PRACE CO PAN • CC PAS REPORTS

Wiktor Marek, Zdzisław Pawlak

Mathematical foundations of information storage and retrieval

Part 2

136

1973

WARSZAWA

CENTRUM OBLICZENIOWE POLSKIEJ AKADEMII NAUK
COMPUTATION CENTRE POLISH ACADEMY OF SCIENCES
WARSAW PKIN, P. O. Box 22, POLAND

Wiktor Marek, Zdzisław Pawlak

MATHEMATICAL FOUNDATIONS OF INFORMATION STORAGE AND RETRIEVAL

Part 2

136

Warszawa 1973

Mailing address: Dr. Wiktor Marek
Institute of Mathematics PAS
ul. Śniadeckich 8
00-950 Warszawa
P.O. Box 187

Prof. Dr. Zdzisław Pawlak
Computation Centre PAS
00-901 Warszawa PKiN
P.O. Box 22

P r i n t e d   a s   a   m a n u s c r i p t

Nakład 450 egz. Ark. wyd. 0,5; ark. druk. 0,750
Papier offset. kl. III, 70 g, 70 x 100. Oddano do
druku w październiku 1973 r. W. D. N. Zam. nr 756
R-30

Abstract • Содержание • Streszczenie

In the part I we introduced a syntax and semantics for the theory of information storage and retrieval. Here we present a development of this theory.

Математическое описание процесса поиска
и хранения информации. Вторая часть

В первой части был введен синтаксис и модели хранения и поиска информации. Сейчас приводится расширение введенной теории.

Matematyczne podstawy gromadzenia
i wyszukiwania informacji. Część 2

W części pierwszej wprowadziliśmy składnię i modele teorii przechowywania i wyszukiwania informacji. Przedstawiamy rozwinięcie wprowadzonej teorii.

§ 1. DESCRIBABLE SETS

Definition 1.1. Let $S = \langle X, A, I, U \rangle$ be a system of i.s.r. A set $Y \subseteq X$ is called describable within S iff there is $t \in T$ such that $\|t\|_S = Y$.

Lemma 1.1. Describable subsets of X form boolean algebra.

Proof: It follows by the choice of axioms for terms in our system.

Lemma 1.2. If S is finite selective system then every subset $Y \subseteq X$ is describable.

Proof: Assume t_x is a term describing $\{x\}$ (such a term exists by selectiveness of S). Form $\sum_{x \in A} t_x$.

Then $\|\sum_{x \in A} t_x\| = \bigcup_{x \in A} \{x\} = A$.

Remark. In this point there is a difference between finite and infinite i.s.r. systems. Indeed assuming the language $\mathcal{L}_A$ finitary (i.e. allowing only for finite conjunctions and disjunctions) with A infinite it is easy to produce infinite selective system with nondescribable subset (using cardinality argument). Since the fact that every subset is describable implies selectiveness we get - by lemma 1.2.

Theorem 1. If S is finite i.s.r. system then S is selective iff every subset of X is describable within S.

§ 2. OPERATIONS ON I.S.R. SYSTEMS

Definition 2.1. Let $S = \langle X,A,I,U\rangle$ be an i.s.r. system.
Let $\{I_j\}_{j\in I}$ be a partition of the set I.
An induced family of systems $\{S_j\}_{j\in J}$ is formed as follows:

$S_j = \langle X,A^j,I_j,U_j\rangle$ where

(a) $A^j = \bigcup_{i\in I_j} A_i$

(b) $U_j = U \restriction A^j$

In the same way the family of languages $\{\mathcal{L}_j\}$ is induced clearly $\mathcal{L}_j$ corresponds to S_j.

Definition 2.2. Let $\{S_j\}_{j\in J}$ be a family of i.s.r. systems ($S_j = \langle X,A^j,I_j,U_j\rangle$) and moreover $i \neq j \Longrightarrow A^i \cap A^j = \emptyset = I_i \cap I_j$.
We define:

$\bigoplus_{j\in J} S_j = \langle X,A,I,U\rangle$ where $A = \bigcup_{j\in J} A^j$, $I = \bigcup_{j\in J} I_j$, $U = \bigcup_{j\in J} U_j$.

