BULLETIN DE L'ACADÉMIE
POLONAISE DES SCIENCES
série des sciences math., astr.
et phys. — Vol. XXII, No. 4, 1974

MATHEMATICS
(COMPUTER SCIENCE)

On the Foundations of Information Retrieval

by

W. MAREK and Z. PAWLAK

Presented by A. MOSTOWSKI on October 5, 1973

Summary. In this paper we present a logical approach towards a theory of Information Storage and Retrieval (ISR) systems. First a formal language for describing properties of document sets is introduced and the basic properties of this language are investigated. Then ISR system is defined in such a way that with each document set an expression of the language is associated (which "describes" it). The properties of IRS system are stated and operations on ISR systems are defined (like extension and reduction of systems). The presented theory leads to a new computer implementation method of ISR systems which will be published elsewhere.

1. Syntax.

DEFINITION 1.1. Let A, I be two given nonempty, disjoint sets, let $\{A_i\}_{i\in I}$ be some fixed partition of A (that is, $(\forall i)(\forall i')\ (i\neq i' \Rightarrow A_i\cap A_{i'}=\emptyset)$, $\bigcup_{i\in I} A_i = A$).

For a given set A we define the language $\mathscr{L}_A$ as follows: The alphabet of $\mathscr{L}_A$ contains:

1° constants c_a (for $a\in A$),
2° symbols $T, F, \vee, \wedge$,
3° auxiliary symbols $\neg, \vee, \wedge, \Rightarrow, \sim, +, \cdot, \rightarrow$,
4° symbol $=$.

DEFINITION 1.2. The set $\mathscr{T}$ of terms is the least set satisfying 1° and 2°.

1° $T\in\mathscr{T}$, $F\in\mathscr{T}$, $c_a\in\mathscr{T}$ $\quad(a\in A)$.
2° If $t_1, t_2\in\mathscr{T}$ then $\sim t_1, t_1+t_2, t_1\cdot t_2, t_1\rightarrow t_2\in\mathscr{T}$. In the sequel t, s (possibly with indices) range over terms.

DEFINITION 1.3. The set $\mathscr{F}$ of formulas is the least set satisfying 1° and 2°.

1° If t_1, t_2 are terms then $\ulcorner t_1=t_2\urcorner\in\mathscr{F}$.
2° If $\varphi_1, \varphi_2\in\mathscr{F}$ then $\neg\varphi_1, \varphi_1\vee\varphi_2, \varphi_1\wedge\varphi_2, \varphi_1\Rightarrow\varphi_2\in\mathscr{F}$.

We assume as axioms:

1° Substitutions of proposition calculus axioms (cf. [2]) — for formulas
2° Substitutions of the axioms of Boolean Algebra — for terms
3° If $i\in I$, $a\in A_i$, then

$$c_a = \sim\left(\sum_{\substack{b\in A_i\\ b\neq a}} c_b\right),$$

where $\sum$ is an abbreviation for sum of bigger amount of terms. (In case when each A_i, $i \in I$ is finite. If we admit A_i infinite, we need some modifications in syntax.) A is called basic dictionary and I — family of features.

2. Semantics, connections with the syntax.

DEFINITION 2.1. A system of information storage (i.s.r. system) over basic dictionary A and family of features I is a quadruple $\mathfrak{S}=\langle X, A, I, \mathscr{U}\rangle$ where

$$\mathscr{U}: A \to \mathscr{P}(X)$$

satisfies conditions:

1° If $i \in I$, $a, b \in A_i$, $a \neq b$ then

$$\mathscr{U}(a) \cap \mathscr{U}(b) = \emptyset .$$

2° If $i \in I$ then

$$\bigcup_{a \in A_i} \mathscr{U}(a) = X .$$

DEFINITION 2.2. Valuation of terms. Given system $\mathfrak{S}=\langle X, A, I, \mathscr{U}\rangle$ we define inductively $\|t\|_{\mathfrak{S}}$, $\|\varphi\|_{\mathfrak{S}}$ as follows:

