Theoretical Computer Science 1 (1976) 331–354. © North-Holland Publishing Company

INFORMATION STORAGE AND RETRIEVAL SYSTEMS: MATHEMATICAL FOUNDATIONS

Wiktor MAREK and Zdzisław PAWLAK

Institute of Mathematics, Polish Academy of Sciences, Warsaw, Poland
Computation Center, Polish Academy of Siences, Warsaw, Poland

Communicated by E. Engeler
Received April 1974
Revised January 1975

Abstract. We introduce in this paper a certain mathematical model of information storage and retrieval system. This is based on a certain family of languages intermediate between sentential and predicate calculi. To each language there is a corresponding semantics. We investigate our languages from the logical point of view showing their completeness. This is done by exhibiting nice and natural sets of axioms for the i.s.r. systems. The class of models for our languages is then examined by the algebraical means. We introduce our algebraical operations in such a way that they correspond to the actions performed while updating the system (in various ways). We also examine the boolean algebra of describable sets (of documents). Having done all this we propose a new implementational algorithm for the i.s.r. systems based on the form of atoms in the boolean algebra of describable sets. In the appendix we show how to compute the code numbers of these atoms (called generalized components). Let us also note that the class of queries relevant to the system is quite rich in our case; we are not only able to ask questions of the form „what are the documents fitting the following description", but we may also ask certain general questions on the system as a whole.

1. Introduction

In this paper we present a new mathematical approach to some problems occuring in information storage and retrieval (i.s.r.) systems. By an i.s.r. systems $\mathcal{S}$ we mean a quadruple consisting of set of objects X (like books, documents, etc.) together with the set of descriptors A, the set of attributes I, and the function U which associates a subset of X to each descriptor from A. Attributes are to be understood as elements of A, all of the same „type", e.g., descriptors: green, blue, brown, and black form the attribute colour. Thus each object from X may be decribed in our system by a vector of descriptors from A exhausting all possible attributes from I. Sometimes "incomplete" descriptions (in the sense that not all possible attributes are specified in the descrip ion of an object) are of interest, however, we do not consider the case here.

Our first goal is to describe precisely some fundamental facts about i.s.r. systems. To do so we introduce in the first place a formal language tailored to deal with

the problem. This language is a sort of intermediate language between propositional and predicate calculi. We further show that the language is adequately chosen for our aims. We show how the language may be used to prove theorems about i.s.r. systems.

Then we introduce the notion of a describable set of objects and find necessary and sufficient conditions to determine whether all sets are describable in $\mathcal{S}$ or not. Since in general not all subsets X are describable in $\mathcal{S}$, we investigate the structure of the family of describable sets.

Since not all the subsets of X are — in general — describable and we may wish to have a more fine description of objects in our systems, we sometimes have to add some attributes or/and descriptors. If — on the other hand — our system is "too fine", we may remove some attributes and/or descriptors from the system. The set of objects in the systems may also be varying; it may increase or decrease. In order to take into account the dynamics of the system (in the above sense) we introduce some algebraic tools, i.e. operations on i.s.r. systems, and study properties of the systems thus changed.

Finally a computer implementation algorithm resulting from our considerations is briefly discussed and some other problems raised by our theory are stated at the end of the paper.

Let us finally note that rudimentary versions of this paper were circulated as preprints (see [6,7]). After finishing the paper we found that elementary considerations of the similar kind were already performed by Wong and Chiang [8].

Throughout the paper we accept standard mathematical notation and assume the reader to be familiar with it. In particular $\mathcal{P}(X)$ denotes the power set of X, $f \restriction Y$ is the restriction of f to Y and we distinguish descriptors (elements of A) from their names by setting the latter in bold face type.

We express our gratitude to Prof. A. Blikle, Mr. W. Lipski, Jr., and Prof. A. Mazurkiewicz for valuable discussions. Interesting remarks due to Prof. E. Engeler had strong influence on the final draft of the paper.

1. Syntax

Let A be a nonempty set and let R be a fixed equivalence on A. We assume that all equivalence classes of R are finite. Since R generates a partition $\{A_i\}_{i\in I}$ of A into family of equivalence classes, $A = \bigcup_{i\in I} A_i$, $i \neq j \Rightarrow A_i \cap A_j = \emptyset$, it is reasonable to call R, R_I. In the sequel A will be referred to as the set of *descriptors* and I will be called the set of attributes.

With each set A we associate the *description language* $\mathcal{L}_A$

Definiton 1.1 (Alphabet of the language $\mathcal{L}_A$). We define an *alphabet of the language* $\mathcal{L}_A$ as follows:

(1) constans **a** (for each $a \in A$),

(2) constants T, F,

(3) constants $\wedge$, $\vee$ (truth values, respectively truth and falsity),

(4) symbols $\sim$, $\cdot$, $+$, $\rightarrow$,

(5) symbols $\neg$, $\vee$, $\wedge$, $\Rightarrow$,

(6) symbol $=$.

Definition 1.2 (Terms of the language $\mathcal{L}_A$). The *set $\mathcal{T}$ of terms* is the least set $\underline{T}$ satisfying (1) and (2)

(1) $T \in \underline{T}$, $F \in \underline{T}$, $\boldsymbol{a} \in \underline{T}$;

(2) if $t_1, t_2 \in \underline{T}$, then $\sim t_1, \ulcorner t_1 + t_2 \urcorner, \ulcorner t_1 \cdot t_2 \urcorner, \ulcorner t_1 \rightarrow t_2 \urcorner \in T$.

