6	u_2	v_2	w_3

then the representation of S is the system.

X	a	b	c
$\{x_1,x_4,x_5\}$	u_1	v_1	w_2
$\{x_2\}$	u_2	v_3	w_1
$\{x_3,x_6\}$	u_2	v_2	w_3

Thus representation of any system is selective, i.e. each row in the representation occurs only once.

Finally let us remark that the set of a nonempty informations in S defines a relation $R_S \subset V_{a_1} \times V_{a_2} \times \dots \times V_{a_k}$, such that $R_S(v_{i_1},x_{i_2},\dots,v_{i_k})$ iff $(a_1,v_{i_1}),(a_2,v_{i_2}),\dots,(a_k,v_{i_k})$ is non empty information in S.
This corresponds to relstional model of information system introduced by Codd [20] however we shall not discuss this model in this paper.

3. DEPENDENCY OF ATTRIBUTES

Often value of some attribute can be derived from values of another attribute.

For example if the value of an attribute AGE is "two years", then the value of the attribute EDUCATION will be "no education", if both attributes are concerning the same person. The problem of dependency of attributes has been intensively studied in relational model (see Aho, Beeri, Ullman [1]), but we shall define it in somewhat different way.

The formal definition of this relation is the following one.

Let $a,b \in A$ be two attributes in an information system $S = \langle X,A,V,\mathcal{S} \rangle$.

a) Attribute b is said to be <u>dependent</u> on a ($a \to b$) iff $\tilde{a} \subset \tilde{b}$,

b) Attributes a,b are called <u>independent</u> iff neither $\tilde{a} \subset \tilde{b}$ nor $\tilde{a} \supset \tilde{b}$;

c) Attributes a,b are said to be <u>equivalent</u> ($a \sim b$) iff $\tilde{a} = \tilde{b}$.

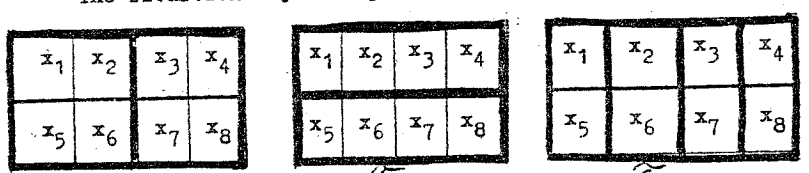
<u>Example 4</u>

Let $S = \langle X,A,V,\mathcal{S} \rangle$ be an information system defined by the table

X	a	b	c
x_1	p_1	q_1	r_1
x_2	p_1	q_1	r_2
x_3	p_2	q_1	r_3
x_4	p_2	q_1	r_4
x_5	p_1	q_2	r_1
x_6	p_1	q_2	r_2
x_7	p_2	q_2	r_3
x_8	p_2	q_2	r_4

It is easy to see that $c \to a$, but a,b and c,b are pairwise independent.

The situation may be depicted as shown below



Similarly we introduce the relations $B \to a$, $a \to B$, $B \to C$, where B,C are subsets of A.

Attribute a is said to be dependent on the set B of attributes, $B \subset A$, iff $\widetilde{B} \subset \widetilde{a}$ and $a \to B$ iff $\widetilde{a} \subset \widetilde{B}$.
In general we may write $B \to C$ iff $\widetilde{B} \subset \widetilde{C}$.

Sets of attributes B,C are equivalent $(B \sim C)$ iff $\widetilde{B} = \widetilde{C}$.

Let us notice that $B \to C$ iff $B \to c_1$ and $B \to c_2$ and $B \to c_k$ and if $b_1 \to C$ or $b_2 \to C$ or $b_1 \to C$ then $B \to C$, where

$B = \{b_1, b_2, \ldots, b_1\}$ and $C = \{c_1, c_2, \ldots, c_k\}$.

The meaning of the "dependency" relation $B \to C$ is obvious. It simply means that values of the left-hand side attributes determine values of the right-hand side attributes.

