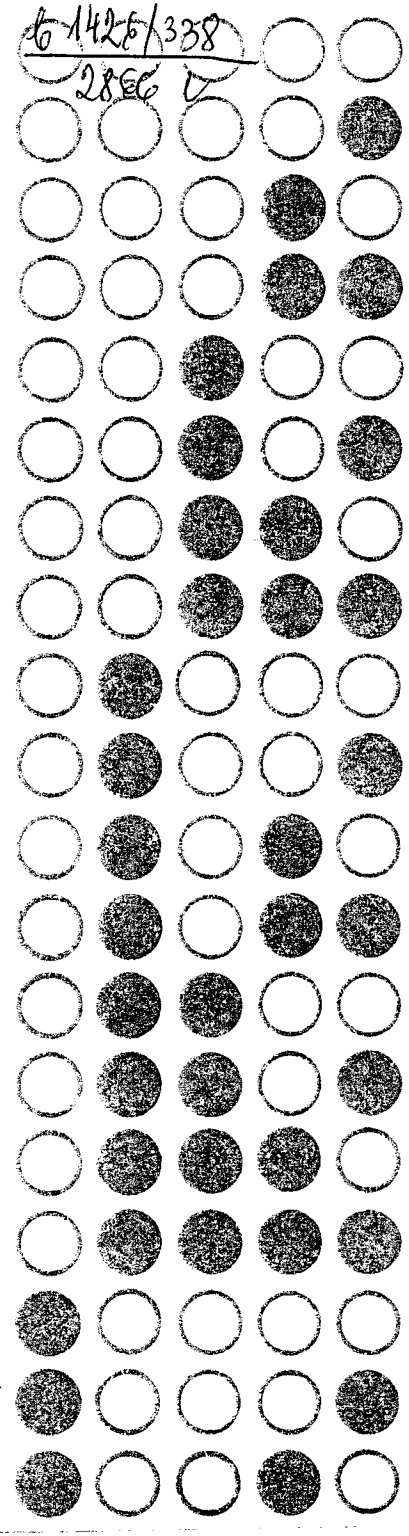

PRACE IPI PAN • ICS PAS REPORTS

Zdzisław Pawlak

Information systems

338

1978

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

INFORMATION SYSTEMS

338

Warsaw 1978

Mailing address: Prof. dr Zdzisław Pawlak
Institute of Computer Science
Polish Academy of Sciences
P.C. Box 22
00-901 Warsaw
Poland

Printed as a manuscript
Na prawach rękopisu

Nakład 700 egz. Ark. wyd. 0,5; ark. druk. 1,00. Papier offset. kl. III, 70 g, 70 x 100. Oddano do druku we wrześniu 1978 r. W. D. N. Zam. nr 618/78
S-03

Abstract . Содержание . Streszczenie

This is new simplified version of ideas given in: W. Marek, Z. Pawlak: Information Storage and Retrieval Systems. Mathematical Foundations, Theoretical Computer Science 2(1978) pp. 331-354, and in Z. Pawlak: Mathematical Foundations of Information Retrieval, CC PAS Reports No 101(1973), concerning information systems.

Presented approach offers more adequate tools for dealing with problems related to information systems. In particular, the notion of independence of attributes is defined, as well as minimization problem for attributes is formulated.

Информационные системы

Приводится новая версия формализации понятия информационной системы, рассматриваемой в работах З.Марка и З.Павляка "Information Storage and Retrieval Systems. Mathematical Foundations". Theoretical Computer Scieces 2 /1978/, pp 331 - 354, и З.Павляка "Mathematical Foundation of Information Retrieval". CC PAS Reports No 101 /1973/, которая дает возможность введения простой формулировки проблем, связанных с данной областью. В частности уточняется понятие зависимости признаков и минимизации множества признаков.

Systemy informacyjne

Podano nową wersję formalizacji pojęcia systemu informacyjnego rozpatrywanego w pracach W. Marka i Z. Pawlaka "Information Storage and Retrieval Systems. Mathematical Foundations". Theorical Computer Sciences 2 (1978) pp. 331-354 i Z. Pawlaka "Mathematical Foundation of Information Retrieval". CC PAS Reports No 101 (1973), która pozwala na prostsze niż poprzednio formułowanie problemów z tego zakresu. W szczególności sprecyzowano pojęcie zależności atrybutów oraz minimalizacji zbioru atrybutów.

This is somewhat modified version of ideas given in [1] and [2].