As it will turn out later the order of the sum is immaterial and so we shall abbreviate finite sums as $\sum_{i \in J} t_i$, similarly for products.

Definition 1.3 (Formulas of the language $\mathcal{L}_A$). The *set $\mathcal{F}$* of formulas is the least set $\underline{F}$ satisfying (1) and (2).

(1) if $t_1, t_2 \in T$, then $\ulcorner t_1 = t_2 \urcorner \in \underline{F}, \in \wedge \in \underline{F}$;

(2) if $\Phi_1, \Phi_2 \in \underline{F}$, then $\sim \Phi_1, \ulcorner \Phi_1 \wedge \Phi_2 \urcorner, \ulcorner \Phi_1 \vee \Phi_2 \urcorner, \ulcorner \Phi_1 \Rightarrow \Phi_2 \urcorner \in \underline{F}$.

In the sequel the letters s,t (possibly with indices) will denote terms and Φ,Ψ possibly with indices) formulas.

Definition 1.3 (Axiomatization). We assume as axioms:

(1) substitution of the proposition calculus axioms for fomulas (see [5]);

(2) Substitution of the axioms of Boolean Algebra for terms (including equality axioms) (see [2]);

(3) $\boldsymbol{a} = \sim \sum \{\boldsymbol{b}: bR_I a \wedge b \neq \boldsymbol{a}\}$,

this is sometimes noted as

$$\boldsymbol{a} = \sum_{\substack{bR_I a \\ b \neq a}} \boldsymbol{b}$$

As an inference rule we take modus ponens.

Note that the restriction of R_I, namely that all equivalence classes of it are finite is essential in (3). In case when some A_i is infinite, the expression $\sum \{\boldsymbol{b}: bR_I a\}$ may be undefined. We could overcome this obstacle allowing infinite sums operator into the language. This leads to a parallel, more general theory, We shall not, however, pursue the matter in this paper.

2. *Semantics, interpretation of terms and formulas*

Definition 2.1. (Basic definition). An information storage and retrieval system (i.s.r. system) is a quadruple

$$\mathcal{I} = \langle X, A, R_I, U \rangle,$$

where X is some set called the carrier of S and elements of X are referred to as objects of $\mathcal{I}$. A is the set of descriptors in $\mathcal{I}$ and R_I is an equivalence on A of finite index. U maps A into $\mathcal{P}(X)$ ($U : A \to \mathcal{P}(X)$) and satisfies the following two conditions:

(1) if $aR_Ib \wedge a \neq b$, then $U(a) \overset{U}{\cap} (b) = \emptyset$;

(2) $\bigcup \{U(b) : bR_Ia\} = X$ (for each $a \in A$).

Conditions (1) and (2) may be expressed equivalently as (1'), (2').

(1') if $i \in I$, $a \in A_i$, $b \in A_i$, $a \neq b$, then $U(a) \cap U(b) = \emptyset$;

(2') $\bigcup\limits_{a \in A_i} U(a) = X$ (for each $i \in I$).

Definition 2.2. (Valuation of terms). Let $\mathcal{I} = \langle X, A, R_I, U \rangle$ be an i.s.r. system. We define inductively the value of a term t in $\mathcal{I}$, $\|t\|_{\mathcal{I}}$, as follows:

(a) $\|a\|_{\mathcal{I}} = U(a)$

(b) $\|\sim t\|_{\mathcal{I}} = X \setminus \|t\|_{\mathcal{I}}$,

(c) $\|t_1 \cdot t_2\|_{\mathcal{I}} = \|t_1\|_{\mathcal{I}} \cap \|t_2\|_{\mathcal{I}}$,

(d) $\|t_1 + t_2\|_{\mathcal{I}} = \|t_1\|_{\mathcal{I}} \cup \|t_2\|_{\mathcal{I}}$,

(e) $\|t_1 \to t_2\|_{\mathcal{I}} = (X \setminus \|t_1\|_{\mathcal{I}}) \cup \|t_2\|_{\mathcal{I}}$,

(f) $\|F\|_{\mathcal{I}} = \emptyset$,

(g) $\|T\|_{\mathcal{I}} = X$.

Definition 2.3 (Valuation of formulas). Unlikely to the terms, formulas will take as values truth values $\bigvee$ and $\bigwedge$, we define inductively $\|\Phi\|$ (we assume that $\|t\|_{\mathcal{I}}$ is already defined):

(a) $\|\bigvee\|_{\mathcal{I}} = \bigvee$, $\|\bigwedge\|_{\mathcal{I}} = \bigwedge$;

(b) $\|t_1 = t_2\|_{\mathcal{I}} = \begin{cases} \bigvee & \text{if } \|t_1\|_{\mathcal{I}} = \|t_2\|_{\mathcal{I}}, \\ \bigwedge & \text{otherwise;} \end{cases}$

(c) $\|{\rightharpoonup}\Phi\|_{\mathcal{I}} = \begin{cases} \bigvee & \text{if } \|\Phi\|_{\mathcal{I}} = \bigwedge, \\ \bigwedge & \text{if } \|\Phi\|_{\mathcal{I}} = \bigvee. \end{cases}$

(d) for other connectives we extend our definition in the natural way.

Theorem 2.4. (Adequacy of axiomatization) *If Φ is an axiom, then* $\|\Phi\|_{\mathcal{I}} = \bigvee$.