That is to say, if $B \to C$, then there exists one function f (dependency function)

$$f: \underset{b \in B}{P} V_b \to \underset{c \in C}{P} V_c$$

such that

$$(\varrho(x,c))_{c \in C} = f(\varrho(x,b)_{b \in B}), \text{ for all } x \in X.$$

(P denotes cartesian product).
In other words there exists one set of functions $(f_c)_{c \in C}$ such that

$$(\underset{c \in C}{\forall})(\varrho(x,c) = f_c(\varrho(x,b)_{b \in B}),$$

and

$$\varrho(x,c) = f_c(\varrho(x,b)_{b \in B})$$

iff

$$X_{c,\varrho(x,c)} \supset X_{b_1,\varrho(x,b_1)} \cap X_{b_2,\varrho(x,b_2)} \cap \ldots \cap X_{b_k,\varrho(x,b_k)}$$

for all $x \in X$, where $X_{c,v} = \{x \in X : \varrho_x(c) = v\}$.

Example 5

Let $S = \langle X, A, V, \varrho \rangle$ be an information system, such that

$X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9\}$,

$A = \{a, b, c\}$,

$V_a = \{p_1, p_2, p_3, p_4\}$,

$V_b = \{q_1, q_2, q_3\}$,

$V_c = \{r_1, r_2, r_3\}$.

Assume that the attributes generate the following partition S on X:

$X_{a,p_1} = \{x_1, x_2, x_4, x_5\}$,

$X_{a,p_2} = \{x_3, x_6\}$,

$X_{a,p_3} = \{x_7, x_8\}$,

$X_{a,p_4} = \{x_9\}$,

$X_{b,q_1} = \{x_1, x_4, x_7\}$,

$X_{b,q_2} = \{x_2, x_5, x_8\}$,

$X_{b,q_3} = \{x_3, x_6, x_9\}$,

$X_{c,r_1} = \{x_1, x_2, x_3\}$,

$X_{c,r_2} = \{x_4, x_5, x_6\}$,

$X_{c,r_3} = \{x_7, x_8, x_9\}$.

The partition is shown below

x_1	x_2	x_3
x_4	x_5	x_6
x_7	x_8	x_9

$\tilde{a}$

x_1	x_2	x_3
x_4	x_5	x_6
x_7	x_8	x_9

$\tilde{b}$

x_1	x_2	x_3
x_4	x_5	x_6
x_7	x_8	x_9

$\tilde{c}$

In this system all attributes are pairwise independent but $\{b,c\} \Rightarrow a$, because $\tilde{b} \cap \tilde{c} \subset \tilde{a}$.

The corresponding "dependency" function f is given by the table below:

V_b	V_c	V_a
q_1	r_1	p_1
q_1	r_2	p_1
q_1	r_3	p_3
q_2	r_1	p_1
q_2	r_2	p_1
q_2	r_3	p_3
q_3	r_1	p_2
q_3	r_2	p_2
q_3	r_3	p_4

Thus for instance $f(q_2,r_3) = p_3$ and $f(q_3,r_2) = p_2$.

Thus knowing values of attributes b and c we may compute by means of dependency function f value of the attribute a.

It is obvious that if $B \Rightarrow C$ in S then also $B \Rightarrow C$ in S^*. So instead of checking whether or not $B \Rightarrow C$ in S we check the dependency in S^*, which is much simpler, because the table of S^* is much simpler than the table of S.

The question arises whether the dependency $B \Rightarrow C$ could be deduced from some other known dependencies in S by means of logical inference rules and not by checking the table of S or S^*. Similar problem has been investigated in relational model of database (see for example Aho, Beeri, Ullman [1]), but we shall assume here another solution (see Orłowska [11]).

4. REDUCED SYSTEMS

As we have stated in the previous paragraph some attributes in the information system may be superfluous in this sense that their values can be "derived" from the values of other attributes in the system. We shall consider this question in this paragraph in some details.