1. INFORMATION SYSTEM

Definition. An information system is a 4-tuple

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

where:

X - is a set of objects,

A - is a set of attributes,

Q - is a set of descriptors,

ϱ - is an information function from $X \times A$ into

Let $Q_a \subset Q$, $a \in A$, denote subset of Q defined as follows:

$$Q_a = \{ q \in Q : \bigvee_x \varrho(x,a) = q \}$$

Thus Q_a is the set of all descriptors assigned to the attribute a. We shall also refer to the elements of Q_a as values of a.

The following notation will be used in the paper:

a) $Inf(S) = Q^A$ - the set of all informations in S

b) For x in X, ϱ_x is to mean a function from A into such that $Q_x(a) = q$ if and only if $\varrho(x,a) = q$

We refer to ς_x as the information about x in S

c) For $\varphi \in Inf(S)$, $X_\varphi = \{x \in X : \varsigma_x = \varphi\}$.

If $\overline{\overline{X}}_\varphi = \phi$, φ is called an empty information, otherwise the information is non empty.

If $\overline{\overline{X}}_\varphi = 1$, then φ is called selective information.

If all informations in S are selective, then S is referred to as an selective system.

If all informations in S are not empty, then S is called complete.

Definition. Let $S = \langle X, A, Q, \varsigma \rangle$ be an information system. Then $\tilde{a}\ (a \in A)$ is a binary relation defined by $x \tilde{a} y$ iff $\varsigma(x,a) = \varsigma(y,a)$. And let $\tilde{S}$ be a binary relation on X defined as:

$$x \tilde{S} y \text{ iff } \varsigma_x = \varsigma_y .$$

Theorem $\tilde{a}$, $\tilde{S}$ are equivalence relations and

$$\tilde{S} = \bigodot_{a_i \in A} \overset{\sim}{a_i},$$

where $\bigodot$ is a product of partitions (equivalence relations) $\overset{\sim}{a_i}$ defined in a usual way.*

$\tilde{a}$ for each $a \in A$ is an equivalence relation, because

1° $x \tilde{a} x$, for $\varsigma_x(a) = \varsigma_x(a)$

2° $x \tilde{a} y \Rightarrow y \tilde{a} x$, because

$\varsigma_x(a) = \varsigma_y(a) \Rightarrow \varsigma_y(a) = \varsigma_x(a)$

---

* Of course every subset $A' \subset A$ generates an equivalent relation

$$\overset{\sim}{A'} = \bigodot_{a_i \in A'} \overset{\sim}{a_i}$$

In particular $\tilde{S} = \tilde{A}$.

3° If $x \tilde{a} y$ and $y \tilde{a} z$, then $x \tilde{a} z$, because if

$$\varsigma_x(a) = \varsigma_y(a) \quad \text{and}$$

$$\varsigma_y(a) = \varsigma_z(a)$$

then

$$\varsigma_x(a) = \varsigma_z(a)$$

Similarly, $\tilde{S}$ is an equivalence relation;

1° $x \tilde{S} x$, because $\varsigma_x = \varsigma_x$

2° if $x \tilde{S} y$, then $y \tilde{S} x$, because

$\varsigma_x = \varsigma_y \Rightarrow \varsigma_y = \varsigma_x$

3° if $x \tilde{S} y$ and $y \tilde{S} z$, then $x \tilde{S} z$, because

$\varsigma_x = \varsigma_y$ and $\varsigma_y = \varsigma_z$ imples, $\varsigma_x = \varsigma_z$.

Finally, if

$$x \tilde{S} y, \text{ then } \varsigma_x = \varsigma_y$$

And this implies

$$\bigwedge_{a \in A} \varsigma_x(a) = \varsigma_y(a)$$

hence

$$\varsigma_x(a_1) = \varsigma_y(a_1)$$

$$\varsigma_x(a_2) = \varsigma_y(a_2)$$

$$\cdots\cdots\cdots$$

$$\varsigma_x(a_n) = \varsigma_y(a_n) \Rightarrow$$

$$\varsigma(x \overset{\sim}{a_1} y) \wedge \varsigma(x \overset{\sim}{a_2} y) \wedge \cdots \wedge \varsigma(x \overset{\sim}{a_n} y) \Rightarrow$$

$$\bigodot_{a_i \in A} x \overset{\sim}{a_i} y ,$$

The converse implication is similar.

2. ATTRIBUTES

Definition. An information system S' is said to be included in a system S ($S' \subset S$) iff $\tilde{S}' \subset \tilde{S}$.

Definition. Two information systems S', S are said to be equivalent iff $\tilde{\varsigma'} = \tilde{\varsigma}$.

Let S', S be two information systems.