Note that $I' \subseteq I$ induces partition $I = I' \cup (I - I')$.
And thus we naturally obtain restriction of S to I'.

Definition 2.3. Let $S_i = \langle X_i,A^i,I_i,U_i\rangle$ $i = 0,1$ be two i.s.r. systems. We say that $S_0 \leq S_1$ iff

(a) $X_0 \subseteq X_1$

(b) $A^0 \subseteq A^1$

(c) $I_0 \subseteq I_1$

(d) $\bigwedge_{a\in A_0} U_1(a) \cap X_0 = U_0(a)$

(e) $\bigwedge_{i\in I_0} A^0_i = A^1_i$

Lemma 2.1. Assume $S = \langle X,A,I,U\rangle$ is i.s.r. system, $\{I_j\}_{j\in J}$ is a partition of I and $\{S_j\}_{j\in J}$ is induced family. Then for each $j \in J$, $S_j \leq S$.

Proof: The conditions (a), (b), (c) and (e) are obvious.
Since the carrier of S_j is X therefore our condition (d) takes form $\bigwedge_{a\in A^j} U_j(a) = U(a)$ which is condition (b) of 2.1.
This shows adequacy of definitions 2.1. and 2.3.

Lemma 2.2. Under obvious assumptions $S_j \leq \bigoplus_{j\in J} S_j$

Theorem 2. Assume $S_0 \leq S_1$ and let t be a term of the language $\mathcal{L}_{A^0}$. Then $\|t\|_{S_0} = \|t\|_{S_1} \cap X_0$.

Proof: By induction on the complexity of t.
If t is an atomic term i.e. t is c_a then the condition 2.3.(d) gives the result. If t is T or F it is equally obvious.
Assume now $t = \neg t_1$ then

$$\|t\|_{S_0} = \|\neg t_1\|_{S_0} = X_0 - \|t_1\|_{S_0} = X_0 - (\|t_1\|_{S_1} \cap X_0)$$

(here inductive assumption is used)

$$X_0 - (\|t_1\|_{S_1} \cap X_0) = (X_1 \cap X_0) - (\|t_1\|_{S_1} \cap X_0) = (X_1 - \|t_1\|_{S_1}) \cap X_0 =$$

$$= \|\neg t_1\|_{S_1} \cap X_0 = \|t\|_{S_1} \cap X_0.$$

Assume $t = t_1 \cdot t_2$ then

$$\|t\|_{S_0} = \|t_1\|_{S_0} \cap \|t_2\|_{S_0} = \|t_1\|_{S_1} \cap X_0 \cap \|t_2\|_{S_1} \cap X_0 =$$

$$\|t_1\|_{S_1} \cap \|t_2\|_{S_1} \cap X_0 = \|t\| \cap X_0.$$

The case $t = t_1 + t_2$ is similar. Implication is eliminated in obvious way.

Definition 2.4. (a) $S_0 \underset{X}{\subseteq} S_1$ iff $S_0 \subseteq S_1$ and $X_0 = X_1$

(b) $S_0 \underset{A}{\subseteq} S_1$ iff $S_0 \subseteq S_1$ and $A^0 = A^1$

Lemma 2.3. If $S_0 \underset{A}{\subseteq} S_1$ then $I_0 = I_1$

Proof: by 2.3.(e) $A_i^0 = A_i^1$ for $i \in I_0$. If $I_1 - I_0 \neq \emptyset$ then by conditions on I and A ($\bigcup_{i\in I_0} A_c = A^0$, $\bigcup_{i\in I_1} A_i = A^1$).

Thus $\bigcup_{i\in I_1} A_i^1 = A \cup (\bigcup_{i\in I_1 - I_0} A_i^1) = A$. Thus $\bigcup_{i\in I_1 - I_0} A_i^1 \subseteq A = \bigcup_{i\in I_0} A_i^1$

which contradicts the fact that I_1 is a partition.