Proof. As our valuation was defined in a way to make the first two groups of axioms true, it is enough to check axiom three. Therefore we need to show that

$$\|a = \sim \sum_{\substack{bR_I a \\ b \neq a}} b\|_{\mathcal{G}} = \vee,$$

i.e., according to Definition 2.3(b) that

$$\|a\|_{\mathcal{G}} = \|\sim \sum_{\substack{bR_I a \\ b \neq a}} b\|_{\mathcal{G}}.$$

easy transformation, according to Definition 2.2(b) and (d) reduces the problem to checking

$$\|a\|_{\mathcal{G}} = X \setminus \bigcup_{\substack{bR_I a \\ b \neq a}} \|b\|_{\mathcal{G}}.$$

This, however, is easily seen to be equivalent to Definition 2.1 (1) and (2) □

There is nothing strange in that we used in our proof only conditions (a), (b) and (d), from Definition 2.2 since other connectives may be expressed with the help of $\sim$ and $+$.

Definition 2.5. Let $\mathcal{G} = \langle X, A, R_I, U \rangle$ be an i.s.r. system. Let $x \in X$.

(a) An *information* on x in $\mathcal{G}$ is a function $f_x : I \to A$ such that for all $i \in I$, $f_x(i) \in A$ and $x \in U(f_x(i))$.

(b) A *description* of x in $\mathcal{G}$ is a term $t_x = \bigsqcap_{i \in I} f_x(i)$.

An information on x in $\mathcal{T}$ determines several terms, all of them provably equivalent (they differ only in that the order of contants occuring in them may be different). This explains our usage of one symbol, t_x.

This leads to:

Definition 2.6. An i.s.r. system $\mathcal{G}$ is *selective* iff for all $x \in X$, $\|t_x\|_{\mathcal{G}} = \{x\}$.

Thus a selective system is one in which different elements have necessarily diffe-descriptions, i.e., are distinguishable.

3. Completeness properties of i.s.r. systems

Since we introduced in Section 1 an axiom system and the rule of inference, we are able to prove formulas. We denote by $\vdash \Phi$ the fact that Φ is provable. It is immediate from Theorem 2.4 and the fact that the rule modus ponens is preseved under $\|\cdot\|_{\mathcal{G}}$

(i.e., $\|\Phi\|_{\mathcal{G}} = \bigvee$ and $\|\Phi \Rightarrow \psi\|_{\mathcal{G}} = \bigvee$ implies $\|\psi\|_{\mathcal{G}} = \bigvee$) cf. Definition 2.2 for a definition of $\|\cdot\|_{\mathcal{G}}$ that the following lemma holds:

Lemma 3.1. (Adequacy of inference). *If* $\vdash \Phi$, *then, for all i.s.r. systems* $\mathcal{G}$, $\|\Phi\|_{\mathcal{G}} = \bigvee$.

We shall also obtain a conseverse result soon.

Definition 3.2. We define relations $\leqslant$ and $\approx$ on $\mathcal{T}$ as follows:

$$t_1 \leqslant t_2 \text{ iff } \vdash t_2 = t_1 + t_2,$$
$$t_1 \approx t_2 \text{ iff } \vdash t_1 = t_2.$$

Lemma 3.3. (a) $\leqslant$ *is reflexive and transitive,* (b) $\approx$ *is an equivalence relation,* (c) $\leqslant$ *has the antisymmetry property with respect to* $\approx$, *i.e.,* $t_1 \leqslant t_2 \wedge t_2 \leqslant t_1 \Rightarrow t_1 \approx t_2$.

Each i.s.r. system $\mathcal{G}$ generates relations $\leqslant_{\mathcal{G}}$ and $\approx_{\mathcal{G}}$ as follows:

Definition 3.4. (a) $t_1 \leqslant_{\mathcal{G}} t_2 \Leftrightarrow \|t_1\|_{\mathcal{G}} \subseteq \|t_2\|_{\mathcal{G}}$,
(b) $t_1 \approx_{\mathcal{G}} t_2 \Leftrightarrow \|t_1\|_{\mathcal{G}} = \|t_2\|_{\mathcal{G}}$.

Lemma 3.5. $t_1 \leqslant t_2 \Rightarrow (\forall \mathcal{G})\, t_1 \leqslant_{\mathcal{G}} t_2$,

$t_1 \approx t_2 \Rightarrow (\forall \mathcal{G})\, t_1 \approx_{\mathcal{G}} t_2$.

Proof. As in Lemma 3.1. □

This leads to the following:

Definition 3.6. $t_1 \leqslant^* t_2 \Leftrightarrow (\forall \mathcal{G}) t_1 \leqslant_{\mathcal{G}} t_2$,

$t_1 \approx^* t_2 \Leftrightarrow (\forall \mathcal{G})\, t_1 \approx_{\mathcal{G}} t_2$.

Notice that $t_1 \leqslant^* t_2$ means that in every interpretation, t_1 determines smaller set then t_2 does. Similarly $t_1 \approx^* t_2$ means that in every interpretation both terms determine the same set.

Thus Lemma 3.5 says that $\leqslant \subseteq \leqslant^*$ and $\approx \subseteq \approx^*$. In the sequel we shall prove converse inclusions.

Definition 3.7. (a) We define $a^0 = a$, $a^1 = \sim a$.

(b) A term t is called *primitive* iff $t = \prod_{j \in J} a_j^{\xi_j}$ where each ξ_j is 0 or 1.