If $X = X'$, $Q' \subset Q$, $A' \subset A$, and

$$\varsigma' = \varsigma /_{A' \times X}$$

then we shall say that S' is weaker than S, or S is stronger than S' ($S' \prec S$). Of course if $S' \prec S$ then $S' \supset S$.

If S_1, S_2 are information systems then $S' = S_1 \cup S_2$ will be called join of S_1 and S_2 - and defined in the following manner:

$$X = X_1 \cup X_2,$$

$$A = A_1 \cup A_2,$$

$$Q = Q_1 \cup Q_2,$$

$$\varsigma = \varsigma_1 \cup \varsigma_2,$$

and

$$\varsigma /_{X_1 \times A_2} = \varsigma_1$$

$$\varsigma /_{X_2 \times X_2} = \varsigma_2$$

moreover

$$\varsigma_1 /_{(X_1 \cap X_2) \times (A_1 \cap A_2)} = \varsigma_2 /_{(X_1 \cap X_2) \times (A_1 \cap A_2)}.$$

Every information system $S = \langle X, A, Q, \rangle$ may be represented as join of one attribute systems $S = \bigcup_{1 \leq i \leq k} S_i$, where $S_i =$

$$= \langle X, a_i, Q_{a_i}, \varsigma_i \rangle, \quad a_i \in A = \{a_1, \ldots, a_k\}.$$

Let S_1, S_2 be information systems such that

$$X_1 = X_2, \quad A_1 = A_2, \quad Q_1 \neq Q_2.$$

If $\tilde{a}^1 \subset \tilde{a}^2$ for all $a \in A(A')$ we shall call S_1 finer than S_2, or S_2 cruder than S_1 ($S_1 \sqsubset S_2$).

If $S_1 \sqsubset S_2$, then $\tilde{S}_1 \subset \tilde{S}_2$.

($\tilde{a}^1$, $\tilde{a}^2$ are equivalence relations defined by attribute a systems S_1, S_2 respectively).

Definition. Let $a, b \in A$ be two attributes in an information system S. Attribute a is said to be included in b ($a \rightarrow b$) (or b is dependent on a) iff $\tilde{a} \subset \tilde{b}$; a and b are equivalent ($a \sim b$) iff $\tilde{a} = \tilde{b}$.

If $a \rightarrow b$ then there exist a function

$$f : Q_a \longrightarrow Q_b;$$

which means that the value at attribute a uniquely defines the value of attribute b.

Definition. The set A of attributes in S is called independent in S iff for every $A' \subset A$ $\tilde{A} \supset \tilde{A}$; if there is a subset $A' \subset A$ such that $\tilde{A}' = \tilde{A}$, then the set of attributes A will be called dependent in S.

Definition. The smallest set $A' \subset A$ such that A' is independent in $S = \langle X, A, Q, \varsigma \rangle$ will be called reduct of A, and the corresponding system, $S' = \langle X, A', Q, \varsigma' \rangle$ - reduced system (ς' is the restriction of the function ς to the set A').

Theorem For every information system S, there exists an equivalent reduced information system S'.

Theorem If an information system S is complete, then S is reduced (the converse implication is not true).

Theorem If S is a reduced information system, then all its different attributes are pairwise independent (converse is not true).

Problem. Given an information system S.
Find an algorithm transfering S to a reduced form.

This problem may be also stated as follows:

Let X be a set and $\Pi = \{\pi_1, \ldots, \pi_n\}$ a finite family of equivalence relations on X (partitions of X) and let $\tilde{\Pi}$ be defined as

$$\tilde{\Pi} = \pi_i \in \Pi \textcircled{\cdot} \tilde{\pi}_i$$

$\textcircled{\cdot}$ - is a product of partitions.
Find the smallest subset $\Pi' \subset \Pi$ such that $\tilde{\Pi}' = \tilde{\Pi}$.

3. THE LANGUAGE OF S

With the system S we associate the language $\mathcal{L}_S$.
Let us first define syntax of $\mathcal{L}_S$.
An alphabet of $\mathcal{L}_S$ (denoted Σ_S) is the following:

1° $Q \subset \Sigma_S$, $A \subset \Sigma_S$

2° $0,1 \in \Sigma_S$

3° $T,F \in \Sigma_S$

4° $\sim, \cdot, + \in \Sigma_S$

5° $\neg, \wedge, \vee \in \sum_S$

6° $= \in \sum_S$

Terms $\mathcal{T}_S$ are defined as follows

1° $0,1 \in \mathcal{T}_S$

2° $A \times Q \subseteq \mathcal{T}_S$

3° if $t, t' \in \mathcal{T}_S$, then
$t, t \cdot t', t + t' \in \mathcal{T}_S$

4° Nothing else is in $\mathcal{T}_S$.