Let us first introduce basic definitions.

A subset $B \subseteq A$ is said to be <u>independent in S</u> iff, for every $B' \not\subseteq B$, $\widetilde{B} \neq \widetilde{B'}$.

A subset $B \subseteq A$ is said to be <u>dependent</u> in S iff there exists a $B' \subseteq B$ such that $\widetilde{B'} = \widetilde{B}$.

The set B is said to be <u>derivable from C</u> in S iff $B, C \subset A$, $C \subset B$ and $\widetilde{B} = \widetilde{C}$.

One can easily verify the following properties:

a) If $B \subset A$ is the greatest independent in S then for every $a \in A$, $B \to a$.

b) If B is dependent in S then there exists $B' \subset B$ independent in S such that for every $a \in B - B'$, $B \to a$.

c) If $B \subset A$ then $A \to B$.

<u>Example 6</u>

Let $S = \langle X, A, V, \mathcal{S} \rangle$ be an information system such that $X = \{x_1, x_2, x_3, x_4, x_5\}$ and $A = \{a, b, c, d\}$.

Assume that the attributes generate the following partitions on X:

$\widetilde{a} = \{x_1, x_2, x_5\}, \{x_3, x_4\}$,
$\widetilde{b} = \{x_1\}, \{x_2, x_3, x_4, x_5\}$,
$\widetilde{c} = \{x_1, x_2, x_3, x_4\}, \{x_5\}$,
$\widetilde{d} = \{x_1\}, \{x_3, x_4\}, \{x_2, x_5\}$.

Of course, the partition generated by attribute a is the set of all equivalence classes of the relation $\widetilde{a}$ i.e., the partition on X defined by the relation $\widetilde{a}$.

It is easy to see that the whole set of attributes A determine the partition $\{x_1\}, \{x_2\}, \{x_3, x_4\}, \{x_5\}$. Now, we have the following relationship between the attributes: $d \Rightarrow b$ and $d \Rightarrow a$ (because $\widetilde{d} \subset \widetilde{b}$ and $\widetilde{d} \subset \widetilde{a}$.

Also $\{a, b, c\} \Rightarrow d$ because $\widetilde{a} \cap \widetilde{b} \cap \widetilde{c} \subset \widetilde{d}$; and $\{c, d\} \to \{a, b\}$, because $\{c, d\} \Rightarrow a$, and $\{c, d\} \Rightarrow b$, i.e. $\widetilde{c} \cap \widetilde{d} \subset \widetilde{a}$ and $\widetilde{c} \cap \widetilde{d} \subset \widetilde{b}$.

The set A is dependent in S because there exists $B \subset A$, $B = \{a, b, c\}$ such that $\widetilde{B} = \widetilde{A}$. There are also other subsets $C, D \subset A$ of attributes $C = \{a, c, d\}$, $D = \{c, d\}$, with the same property, e.g., $\widetilde{C} = \widetilde{D} = \widetilde{A}$.

Sets B and D are independent in S whereas C is not because $\widetilde{C} = \widetilde{D}$.

As we have seen from previous consideration some attributes may be sometimes eliminated from the system, and their values one can derive from the remaining set of attributes. This is to mean that they are superfluous in the system. This leads to the following definition.

Let $S = \langle X, A, V, \mathcal{S} \rangle$ be an information system. A set $A' \subset A$ will be called a <u>reduct of A</u> iff $\widetilde{A} = \widetilde{A'}$, and there does not exist a proper subset B of A' such that $\widetilde{B} = \widetilde{A'}$. The corresponding system $S' = \langle X, A', V, \mathcal{S}' \rangle$ is called <u>reduced system</u>. ($\mathcal{S}'$ is the restriction of the function $\mathcal{S}$ to the set $X \times A'$).

It is clear that a system can turn out to have more than one reduct. In example 6, we have two reducts of A, namely B and D.