(a) $\|c_a\|_{\mathfrak{S}} = \mathscr{U}(a)$,
(b) $\|\sim t\|_{\mathfrak{S}} = X - \|t\|_{\mathfrak{S}}$,
(c) $\|t_1 \cdot t_2\|_{\mathfrak{S}} = \|t_1\|_{\mathfrak{S}} \cap \|t_2\|_{\mathfrak{S}}$,
(d) $\|t_1 + t_2\|_{\mathfrak{S}} = \|t_1\| \cup \|t_2\|_{\mathfrak{S}}$,
(e) $\|F\|_{\mathfrak{S}} = \emptyset$,
(f) $\|T\|_{\mathfrak{S}} = X$,
(g) $\|t_1 \to t_2\|_{\mathfrak{S}} = (X - \|t_1\|_{\mathfrak{S}}) \cup (\|t_2\|_{\mathfrak{S}})$.

Now assume $\|t\|_{\mathfrak{S}}$ is defined for all $t \in \mathscr{T}$

$$\|t_1 = t_2\|_{\mathfrak{S}} = \begin{cases} \vee & \text{if } \|t_1\|_{\mathfrak{S}} = \|t_2\|_{\mathfrak{S}}, \\ \wedge & \text{if } \|t_1\|_{\mathfrak{S}} \neq \|t_2\|_{\mathfrak{S}}, \end{cases}$$

$$\|\neg \varphi\|_{\mathfrak{S}} = \begin{cases} \wedge & \text{if } \|\varphi\|_{\mathfrak{S}} = \vee , \\ \vee & \text{if } \|\varphi\|_{\mathfrak{S}} = \wedge . \end{cases}$$

For other connectives we extend our definition in a natural way.

THEOREM 2.1. *If φ is an axiom then $\|\varphi\|_{\mathfrak{S}} = V$.*

DEFINITION 2.3. Let $\mathfrak{S}=\langle X, A, I, \mathscr{U}\rangle$ be a system $x \in X$.

(a) An information on x in $\mathfrak{S}$ is a function $f: I \to A$ such that $f(i) \in A_i$ and $(i)_I\big(x \in U(f(i))\big)$.
(b) A description of x in $\mathfrak{S}$ is a term

$$\prod_{i \in I} c_{f(i)} .$$

Clearly an information on x determines a description of x (up to a possible order of I).

DEFINITION 2.4. A system $\mathfrak{S}$ is selective iff:
For all $x \in X$, if t is a description of x in $\mathfrak{S}$ then $\|t\|_{\mathfrak{S}} = \{x\}$.
Thus a selective system is one in which any two elements are distinguishable.

3. Completeness property of informational systems.

DEFINITION 3.1. (a) We define $c_a^0 = c_a$, $c_a^1 = \sim c_a$.

(b) A term t is called primitive if $t = c_{a_0}^{\varepsilon_0} \dots c_{a_n}^{\varepsilon_n}$ where each ε_j is 0 or 1.

(c) A term t is in normal additive form if $t = \sum t_j$ where each t_j is primitive term.

(d) A term t is in positive form if $\sim, \rightarrow$ does not occur in t.

The axioms accepted in Sec. 1 allow us to prove formulas (in the theory of information systems).

We use $\vdash$ to denote the existence of a proof of the formula.

THEOREM 3.1.

(a) *If t is a term then there is a term t_1 in normal additive form such that $\vdash t = t_1$.*

(b) *If t is a term then there is a term t_2 in positive normal additive form such that $\vdash t = t_2$.*

DEFINITION 3.2.

(a) A primitive term is called complete iff for every $i \in I$ there is exactly one $a \in A_i$ such that c_a occurs in t.

(b) A term t is in complete positive normal additive form iff $t = \sum_k t_k$ and each t_k is complete positive primitive term.