(c) A term t is in *normal additive* form iff $t = \sum_{j \in J} t_j$, where each t_j is primitive.

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

Theorem 3.8. (Normal form I). *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 term t_2 in positive normal additive form such that* $\vdash t = t_2$.

Proof. (a) A reasoning used in this case is a standard one; we refer the reader to [5].

(b) By (a) we may assume that t is already in normal additive form. Using the axioms $x = y \Rightarrow \sim x = \sim y$, $\sim\sim x = x$, from Definition 1.3 (3) we get

$$\sim \boldsymbol{a} = \sum_{\substack{b R_I a \\ b \neq a}} \boldsymbol{b}.$$

Putting right-hand side in every place where the left-hand side occurs we eliminate negation from t_1. Consecutive applications of the distributive law finish the proof. □

Definition 3.9. (a) A primitive term t is called *complete* iff for every $i \in I$ there is exactly one $a \in A_i$ such that $\boldsymbol{a}$ occurs in t.

(b) A term t is in complete positive normal additive form iff $t = \sum_{k \in K} t_k$ and each t_k is a complete positive primitive term.

Theorem 3.10. (Normal form II). *For each term t there is a term t_3 in complete positive normal additive form such that* $\vdash t = t_3$.

Proof. It is clear that it is enough to find such a term for a positive primitive term, by Theorem 3.8 (b). Since $\vdash \sim \boldsymbol{a} = \sum_{\substack{b R_I a \\ b \neq a}} \boldsymbol{b}$, we have $\vdash \sum_{b R_I a} \boldsymbol{b} = T$. Using in turn $\vdash t \cdot T = t$ we get $t \cdot \sum_{b \in A_i} b = t$. Assume that no b (for $b \in A_i$) occurs in t, then $t = \sum_{b \in A_i} t \cdot b$. Thus we diminished in t a number of i such that no b (for $b \in A_i$) occurs in t. Since I is finite this gives an inductive procedure. □

Notice that t_3 is unique up to a possible order of primitive terms and a possible order within the terms.

Theorem 3.11. (Completeness property for terms).

(a) $t_1 \leqslant t_2$ iff $t_1 \leqslant^* t_2$,

(b) $t_1 \approx t_2$ iff $t_1 \approx^* t_2$.

Proof. Clearly (b) follows from (a).

(a) $\Rightarrow$ was already proved in Lemma 3.5.

(a) $\Leftarrow$ Assume $t_1 \leqslant^* t_2$. We may assume that both t_1 and t_2 are in complete normal positive additive form. It is clear that if every primitive term occuring in t_1

occurs also in t_2, then $t_1 \leqslant t_2$. Thus it is enough to show the this first property holds. Assume it is not true.

Let t_0 be a primitive term occurring in t_1 but not in t_2. We construct an i.s.r. system $\mathcal{S}'$ in which $\|t_0\|_{\mathcal{S}'} \neq \emptyset$. Using the fact (which we leave to the reader) that different primitive complete positive terms have disjoint values, we find that $\|t_1\|_{\mathcal{S}'}$ is not included in t_2, which contradicts $t_1 \leqslant^* t_2$. □

Dually to the additive normal form one may — as usually — introduce a multiplicative normal form.

A construction from the proof of theorem 3.11 suggests the following question: Is there an i.s.r. system $\mathcal{S}$ such that $\approx_{\mathcal{S}}$ is identical with $\approx$?

In fact there is one. A construction of it strongly resembles the construction of a family $\{A_n\}$ such that all components corresponding to it are nonempty (cf. [2]).

Construction: Let each A_i be in the form $\{a^i_1 \ldots a^i_{k_i}\}$. We produce the Cartesian product $\mathbf{P}_{i\in I} A_i$ and define $\mathcal{S}_{\max}$ as follows

$$\mathcal{S}_{\max} = \langle \mathbf{P}_{i\in I} A_i A, R_I, U\rangle,$$

where $U(a) = \{f \in \mathbf{P}_{i\in I} A_i : f(i) = a$, for the unique i such that $a \in A_i\}$.

We leave it to the reader to check that the system $\mathcal{S}_{\max}$ has the property that each complete primitive positive term has in $\mathcal{S}_{\max}$ a non void value.

Before we prove the completeness theorem for formulas we need some terminology and facts.

(1) By a procedure similar to that applied in case of transformation of a term into normal form, we are able to transform every formula into the following form: $\Phi := \Phi_1 \wedge \ldots \wedge \Phi_k$, where each Φ_j is of the form $\psi_{i_1} \vee \ldots \vee \psi_{i_j}$ and each ψ_j is of the form $t_1 = t_m$ or of the form $t_k \neq t_r$ for some terms t being in normal, positive, additive, complete form. We describe this fact symbolically as $\Phi = \bigwedge \bigvee \psi_j$.

(2) Another fact needed in proof is the following: $t = s$ is equivalent to the conjunction of equations of the form $t'_1 = F, \ldots, t'_k = F$. Indeed assume that both t and s are in the positive normal additive complete form. Then $t = t_1 + \ldots + t_m$ and $s = s_1 + \ldots + s_n$. There are possibly some primitive terms which appear in both expansions. Let $t'_1, \ldots, t'_k$ be primitive terms which appear in either t or s but not in both. We leave it to the reader to prove that $\vdash (t = s) \Leftrightarrow (t'_1 = F \wedge \ldots \wedge t'_k = F)$. Similarly $t \neq s \Leftrightarrow (t'_1 \neq F \vee \ldots \vee t'_k \neq F)$.