It is easy to prove the following properties:

a) If an information system is complete then it is also reduced (the converse implication does not hold (see example 7).

b) If an information system is reduced then all its different attributes are pairwise independent. (The converse implication does not hold see example 8).

c) Two information systems S, S' with the same set of objects X, are said to be equivalent iff $\widetilde{S} = \widetilde{S}'$.

For every information system S there exists a reduced system S' equivalent to S.

Let us also notice, that if S is reduced, then also S is reduced.

Example 7

Let $S = \langle X, A, V, \varrho \rangle$ an information system

X	a	b
x_1	p_1	q_1
x_2	p_2	q_1
x_3	p_1	q_2

The system is reduced but is not complete since for a,b such that $\varphi(a) = p_2$ and $\varphi(b) = p_2$, $X\varphi = \emptyset$.

Example 8

Let $S = \langle X, A, V, \varrho \rangle$, where $X = \{x_1, x_2, x_3, x_4\}$, $A = \{a, b, c\}$ and the attributes determine the following partition of the set X:

$\widetilde{a} = \{x_1, x_2\}, \{x_3, x_4\}$,
$\widetilde{b} = \{x_1\}, \{x_2, x_3, x_4\}$,
$\widetilde{c} = \{x_2\}, \{x_1, x_3, x_4\}$.

The attributes a,b,c are pairwise independent, while $\{a,b\}$, $\{a,c\}$ and $\{b,c\}$ are reducts of A.

The idea of reduction of an attribute set in a system is of great practical importance, because it shows that one can get sometimes the same information from the system with smaller set of attributes. This may have special meaning in the case when attributes are symptoms of some illness but in order to get the proper diagnosis it is not necessary to investigate all symptoms, but try to find only those which are really necessary. In fact there can be more than one set of such minimal symptoms (see example 6).

The problem arises how to find effectively reducts of a given information system. Because all sets in the system are finite such an algorithm always exists, however it may be not very efficient in general.

Some considerations concerning this subject one can find in Łoś [7], Truszczyński [17], Grzymała-Busse [2].

5. SUBSYSTEMS

In this section we shall introduce and discuss the notion of subsystem of a given information system.

Let $S = \langle X, A, V, \varrho \rangle$ and $S' = \langle X', A', V', \varrho' \rangle$ be two information systems. We say that S' is a subsystem of S if $X' \subset X$, $A' \subset A$, $V' \subset V$ and $\varrho' = \varrho\, X' \times A'$.

If S' is subsystem of S, then we shall write $S' < S$ or $S' \underset{X',A'}{<} S$, or $S' = S/_{X',A'}$.

In other words if we remove from the table S some columns or rows then the remaining table is the subsystem of the system S.

For example if in the system

X	a	b	c
x_1	v_1	u_2	w_1
x_2	v_2	u_1	w_2
x_3	v_1	u_2	w_2
x_4	v_1	u_2	w_1

we dropp the column b and the row x_3 then we obtain a subsystem of S

X	a	c
x_1	v_1	w_1
x_2	v_2	w_2
x_4	v_1	w_1

We shall introduce two kinds of subsystems.

If $S' < S$ and $X' = X$, then we shall say that S' is an attribute restricted subsystem of S, in symbols $S' \underset{A'}{<} S$ or $S' = S/_{A'}$.

If $S' < S$ and $A' = A$, then we shall say that S' is an object restricted subsystem of S, in symbols $S' \underset{X'}{<} S$ or $S' = S/_{X'}$.

Thus if S is an information system and we dropp some column from it, then the obtained system is an attribute restricted subsystem of S, and if we remove some rows from the system S - we obtain object restricted subsystem of S.