THEOREM 3.2. *If I is finite then for each term t there is exactly one term t_3 (being in complete positive normal additive form) such that $\vdash t = t_3$.*

DEFINITION 3.3. We introduce relations $\leqslant, \approx$ on $\mathscr{T}$ as follows

1° $t_1 \leqslant t_2 \Leftrightarrow$ there is t such that $\vdash t + t_1 = t_2$

2° $t_1 \approx t_2 \Leftrightarrow \vdash t_1 = t_2$.

This is nothing else but Lindenbaum algebra on $\mathscr{T}$.

LEMMA 3.3.

(a) *$\leqslant$ is a partial ordering in $\mathscr{T}$,*

(b) *$\approx$ is an equivalence relation in $\mathscr{T}$,*

(c) *$\leqslant$ generates $\approx$ i.e. $t_1 \leqslant t_2$ and $t_2 \leqslant t_1 \Rightarrow t_1 \approx t_2$.*

DEFINITION 3.4.

(a) A term t is semantically less than term s if for all information systems $\mathfrak{S}$

$$\|t\|_{\mathfrak{S}} \subseteq \|s\|_{\mathfrak{S}}.$$

(b) Term t is semantically equal to the term s iff for all information systems $\mathfrak{S}$: $\|t\|_{\mathfrak{S}} = \|s\|_{\mathfrak{S}}$.

THEOREM 3.4. (Completeness property for terms).

(a) *The term t is semantically less than the term s iff $t \leqslant s$.*

(b) *The term t is semantically equal to the term s iff $t \approx s$.*

4. Describable sets.

DEFINITION 4.1. Let $\mathfrak{S} = \langle X, A, I, U \rangle$ be a system of i.s.r. A set $Y \subseteq X$ is called describable within $\mathfrak{S}$ iff there is $t \in T$ such that $\|t\|_{\mathfrak{S}} = Y$.

LEMMA 4.1. *Describable subsets of X form boolean algebra.*

LEMMA 4.2. *If $\mathfrak{S}$ is finite selective system then every subset $Y \subseteq X$ is describable.*

Since the fact that every subset is describable implies selectiveness we get — by Lemma 1.2.

THEOREM 4.1. *If $\mathfrak{S}$ is finite i.s.r. system then $\mathfrak{S}$ is selective iff every subset of X is describable within $\mathfrak{S}$.*

5. Operations on i.s.r. systems.

DEFINITION 5.1. Let $\mathfrak{S}=\langle X, A, I, U\rangle$ be an i.s.r. system. Let $\{I_j\}_{j\in J}$ be a partition of the set I. An induced family of systems $\{\mathfrak{S}_j\}_{j\in J}$ is formed as follows: $\mathfrak{S}_j=\langle X, A^j, I_j, U_j\rangle$ where

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

(b) $U_j=U \restriction A^j$ (where $\restriction$ is a restriction sign).

In the same way the family of languages $\{\mathscr{L}_j\}$ is induced. Clearly $\mathscr{L}_j$ corresponds to $\mathfrak{S}_j$.

DEFINITION 5.2. Let $\{\mathfrak{S}_j\}_{j\in J}$ be a family of i.s.r. systems $(\mathfrak{S}_j=\langle X, A^j, I_j, U_j\rangle)$ and moreover $i\neq j \Rightarrow A^i\cap A^j=\emptyset=I_i\cap I_j$. We define:

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

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

DEFINITION 5.3. Let $\mathfrak{S}_i=\langle X_i, A^i, I_i, U_i\rangle$ $(i=0, 1)$ be two i.s.r. systems. We say that $\mathfrak{S}_0\subseteq\mathfrak{S}_1$ iff

(a) $X_0\subseteq X_1$,

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

(c) $I_0\subseteq I_1$,

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

(e) $\forall_{i\in I_0} A_i^0=A_i^1$.

LEMMA 5.1. *Assume $\mathfrak{S}=\langle X, A, I, U\rangle$ is i.s.r. system, $\{I_j\}_{j\in J}$ is a partition of I and $\{\mathfrak{S}_j\}_{j\in J}$ is induced family. Then for each $j\in J$, $\mathfrak{S}_j\subseteq\mathfrak{S}$.*

This shows adequacy of Definitions 2.1. and 2.3.