Formulae $\mathcal{F}_S$ are the following expressions

1° $T,F \in \mathcal{F}_S$

2° if $t,t' \in \mathcal{T}_S$, then $t = t' \in \mathcal{F}_S$

3° if $\psi, \psi' \in \mathcal{F}_S$, then
$\neg\psi, \psi \vee \psi', \psi \wedge \psi' \in \mathcal{F}_S$

4° Nothing else is in $\mathcal{F}_S$.

Semantics of the language will consist of two parts: semantics of terms and semantics of formulae.

Semantics of terms is the function

$$\sigma_{\mathcal{T}}: \mathcal{T}_S \to \mathcal{P}(X),$$

where

1° $\sigma_{\mathcal{J}}(0) = \emptyset$, $\sigma_{\mathcal{J}}(1) = X$

2° $\sigma_{\mathcal{J}}(a,q) = \{x \in X : x(a) = q\}$

3° $\sigma_{\mathcal{J}}(\sim t) = X \setminus \sigma_{\mathcal{J}}(t)$

$\sigma_{\mathcal{J}}(t \cdot t') = \sigma_{\mathcal{J}}(t) \cap \sigma_{\mathcal{J}}(t')$

and

$$\sigma_{\mathcal{J}}(t + t') = \sigma_{\mathcal{J}}(t) \cap \sigma_{\mathcal{J}}(t').$$

Semantics of formulae is the following function

$$\sigma_{\mathcal{F}} : \mathcal{F} \longrightarrow \{T,F\},$$

where

1° $\sigma_{\mathcal{F}}(T) = T$, $\sigma_{\mathcal{F}}(F) = F$

2° $(T=t') =$

2° $\sigma_{\mathcal{F}}(t=t') = \begin{cases} T & \text{if } \sigma_{\mathcal{J}}(t) = \sigma_{\mathcal{J}}(t') \\ F & \text{if } \sigma_{\mathcal{J}}(t) \neq \sigma_{\mathcal{J}}(t') \end{cases}$

3° $\sigma_{\mathcal{F}}(\neg\Psi) = \begin{cases} T & \text{if } \sigma_{\mathcal{F}}(\Psi) = F \\ F & \text{if } \sigma_{\mathcal{F}}(\Psi) = T \end{cases}$

4° $\sigma_{\mathcal{F}}(\Psi \vee \Psi') = \sigma_{\mathcal{F}}(\Psi) \vee \sigma_{\mathcal{F}}(\Psi')$

5° $\sigma_{\mathcal{F}}(\Psi \wedge \Psi') = \sigma_{\mathcal{F}}(\Psi) \wedge \sigma_{\mathcal{F}}(\Psi')$

Definition . A term t is primitive if t is of the form

$$(a_1,q_1) \cdot (a_2,q_2) \cdot \ldots \cdot (a_n,q_n),$$

where $A = \{a_1,\ldots,a_n\}$ and $q_1,\ldots,q_n \in Q$.

Definition A term t is in normal form if t is of the form

$$t_1 + t_2 + \ldots + t_r,$$

with $r \geq 1$ and $t_1,\ldots,t_r$ primitive terms.

Definition Let $Y \subseteq X$. We say that Y is atomic, if there exists a primitive term t such that

$$\sigma_{\mathcal{J}}(t) = Y$$

(Note that non empty atomic sets are exactly equivalence classes of the relation $\tilde{S}$.)

Definition. Let $Y \subseteq X$. We say that Y is describable if there exists a term t such that

$$\sigma_{\mathcal{J}}(t) = Y.$$

We will use $\mathcal{T}_{prim}$ to denote the set of all primitive terms and $\mathcal{T}_{nor}$ to denote the set of all normal terms.

Theorem. Let $t,t' \in \mathcal{T}_{prim}$ be two different primitive terms. Then

1° $\sigma_{\mathcal{J}}(t) \cap \sigma_{\mathcal{J}}(t') = \emptyset$

2° $\bigcup_{t \in \mathcal{T}_{prim}} \sigma_{\mathcal{J}}(t) = X$

Theorem. There exists an algorithm which for an arbitrary term t constructs a normal term t' such that

$$\sigma_J(t) = \sigma_J(t')$$

Corollary. Let $Y \subseteq X$. Then Y is describable if and only if Y is a set theoretical union of atomic sets.

References

[1] W. Marek, Z. Pawlak: Information Storage and Retrieval Systems. Mathematical Foundations. Theoretical Computer Sciences 2 (1976), pp. 331-354.

[2] Z. Pawlak: Mathematical Foundation of Information Retrieval. CC PAS Reports, No 101 (1973).