For example if in the system

X	a	b	c
x_1	v_1	u_2	w_1
x_2	v_2	u_1	w_2
x_3	v_1	u_2	w_2
x_4	v_1	u_1	w_1

we remove column b the obtained attribute restricted subsystem of S

X	a	c
x_1	v_1	w_1
x_2	v_2	w_2
x_3	v_1	w_2
x_4	v_1	w_1

and if we remove from S the row x_3 we obtain object restricted subsystem of S

X	a	b	c
x_1	v_1	u_1	w_1
x_2	v_2	u_1	w_2
x_4	v_1	u_2	w_1

Now we shall give some elementary properties of subsystems.

If $S' = S/X'$ then

$$\widetilde{S'} = \widetilde{S} \cap (X')^2$$

If $S' = S/A'$, then $\widetilde{S'} \supset S$.

If $S' < S$ and S is reduced then S' is also reduced.

If $S' = S/X'$ and S is complete then S' may be not complete.

If $S' = S/A'$ and S is complete then S' is complete.

If $S' = S/A'$ and S is selective then S' may not be selective.

If $S' = S/X'$ and S is selective then S' is selective.

If $S' = S/X'$ then

$$(S')^* = S^*/X'.$$

If $S' = S/A'$, then

$$(S')^* \neq S^*/A'.$$

If $S' = S/X',A'$ then there exist exactly one system $S_1 = S/X'$ and $S_2 = S/A'$ such that $S' = S_1/A' = S_2/X'$.

6. CONNECTION OF INFORMATION SYSTEMS

Very often we face the following problem. We are given some information systems $S_1,S_2,\ldots,S_k$ and we want to have one "common" information system S combining all systems $S_1,S_2,\ldots,S_k$ into one. The system S will be called connection of systems S_i, $i = 1,2,\ldots,k$, and will be denoted as $S = \bigcup_{i=1}^{k} S_i$.

Let $S = \langle X,A,V,\varrho\rangle$ and $S_i = \langle X_i,A_i,V_i,\varrho_i\rangle$, $i = 1,\ldots,k$.

The system S is a connection of system S_i if the following conditions are satisfied:

$$X = \bigcup_{i=1}^{k} X_i ,$$

$$A = \bigcup_{i=1}^{k} A_i ,$$

$$V = \bigcup_{i=1}^{k} V_i ,$$

$$\varrho_{/X_i \times A_i} = \varrho_i , \quad i = 1,\ldots,k,$$

$$\varrho_x = \bigcup_{i=1}^{k} \varrho_{i_x} , \quad x \in X.$$

Connection $S = \bigcup_{i=1}^{k} S_i$ is well defined if the following two conditions are valid:

1) If $(X_i \cap X_j) \neq \emptyset$ and $(A_i \cap A_j) \neq \emptyset$ then

$$\varrho_{i/(X_i \cap X_j)\times(A_i \cap A_j)} = \varrho_{j/(X_i \cap X_j)\times(A_i \cap A_j)} ,$$

for all $i,j = 1,\ldots,k$, and

2) $\varrho_x = \bigcup_{i=1}^{k} \varrho_{i_x}$ is defined for all $x \in X$ and $a \in A$.

Of course systems S_i are subsystems of S.

The first condition is obvious and the second needs some explanation.

Let S_1 be a system with only one attribute, say color, and S_2 a system also with one attribute, for example, length, and assume that $X_1 \cap X_2 = \emptyset$. The second condition says that we are not allowed to define connection S of S_1 and S_2 because we do not have any information about lengths of objects in S_1 or about colors of objects in S_2. Thus we are unable to define for all $x \in X$ the information about color and length of x. In other words, we are not able to define the function ϱ_x for the connection $S = S_1 \cup S_2$.

This seems to have natural justification in real life systems. If we have two information systems, say first concerning insurance and the second medical care with different

sets of population, for example, one in London and the second in Warsaw, then combining those two systems into one connected system is justified only in the case when we have insurance data in the medical system and conversely. Otherwise we are unable to define for all $x \in X$ the information f_x about insurance and medical care and, consequently, according to our definition, the connection of these two systems is not an information system.