LEMMA 5.2. *Under obvious assumptions $\mathfrak{S}_j\subseteq\bigoplus_{j\in J}\mathfrak{S}_j$.*

THEOREM 5.1. *Assume $\mathfrak{S}_0\subseteq\mathfrak{S}_1$ and let t be a term of the language $\mathscr{L}_{A^0}$. Then $\|t\|_{\mathfrak{S}_0}=\|t\|_{\mathfrak{S}_1}\cap X_0$.*

DEFINITION 5.4. (a) $\mathfrak{S}_0\underset{*}{\subseteq}\mathfrak{S}_1$ iff $\mathfrak{S}_0\subseteq\mathfrak{S}_1$ and $X_0=X_1$

(b) $\mathfrak{S}_0\underset{A}{\subseteq}\mathfrak{S}_1$ iff $\mathfrak{S}_0\subseteq\mathfrak{S}_1$ and $A^0=A^1$.

LEMMA 5.3. *If $\mathfrak{S}_0\underset{A}{\subseteq}\mathfrak{S}_1$ then $I_0=I_1$.*

THEOREM 5.2. *If $\mathfrak{S}_0\subseteq\mathfrak{S}_1$ then there are i.s.r. systems $\mathfrak{S}_2$ and $\mathfrak{S}_3$ such that*

$$\mathfrak{S}_0 \begin{array}{l} \underset{A}{\subseteq}\ \mathfrak{S}_2\ \underset{*}{\subseteq} \\ \underset{*}{\subseteq}\ \mathfrak{S}_3\ \underset{A}{\subseteq} \end{array} \mathfrak{S}_1.$$

THEOREM 5.3. (a) *If $\mathfrak{S}$ is i.s.r. system and $Y \subseteq X$ then there is $\mathfrak{S}$ such that $\mathfrak{S} \subseteq_* \mathfrak{S}'$ and Y is describable within $\mathfrak{S}'$.*

(b) *If $\mathfrak{S}$ is finite i.s.r. system, $\mathfrak{A}$ is a boolean algebra of describable sets (within $\mathfrak{S}$) and $\mathfrak{B}$ is any boolean algebra of subsets of X such that $\mathfrak{A} \subseteq \mathfrak{B} \subseteq P(X)$ then there is $\mathfrak{S}'$ such that $\mathfrak{S} \subseteq_* \mathfrak{S}'$ and $\mathfrak{B}$ is exactly boolean algebra of describable subsets of $\mathfrak{S}'$.*

6. Implementation restrictions.

DEFINITION 6.1. A family $\mathfrak{A} \subseteq \mathfrak{P}(X)$ is called conus iff $(A)_{\mathfrak{A}}(B)(B \subseteq A \to B \in \mathfrak{A})$.

DEFINITION 6.2. A sufficient information in selective i.s.r. system $\mathfrak{S}$ is a conus $\mathfrak{A}$ containing all singeltons.

DEFINITION 6.3. A term t is conforming s.i. $\mathfrak{A}$ iff $\|t\| \in \mathfrak{A}$.

LEMMA 6.1. *The set of terms conforming s.i. $\mathfrak{A}$ forms a subfamily of $\mathfrak{F}$ closed under $\cdot$.*

DEFINITION 6.4. If $\mathfrak{A}$ is a s.i. for $\mathfrak{S}$ we define

$$t_1 \sim_{\mathfrak{A}} t_2 \leftrightarrow (\|t_1\|_{\mathfrak{S}} = \|t_2\|_{\mathfrak{S}} \in \mathfrak{A}) \vee (\|t_1\|_{\mathfrak{S}} \notin \mathfrak{A} \quad \text{and} \quad \|t_2\|_{\mathfrak{S}} \notin \mathfrak{A}).$$

LEMMA 6.3. *$\sim_{\mathfrak{A}}$ is an equivalence.*

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

DEFINITION 6.5. Every term $t_1 \leqslant t$ such that $\|t_1\| \in \mathfrak{A}$ is called sufficient extension of t for $\mathfrak{A}$.