(3) Finally let us note that if $\Phi := \Phi_1 \wedge \ldots \wedge \Phi_k$, then $\vdash \Phi$ iff for all $1 \leqslant j \leqslant k$ $\vdash \Phi_j$.

Theorem 3.12. (Completeness theorem for formulas). $\vdash \Phi$ *iff for all* $\mathcal{S}, \|\Phi\|_{\mathcal{S}} = V$.

Proof. $\Rightarrow$ was already proved in Theorem 2.4 (adequacy theorem). $\Leftarrow$ By our remark (1) we may assume that Φ is $\bigwedge \bigvee \psi_j$ where each ψ_j is of the form $t_m = t_s$ or of the form $t_n \neq t_k$. We want to prove that $\vdash \Phi$. By remark (3) it is sufficient to show that $\vdash \bigvee \psi_j$. We shall transform $\bigvee \psi_j$ to certain form which

finally allows to find a proof for it. Indeed, using remark (2) we may substitute for ψ_j either a conjuction $(t_{l_1} = F \wedge \ldots \wedge t_{l_k} = F)$ or alternative $(t_{j_1} \neq F \vee \ldots \vee t_{j_m} \neq F)$ depending whether ψ_j is $t = s$ or else $t \neq s$.

Thus instead of the proof of $\mathbb{W}\psi_j$ we need a proof of a certain formula Ξ built from the primitive formulas of the form $t_r = F$ and $t_u \neq F$ where each t_v is in primitive, positive, complete form.

Using our remark (1) once more we find a formula Ξ_1 in conjunctive normal form equivalent to Ξ. Since in the process of building conjunctive normal form no new atomic formula is used, we find that our formula Ξ has the form $\mathbb{M} \mathbb{W} \Theta_m$ where each Θ_m is of the form $t_1 = F$ or $t_j \neq F$ and t_k are in primitive, positive, complete, normal form. Since $\mathbb{M} \mathbb{W}\psi_j$ was valid in every i.s.r. system, so is $\mathbb{W}\psi_j$. Thus also Ξ and Ξ_1 are valid in all i.s.r. systems since the transformation used in the above reasoning preserves validity. Thus $\mathbb{M} \mathbb{W} \Theta_m$ is valid in every i.s.r. system. This in turn is equivalent to the fact that $\mathbb{W}\Theta_m$ is valid in every i.s.r. system. But $\mathbb{W} \Theta_m$ is of the form.

$$t_1' = F \vee \ldots \vee t_k' = F \vee t_{k+1}'' \neq F \vee \ldots \vee t_{k+m}'' \neq F.$$

If we show a proof for $\mathbb{W}\Theta_m$, then we are done.

Now the main step of the proof follows: We claim that if $\mathbb{W}\Theta_m$ is valid in every i.s.r. system, then there must be a primitive term t such that both $t = F$ and $t \neq F$ occur among Θ_m. Assume this is not true. We construct a i.s.r. system $\mathcal{G}$ in which $\|\mathbb{W}\Theta_m\|_{\mathcal{G}} = F$.

Indeed such a system is produced from the previously constructed system $\mathcal{G}_{max}$ by throwing out generalized components corresponding to $t_{k+1}'', \ldots, t_{k+m}''$. Then since no t_j' is t_r'', in this particular system $\mathcal{G}$, $t_j' \neq F$ for all $1 \leqslant j \leqslant k$, and at the same time $t_r'' = F$ for all $k+1 \leqslant r \leqslant k+m$.

Thus $\|t_1' = F \vee \ldots \vee t_k' = F \vee t_{k+1}'' \neq F \vee \ldots \vee t_{k+m}'' \neq F\|_{\mathcal{G}} = \wedge$, contradicting the validity of $\mathbb{W}\Theta_m$. Therefore there must be a primitive term t such that both $t = F$ and $t \neq F$ occur among Θ_m. Since, however, formula $t = F \vee t \neq F$ is provable, so is formula $\mathbb{W}\Theta_m$. Since all alternatives $\mathbb{W}\Theta_m$ are provable, so is $\mathbb{M} \mathbb{W}\Theta_m$. Thus $\mathbb{W}\psi_j$ is provable, being equivalent to a provable formula. So finally $\mathbb{M} \mathbb{W}\psi_j$ is provable and thus $\vdash \Phi$. □

Let us notice, that as in any formalized system we may consider theories based on some additional axioms. (We shall encounter this situation in the sequel). In such enriched system we may again prove theorems. Let us denote the fact that a proof, using possibly additional axioms from T, exists, by $T \vdash \Phi$. By the reasonings virtually identical with that of Theorems 3.11 and 3.12 we prove:

Theorem 3.11. (Generalized completeness property for terms). $T \vdash t_1 = t_2$ *iff* $\|t_1 = t_2\|_{\mathcal{G}} = \vee$ *for every i.s.r.* system $\mathcal{G}$ such that $(\forall \Phi)\,(\Phi \in T \Rightarrow \|\Phi\|_{\mathcal{G}} = \vee)$.

Theorem 3.12. (*Generalized completeness property for formulas*). $T \vdash \Psi$ *iff for every i.s.r. system* $\mathcal{G}$ *such that* $(\forall \Phi)\,(\Phi \in T \Rightarrow \|\Phi\|_{\mathcal{G}} = V \Rightarrow \|\Psi\|_{\mathcal{G}} = V$.