Let us consider very simple formal example depicting above situation more clearly.

The connection of the two following systems

X	a	b	c
x_1	v_1	u_1	w_2
x_2	v_1	u_2	w_1
x_3	v_2	u_1	w_2
x_4	v_2	u_1	w_2

Y	c	d	e
x_3	w_2	p_1	q_1
x_4	w_2	p_2	q_1
y_1	w_1	p_3	q_1
y_2	w_2	p_1	q_2

is the table

X ∪ Y	a	b	c	d	e
x_1	v_1	u_1	w_2	-	-
x_2	v_1	u_2	w_1	-	-
x_3	v_2	u_1	w_2	p_1	q_1
x_4	v_2	u_1	w_2	p_2	q_1
y_1	-	-	w_1	p_3	q_1
y_2	-	-	w_2	p_1	q_2

which is not an information system according to our definition because some values of attributes are undefined in the table. That is to say function defined by the table is not total but partial, which is not allowed in our definition of an information system.

This property leads to a definition of two special kinds of connections of information systems.

If $S = \bigcup S_i$ and $S_i = S/A_i$ then S will be called attribute connected system.

If $S = \bigcup S_i$ and $S_i = S/X_i$, then S will be called object connected system.

This two kinds of connections are depicted by the following example.

Example 3

Let S_1, S_2 be two information systems with the some set of objects and different sets of attributes as shown below:

X	a	b	c
x_1	u_1	v_1	w_2
x_2	u_1	v_2	w_1
x_3	u_2	v_1	w_2
x_4	u_1	v_1	w_2

X	a	d	e
x_1	u_1	p_1	q_2
x_2	u_1	p_2	q_1
x_3	u_2	p_1	q_1
x_4	u_1	p_1	q_2

Connection of S_1 and S_2 is given below

X	a	b	c	d	e
x_1	u_1	v_1	w_2	p_1	q_2
x_2	u_1	v_2	w_1	p_2	q_1
x_3	u_2	v_1	w_2	p_1	q_1
x_4	u_1	v_1	w_2	p_1	q_2

Let S_3, S_4 be two information systems with different sets of objects but the some set of attributes as shown in the tables

X	a	b	c
x_1	u_1	v_1	w_2
x_2	u_2	v_2	w_1
x_3	u_1	v_2	w_1
x_4	u_1	v_1	w_1

Y	a	b	c
x_3	u_1	v_2	w_1
x_4	u_1	v_1	w_1
y_1	u_2	v_2	w_2
y_2	u_1	v_2	w_1

Connection of S_3 and S_4 is the system

X∪Y	a	b	c
x_1	u_1	v_1	w_2
x_2	u_2	v_2	w_1
x_3	u_1	v_2	w_1
x_4	u_1	v_1	w_1
y_1	u_2	v_2	w_2
y_2	u_1	v_2	w_1

Attribute connected system corresponds to the situation when all constituent information systems have the same set of objects but different set of attributes. For example if we have in some town different information systems owned by insurance company, medical care service, bank office, police etc. then we may combine them into one information system. The set of objects in those systems are the same (all inhabitant of the town) but the set of attributes in all systems are different.

Object connected information system refers to the situation when all constituent systems have the same set of attributes but different sets of objects. For example if the same company, say insurance company, own information systems in different districts. Thus we have the case when the set of attributes in each system is the same but the objects (inhabitants of the districts) are different. So we can consider all these systems as an attribute connected system.

Now we shall give some elementary properties of the "connection" operation.

Let $S = \langle X, A, V, \varrho \rangle$ and $S_i = \langle X_i, A_i, V_i, \varrho_i \rangle$, $i=1,\ldots,k$ be information systems and let $S = \bigcup_{i=1}^{k} S_i$.

If $S = \bigcup S_i$, $S_i = S/A_i$ and each S_i is reduced, then S may not be reduced.