LEMMA 6.4. *The set of sufficient extensions of t for $\mathfrak{A}$ is closed under $\cdot$. However it needs not be closed under $+$.*

In practical situations we consider systems with numeration.

DEFINITION 6.6. Let $\langle T, \leqslant \rangle$ be an ordered set. If $\mathfrak{S} = \langle X, A, I, U \rangle$ 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 $\varphi * X$ (that is, an image of X under φ), isomorphic to $\mathfrak{S}$.

DEFINITION 6.7. Term t is called segmental is ordered i.s.r. system $\langle \mathfrak{S}, \varphi \rangle$ iff $\varphi * (\|t\|_{\mathfrak{S}})$ is a segment in $\langle T, \leqslant \rangle$

The segments are particularily convenient in the process of retrieval.

Thus we may wish to have certain terms in segmental forms.

LEMMA 6.5. *The family of segmental terms in $\langle \mathfrak{S}, \varphi \rangle$ is closed under $\cdot$.*

Let $\mathfrak{M}$ be a family of terms such that $(t_1, t_2)_{\mathfrak{M}}(t_1 \neq t_2 \to \|t_1\|_{\mathfrak{S}} \cap \|t_2\|_{\mathfrak{S}} = \emptyset)$ then we have

LEMMA 6.6. *There is a well ordered set $\langle T, \leqslant \rangle$ and enumeration $\varphi: X \xrightarrow[onto]{1-1} T$ such that each term $t \in \mathfrak{M}$ is segmental. Moerover we may order that fixed term $t \in \mathfrak{M}$ generates an initial segment of T.*

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

Let $\mathfrak{M}$ be a family of terms, $\mathfrak{S}$ an i.s.r. system. $\mathfrak{M}$ is said to satisfy condition C with respect to $\mathfrak{S}$ iff $\mathfrak{M}$ decomposes into two subfamilies $\mathfrak{M}'$ and $\mathfrak{M}''$ such that:

(a) Every two different terms in $\mathfrak{M}'$ have disjoint values (in $\mathfrak{S}$).

(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 $\|t\|_{\mathfrak{S}} \subseteq \left\| \sum_{t \in W} t \right\|_{\mathfrak{S}}$ 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 $\mathfrak{S}$) are not included in that of t.

We have:

THEOREM. *If $\mathfrak{M}$ satisfies condition C with respect to $\mathfrak{S}$ then there is ordering $\langle T \leqslant \rangle$ and φ: $X \xrightarrow[\text{onto}]{1-1}$ such that all terms in t are segmental in $\langle \mathfrak{S}, \varphi \rangle$.*

Proof of this theorem will be published elsewhere.

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

INSTITUTE OF MATHEMATICS, POLISH ACADEMY OF SCIENCES, 00-950 WARSAW
(INSTYTUT MATEMATYCZNY, PAN)
COMPUTATION CENTRE, POLISH ACADEMY OF SCIENCES, 00-901 WARSAW
(CENTRUM OBLICZENIOWE PAN, WARSZAWA)

REFERENCES

[1] Z. Pawlak, *Mathematical foundations of information retrieval*, CC PAS Report 101.
[2] R. Lyndon, *Notes on logic*, Princeton, 1966.
[3] K. Kuratowski, A. Mostowski, *Set theory*, Warszawa, 1967.
[4] W. Marek, Z. Pawlak, *Mathematical foundations of information storage and retrieval I*, CC PAS, Reports 131.
[5] — , — , *Mathematical foundations of information storage and retrieval II*, ibid., 132.

В. Марэк, З. Павляк, **Об основах нахождения информации**

Содержание. В настоящей работе дается логический подход к теории нахождения информации. Представленная в работе теория приводит к новому методу машинного восстановления систем нахождения информации.