As a corollary we get the following:

Theorem 3.13. *Let* $\mathcal{G}$ *be an i.s.r. system with A finite. Then there is a single formula* $\Phi_{\mathcal{G}}$ *such that, for all formulas* ψ,

$$\|\psi\|_{\mathcal{G}} = V \Leftrightarrow \Phi_{\mathcal{G}} \vdash \psi.$$

Theorem 3.13. was also proved — using different reasoning — by W. Lipski.

Definition 3.14. Let $\mathcal{G}$ and $\mathcal{G}'$ be i.s.r. systems. We say that $\mathcal{G}$ is equivalent to $\mathcal{G}'$ ($\mathcal{G} \equiv \mathcal{G}'$) iff for every Φ,

$$\|\Phi\|_{\mathcal{G}} = \|\Phi\|_{\mathcal{G}'}.$$

Obviously $\equiv$ is an equivalence relation.

The equivalence classes of $\equiv$ are determined (according to Theorem 3.13) by some special formulas. In fact the formula ψ determining the equivalence class of $\equiv$ is of the form $(\mathbb{M} t_i = F) \wedge (\mathbb{M} t_j \neq F)$, where t_i are those primitive complete positive terms whose value in $\mathcal{G}$ is empty, whereas t_j' are those whose value in $\mathcal{G}$ is nonempty. Using this remark we have Theorem 3.15. In every $\equiv$ equivalence class there is exactly one (up tc isomorphism) selective system. Thus for every system $\mathcal{G}$ there is a selective system $\mathcal{G}$ such that $\mathcal{G} \equiv \mathcal{G}'$.

We also get — as a corollary — the following fact: We are not able to express within the formal language of an i.s.r. system the number of elements of the universe of the i.s.r. system.

4. Algebraic properties of i.s.r. systems

Definition 4.1. Let $\mathcal{G}_X = \langle X, A, R_I, U_X \rangle$ and $\mathcal{G}_Y = \langle Y, B, R_J, U_Y \rangle$ be two i.s.r. systems. We say that $\mathcal{G}_X \subseteq \mathcal{G}_Y$ iff

(a) $X \subseteq Y$,

(b) $A \subseteq B$,

(c) $R_J \cap A^2 = R_I$,

(d) $(\forall a \in A) U_Y(a) \cap X = U_X(a)$.

Let us note that whenever $A \subseteq B$ then $\mathcal{L}_A \subseteq \mathcal{L}_B$. Thus the terms of $\mathcal{L}_A$ are, in particular, terms of $\mathcal{L}_B$.

The adequacy of our definition is shown by the following:

Lemma 4.2. *If* $\mathcal{G}_X \subseteq \mathcal{G}_Y$ *and* t *is a term of the language* $\mathcal{L}_A$, *then* $\|t\|_{\mathcal{G}_X} = \|t\|_{\mathcal{G}_Y} \cap X$.

Proof. By induction on complexity of t. If t is a, then the desired equality is nothing else but Definition 4.1(d). In case of F and T the condition is seen immediately.

Assume now that t is $\sim t_1$. We have

$$\|t\|_{\mathcal{G}_X} = \|\sim t_1\|_{\mathcal{G}_X} = X - \|t_1\|_{\mathcal{G}_X},$$

by inductive assumption we have $\|t\|_{\mathcal{G}_X} = X - (X \cap \|t_1\|_{\mathcal{G}_Y})$.

But

$$X - (X \cap \|t_1\|_{\mathcal{G}_Y}) = X \cap Y - (X \cap \|t_1\|_{\mathcal{G}_Y}) =$$

$$= (Y - \|t_1\|_{\mathcal{G}_Y}) \cap X = X \cap \| \sim t_1\|_{\mathcal{G}_Y}.$$

If $t = t_1 \cdot t_2$, then we have

$$\|t\|_{\mathcal{G}_X} = \|t_1\|_{\mathcal{G}_X} \cap \|t_2\|_{\mathcal{G}_X} = \|t_1\|_{\mathcal{G}_Y} \cap X \cap \|t_2\|_{\mathcal{G}_Y} \cap X$$

(here inductive assumption is used) thus

$$\|t\|_{\mathcal{G}_X} = \|t_1\|_{\mathcal{G}_Y} \cap \|t_2\|_{\mathcal{G}_Y} \cap X = \|t_1 \cdot t_2\|_{\mathcal{G}_Y} \cap X = \|t\|_{\mathcal{G}_Y} \cap X.$$

The case when $t = t_1 + t_2$ is similar. Finally when $t = t_1 \to t_2$ we eliminate the case using equality $t_1 \to t_2 = (\sim t_1) + t_2$ and then applying inductive assumption. □

Definition 4.3. (a) $\mathcal{G}_X \overset{*}{\subseteq} \mathcal{G}_Y$ iff $\mathcal{G}_X \subseteq \mathcal{G}_Y$ and $X = Y$,

(b) $\mathcal{G}_X \underset{*}{\subseteq} \mathcal{G}_Y$ iff $\mathcal{G}_X \subseteq \mathcal{G}_Y$ and $A = B$.

Fact. 4.4. *If* $\mathcal{G}_X \underset{*}{\subseteq} \mathcal{G}_Y$, *then* $R_I = R_J$.