Theorem 3. If $S_0 \subseteq S_1$ then there are i.s.r. systems S_2 and S_3 such that

$$\begin{array}{ccccc} & \underset{A}{\subseteq} & S_2 & \underset{X}{\subseteq} & \\ S_0 & & & & S_1 \\ & \underset{X}{\subseteq} & S_3 & \underset{A}{\subseteq} & \end{array}$$

Proof: We define S_2 as follows: $X_2 = X_1$, $A_2 = A_0$, $U_2 = U_1 \upharpoonright A_0$.
Similarily S_3 is defined as follows $X_3 = X_0$ $A_3 = A_1$
$U_3(a) = U_1(a) \cap X_0$.
We leave to the reader checking of the details.

Theorem 4: (a) If S is i.s.r. system and $Y \subseteq X$ then there is S' such that $S \underset{X}{\subseteq} S'$ and Y is describable within S'.
(b) If S is finite i.s.r. system, $\mathcal{A}$ is a boolean algebra of describable sets (within S) and $\mathcal{B}$ is any boolean algebra of subsets of X such that $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{P}(X)$ then there is S' such

that $S \underset{A}{\subseteq} S'$ and $\mathcal{B}$ is exactly boolean algebra of describable subsets of S'.

Proof: (a) If Y is describable within S then put $S' = S$. Assume Y is not describable within S. Add new element $i' \notin I$ and two another elements a', a'' both not in $A \cup I$. Form $A' = A \cup \{a', a''\}$, $I' = I \cup \{i'\}$, $A_{i'} = \{a', a''\}$, $U'(a') = Y$, $U'(a'') = X - Y$.
Clearly Y is describable in S'.
(b) Follows from (a) and following observation:
If $Y \in \mathcal{B}$ and $\mathcal{A} \subseteq \mathcal{B}$ ($\mathcal{A}, \mathcal{B}$ boolean algebras of sets) then the smallest boolean algebra $[\mathcal{A}, Y]$ containing $\mathcal{A}$ and Y is included in $\mathcal{B}$.

§ 3. IMPLEMENTATION RESTRICTIONS

Definition 3.1. A family $\mathcal{A} \subseteq \mathcal{P}(X)$ is called conus iff

$$(A)_{\mathcal{A}} (B) (B \subseteq A \longrightarrow B \in \mathcal{A})$$

Definition 3.2. A sufficient information in selective i.s.r. system S is a conus $\mathcal{A}$ containing all singeltons.

Definition 3.3. A term t is conforming s.i. $\mathcal{A}$ iff $\|t\| \in \mathcal{A}$

Lemma 3.1. The set of terms conforming s.i. $\mathcal{A}$ forms a subfamily of T closed under •.

Definition 3.4. If $\mathcal{A}$ is a s.i. for S we define

$$t_1 \sim_{\mathcal{A}} t_2 \longleftrightarrow (\|t_1\|_S = \|t_2\|_S \in \mathcal{A}) \vee (\|t_1\|_S \notin \mathcal{A} \wedge \|t_2\|_S \notin \mathcal{A})$$

Lemma 3.2. $\sim_{\mathcal{A}}$ is an equivalence.

Note that in practical applications relation $\sim_{\alpha}$ plays important role.

Definition 3.5. Every term $t_1 \leq t$ such that $\|t_1\| \in \mathfrak{A}$ is called sufficient extension of t for $\mathfrak{A}$.

Lemma 3.3. The set of sufficient extensions of t for $\mathfrak{A}$ is closed under $\bullet$. However it needs not to be closed under $+$.
In practical situations we consider systems with numeration.

Definition 3.6. Let $\langle T, \leq \rangle$ be an ordered set. If $S = \langle X, A, I, U \rangle$ is an i.s.r. system and $\varphi : X \xrightarrow{1-1} T$ then φ is called enumeration on T.
Clearly φ induces an i.s.r. on $\varphi * X$, isomorphic to S.

Definition 3.7. Term t is called segmential in ordered i.s.r. system $\langle S, \varphi \rangle$ iff $\varphi * (\|t\|_S)$ is a segment in $\langle T, \leq \rangle$.
The segment are particularily convenient in the process of retrieval. Thus we may wish to have certain terms in segmential form.