If $S = \bigcup S_i$, $S_i = S/X_i$ and each S_i is reduced, then S is also reduced.

If $S_k = \bigcup S_i$, $S_i = S/A_i$, then $\widetilde{S} = \bigcap_{i=1}^{k} \widetilde{S}_i$.

If $S = \bigcup S_i$, $S_i = S/X_i$, then $X_\varphi = \bigcup_{i=1}^{k} X_{i,\varphi}$ for all $\varphi \in Inf(S)$.

If $S = \bigcup S_i$, $S_i = S/A_i$ and each S_i is reduced then S may not be reduced.

If $S = \bigcup S_i$, $S_i = S/X_i$, and each S_i is reduced, then S is also reduced.

If $S = \bigcup S_i$, $S_i = S/X_i$, then $S^* \neq \bigcup S_i^*$,

If $S = \bigcup S_i$, $S_i = S/A_i$, then $S^{**} = S_i^{**}$.

If $S = \bigcup S_i$, $S_i = S/A_i$ and each S_i is selective, then S is also selective.

If $S = \bigcup S_i$, $S_i = S/X_i$ and each S_i is selective then S may not be selective.

If $S = \bigcup S_i$ and each S_i is complete then S is also complete.

There are systems S_i such that

$$(\bigcup S_i)^{**} \neq \bigcup (S_i)^{**}.$$

Received July 15, 1980

REFERENCES

1. Aho A.V., Beeri, Ullman J.D., The theory of Joins in Relational Databases, ACM Transactions on Database Systems, vol. 4, No 3 (1979), 297-314.
2. Grzymała-Busse J., manuscript, 1978.
3. Jaegerman M., Information Storage and Retrieval Systems with Incomplete Information, Part I - Fundamenta Informaticae, II.1 (1978), 17-41, Part II, II.2 (1979), 141-166.
4. Konikowska B., Traczyk T., A Query Language of Stochastic Information Systems, Fundamenta Informaticae, II.3 (1978), 351-363.
5. Lipski W., On Semantic Issues Connected with Incomplete Information Databases, ACM Transaction on Database Systems, vol. 4, No 3 (1979), 262-298.
6. Lipski W., Marek W., On Queries Involving Cardinalities, Information Systems. Vol. 4, No 3 (1979), 241-246.
7. Łoś J., Characteristic Sets of a System of Equivalence Relations, Colloquium Mathematicum, XLII (1979), 291-293.
8. Marek W., Pawlak Z., Information Storage and Retrieval Systems: Mathematical Foundations, Theoretical Computer Science No 1(1976), 331-354.
9. Margański E., Implementation of Retrieval System by the Method of Atomic Constituents (in Polish), PWN (1979).
10. Orłowska E., Dynamic Information Systems (to appear).
11. Orłowska E., On Dependency of Attributes in Information Systems, (to appear).
12. Orłowska M., Algebraical and Topological Properties of Database Systems with Incomplete Information. (in Polish) PWN (1980).

13. Pawlak Z., Mathematical Foundation of Information Retrieval, CC PAS Reports, No 101, (1973).

14. Pawlak Z., Information Systems, ICS PAS Reports, No 338 (1978).

15. Salton G., Automatic Information Organization and Retrieval, McGrow Hill Book Company, New York (1968).

16. Traczyk T., Common Extension of Boolean Information Systems, Fundamenta Informaticae, II.1 (1978), 63-70.

17. Truszczyński M., Algorithmic Aspects of the Minimization of the Set of Attribute Problem, Fundamenta Informaticae, Vol. 4, No 4 (to appear).

18. Wakulicz - Deja A., Time Varing Information Systems (to appear).

19. Wong E., Chiang T.C., Canonical Structure in Attribute Based File Organization, Comm. ACM., 14 (1971), 593-597.

20. Codd E. F., A Relational Model of Data for Large Shared Data Banks, Comm. ACM, Vol. 13 (1970), 377-387.

CONTENTS