Proof. Immediate by Definition 4.1(c). □

Theorem 4.5. (Interpolation property). *If* $\mathcal{G}_X \subseteq \mathcal{G}_Y$, *then there are systems* $\mathcal{G}'$ *and* $\mathcal{G}''$ *such that the following holds*:

(1°) $\mathcal{G}_X \underset{*}{\subseteq} \mathcal{G}' \overset{*}{\subseteq} \mathcal{G}_Y$,

(2°) $\mathcal{G}_X \overset{*}{\subseteq} \mathcal{G}'' \underset{*}{\subseteq} \mathcal{G}_Y$.

Proof. Define $\mathcal{G}'$ as follows: $\mathcal{G}' = \langle Y, A, R_I, U' \rangle$, where $U' = U_Y \restriction A$; and $\mathcal{G}''$ as follows: $\mathcal{G}'' = \langle X, B, R_J, U'' \rangle$, where $U''(b) = U_Y(b) \cap X$.

It is straighforward to check that $\mathcal{G}_X \subseteq \mathcal{G}', \mathcal{G}' \subseteq \mathcal{G}_Y, \mathcal{G}_X \subseteq \mathcal{G}'', \mathcal{G}'' \subseteq \mathcal{G}_Y$ because all remaining conditions hold by our construction. □

Definition 4.6. Let $\mathcal{S} = \langle X, A, R_I, U\rangle$ be an i.s.r. system. Let $\{I_j\}_{j\in J}$ be a partition of the set I. An induced family $\{\mathcal{S}_j\}_{j\in J}$ of i.s.r. systems is formed as follows: $\mathcal{S}_j = \langle X, A^j, R_{I_j}, U_j\rangle$, where

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

(b) $R_{I_j} = R_I \cap (A^j \times A^j)$,

(c) $U_j = U \restriction A^j$.

Lemma 4.7. *Under the assumptions of Definition* 4.6, *for each* $j \in J$, $\mathcal{S}_j \overset{*}{\subseteq} \mathcal{S}$.

Proof. Since the universe of $\mathcal{S}_j$ is X, it is sufficient to prove that $\mathcal{S}_j \subseteq \mathcal{S}$. But all the conditions (b), (c) and (d) of Definition 4.1 are easily seen to be satisfied. □

Since each subset $I' \subseteq I$ induces a partition $I = I' \cup (I - I')$, we naturally get a restriction of $\mathcal{S}_I$ or $\mathcal{S}$ to $I' \subseteq I$ and the complementary system $\mathcal{S}_{I-I'}$.

Definition 4.8. Let $\{\mathcal{S}_j\}_{j\in J}$ be a family of i.s.r. systems with the same carrier ($\mathcal{S}_j = \langle X, A^j, R_{I_j}, U_j\rangle$) and suppose moreover that $i \neq j \Rightarrow A^i \cap A^j = \emptyset$.

Define $\bigoplus_{j\in J} \mathcal{S}$ as follows:

$$\bigoplus_{j\in J} \mathcal{S}_j = \langle X, A, R_I, U\rangle,$$

where $A = \bigcup_{j\in J} A^j$, $R_I = \bigcup_{j\in J} R_{I_j}$, $U = \bigcup_{j\in J} U_j$.

Lemma 4.9. *Under the assumptions of Definition* 4.8, $\mathcal{S}_j \subseteq \bigoplus_{j\in J} \mathcal{S}_j$, *moreover if the family* $\{\mathcal{S}_j\}_{j\in J}$ *is obtained as in Definition* 4.6, *then* $\mathcal{S} = \bigoplus_{j\in J} \mathcal{S}_j$.

We leave the proof to the reader.

Definition 4.10. *Let* R, S be equivalences on a set Z; we say that $S \prec R$ iff $S \leqslant R$, i.e.,

$$(\forall x)\,(\forall y)\,(xSy \Rightarrow xRy).$$

It is clear $\prec$ is a partial ordering (i.e., that it is reflexive, antysymmetric and transitive).

Definition 4.11. (a) Let R be an equivalence on Z. $Z/_R$ consists of all equivalence classes of R in Z.

(b) Let S be an equivalence on Z and R an equivalence on $Z/_S$. We define a relation $R * S$ on Z as follows $xR * Sy \Leftrightarrow (x/_S)\, R (y/_S)$.

Lemma 4.12. (a) *Under the assumptions of Definition* 4.11, $R * S$ *is an equivalence relation on* Z.

(b) *Moreover* $R \prec R * S$.

Proof: (a) is a simple computation.

(b) assume xRy, then $x/_R = y/_R$ and so, by reflexivity of S, we get $x/_R \, S \, y/_R$, i.e., $xS * Ry$. □

Lemma 4.13. *Assume* $S * (T * R)$ *is defined. Then* $(S * T) * R$ *is defined and* $S * (T * R) = (S * T) * R$.

Proof. Assume T is defined on $A/_R$ and S defined on $({}^{A}/_{R})/_{T}$. Then $S * T$ is defined on $A/_R$ and so $(S * T) * R$ is defined.

Let $xS * (T * R)\, y$. Then $(x/_T)/_R S (y/_T)/_R$. Having in mind that $(x/_T)/_R$ consists of all $y/_R$ which are (with $x/_R$) in relation T we find that

$$(x/_R)\, S * T\, (y/_R)$$

which is the desired result. □

Lemma 4.14. *If* $S \prec R$, *then there is unique* T *such that*:

$$R = T * S.$$

Proof. We define T as follows:

$$(x/_S) T (y/_S) \quad \text{iff } x \, R \, y.$$

T is an equivalence. Clearly T is both reflexive and symmetric. If $(x/_S)T(y/_S)$ and $y/_S \, T \, z/_S$, then $x \, R \, y$ and $y \, R \, z$, i.e., $(x/_S)T(z/_S)$.