Lemma 3.4. The family of segmential terms in $\langle S, \varphi \rangle$ is closed under $\bullet$.
Let $\mathfrak{M}$ be a family of terms such that $(t_1, t_2)_{\mathfrak{M}} (t_1 \neq t_2 \longrightarrow \|t_1\|_S \cap \|t_2\|_S = \emptyset)$ then we have

Lemma 3.5. There is well ordered set $\langle T, \leq \rangle$ and enumeration $\varphi : X \xrightarrow[\text{onto}]{1-1} T$ such that each term $t \in \mathfrak{M}$ is segmential.
Moreover we may order that fixed term $t \in \mathfrak{M}$ generates an initial segment of T.
Proof of 3.5. being straightforward we omit here.

The problem which families of terms may be segmentalized seems to us to be of big importance. We do not know any sufficient and necessary condition. Yet we give here certain sufficient condition.

Let $\mathfrak{M}$ be a family of terms, S an i.s.r. system.
$\mathfrak{M}$ is said to satisfy condition C with respect to S iff decomposes into two subfamilies $\mathfrak{M}'$ and $\mathfrak{M}''$ such that
(a) Every two different terms in $\mathfrak{M}'$ have disjoint values (in S)
(b) There is a decomposition $\mathfrak{X}$ of $\mathfrak{M}'$ such that for every class W of $\mathfrak{X}$ there is at most one term $t \in \mathfrak{M}''$ such that $\|t\|_S \subseteq \|\sum_{t \in W} t\|_S$ moreover
If W is a class of the decomposition $\mathfrak{X}$ (as before) then there are at most two terms in W which values (in S) are not included in that of t.
We have:

Theorem 5: If $\mathfrak{M}$ satisfies condition C with respect to S then there is ordering $\langle T \leq \rangle$ and $\varphi : X \xrightarrow[\text{onto}]{1-1} T$ such that all terms in t are segmential in $\langle S, \varphi \rangle$.
Proof of this theorem will be published elsewhere.

In the further work we shall present the hierarchical approach within our framework.

Mathematical Institute of Polish Academy of Sciences
Computing Center of Polish Academy of Sciences

Received August 30, 1973

Recently the following papers have appired in this series:

Nr 112 - B. Idzik: Farkas lemma for concave-convex functions with an application to the nonlinear von Neumann model, 1973

Nr 113 - R. Naczke, H. Pytel: Maszyny wieloprogramowane, 1973

Nr 114 - K. Wierzcholski: Wyznaczanie rozwiązań równań różniczkowych n-tego rzędu na EMC, 1973

Nr 115 - L. Czaja: Heurystyczne lokowanie macierzy rzadkich, 1973

Nr 116 - J. Małuszyński: Subproblems of parsing problems, 1973

Nr 117 - A. Idzik: On Markow policies in continuous time discounted dynamic programming, 1973

Nr 118 - J. Ejsmund: Rozkład grafu programu na bloki, 1973

Nr 119 - A. Blikle: An algebraic approach to mathematical theory of programs, 1973

Nr 120 - W. Żakowski: Wielowymiarowe maszyny ciągłe, 1973

Nr 121 - F. Jarosińska, B. Rykaczewska-Kamińska: Język makrogeneracyjny MOL. Zastosowania i implementacja, 1973

Nr 122 - M. Iglewski, R. Krzemień: Analiza leksykalna, 1973

Nr 123 - A. Trybulec: Rozpoznawanie stanów pamięci przez procesor, 1973

Nr 124 - R. Naczke, H. Pytel: A multi-counter machine, 1973

Nr 125 - J. Winkowski: Processes in composed systems, 1973

Nr 126 - T. Skrzypkowski: O zawieraniu maszyn relacyjnych, 1973

Nr 127 - S. Smolik: Rachunek wyrównawczy dla funkcji nieliniowych, 1973

Nr 128 - W. Żakowski: O pewnych własnościach n-wymiarowych prostych maszyn ciągłych, 1973

Nr 129 - R. Wajs: O pewnych własnościach maszyn relacyjnych, 1973

Nr 130 - W. Lipski, W. Marek: On Hamiltonian paths in a graph, 1973

Nr 131 - I. Nabiałek: Pewne własności funkcji τ-obliczalnych, 1973

Nr 132 - Z. B. Miądowicz: On the realization of automata by iterative networks, 1973

Nr 133 - W. Marek, Z. Pawlak: Mathematical foundations of information storage and retrieval. Part 1, 1973 .