The uniqueness of T is easily proved e.g. by contraposition. □

Definition 4.15. Let S be an equivalence on A, $S \prec R_I$, $\mathcal{I} = \langle X, A, I, U \rangle$ be an i.s.r. system. We define the quotient system $\mathcal{I}/_S$ as follows

$$\mathcal{I}/_S = \langle X, A/_S, T_I, U_S \rangle,$$

Where

(a) T_I is the unique relation such that $R_I = T_I * S$,

(b) $U/_S(a/_S) = \bigcup \{U(b) : bSa\}$.

Note that the equivalence classes of T may be indexed by I which makes our Definition 4.15 (a) reasonable.

Clearly $\mathcal{I}/_S$ determines its language $\mathcal{L}_{A/S}(A/_S)$.

Let us form now $\mathcal{I} \oplus \mathcal{I}/_S$. The language corresponding to this system consists of constants $\boldsymbol{a}$ for $\boldsymbol{a} \in A$ and $\boldsymbol{a}/_S$ for $a/_S \in A/_S$.

The system obtained is such a way is denoted by $\mathcal{I}_S$.

Lemma 4.16. *If S satisfies the assumptions of Definition* 4.15, *and $\mathcal{G}_S$ is the resulting system, then* $\|a/_S = \sum_{bSa} b\|_{\mathcal{G}S} = \bigvee$.

Proof. We need only to show that $\|a/_S\|_{\mathcal{G}_S} = \bigcup_{bSa} \|b\|_{\mathcal{G}_S}$, but since $\|b\|_{\mathcal{G}_S} = \|b\|_{\mathcal{G}}$, the right-hand side is $\bigcup_{bSa} \|b\|_{\mathcal{G}_S}$, i.e., $\bigcup \{U(b) : bSa\}$. On the other hand, $\|a/_S\|_{\mathcal{G}_S} = \|a/_S\|_{\mathcal{G}/_S}$, i.e., $U/_S(a/_S)$ which is by definition $\bigcup \{U(b): bSa\}$. □

The full power of the operation $\oplus$ and in the same time the generality of our approach allowing inclusion of the hierarchical approach is seen after Theorem 4.18.

Definition 4.17. Let $S_1 \prec \ldots \prec S_n \prec R_I$ be an increasing sequence of equivalence relations on A, we define $_{S1\ldots Sn}$ as follows:

$$\mathcal{G}_{S_1\ldots S_n} = \mathcal{G} \oplus (\bigoplus_{i=1}^{n} \mathcal{G}/_{S_i}).$$

Let $T_1,\ldots,T_{n-1}$ be equivalences such that $S_{i+1} = T_i * S_i$.

Theorem 4.18.

$$\left\| a/_{S_{i+1}} = \sum_{b/_{S_i}\, T_i a/_{S_i}} b/_{S_i} \right\|_{\mathcal{G}_{S_1\ldots S_n}} = \bigvee.$$

Proof. It is clear that it is enough to give the proof for the case $S_1 \prec S_2 \prec R_I$, $S_2 = T * S_1$. Indeed, for $a \in A$,

$$\left\| a/_{S_2} = \sum_{bS_2a} b \right\|_{\mathcal{G}_{S_2}} = \bigvee$$

and so,

$$\left\| a/_{S_2} = \sum_{bS_2a} b \right\|_{\mathcal{G}_{S_1S_2}} = \bigvee.$$

Similarly,

$$\left\| a/_{S_1} = \sum_{bS_1a} b \right\|_{\mathcal{G}_{S_1S_2}} = \bigvee.$$

Since, however, $S_1 \prec S_2$, we have, for $a \in A$,

$$\left\| a/_{S_1} \right\|_{\mathcal{G}_{S_1S_2}} \subseteq \left\| a/_{S_2} \right\|_{\mathcal{G}_{S_1S_2}}.$$

Using the idempotence laws we get

$$\left\| a/_{S_2} = \sum_{bS_2a} b/_{S_1} \right\|_{\mathcal{G}_{S_1S_2}} = \bigvee.$$

There are identical terms on the right-hand side and grouping them together we find that they correspond exactly to the equivalence classes of the relation T, which gives the desired equation. □

The hierarchical construction is used when our system is "too fine", i.e., when the attributes are too small.

Let us see this on the following example: In the system $\mathcal{S}$ the attribute colour has as its descriptors the following: red_{I}, red_{II}, $green_{\mathrm{I}}$, $green_{\mathrm{II}}$, $green_{\mathrm{III}}$, $blue_{\mathrm{I}}$, $blue_{\mathrm{II}}$. By grouping the descriptors: $\{red_{\mathrm{I}}, red_{\mathrm{II}}\}$, $\{green_{\mathrm{I}}, greeu_{\mathrm{II}}, green_{\mathrm{III}}\}$, $\{blue_{\mathrm{I}}\}$, $\{blue_{\mathrm{II}}\}$, we get now in the appropriate $\mathcal{S}/_S$ the attribute colour having 4 descriptors: *red*, *green*, $blue_{\mathrm{I}}$, $blue_{\mathrm{II}}$. In the system $\mathcal{S} \oplus \mathcal{S}/_S$ the following equalities hold: $red = red_{\mathrm{I}} + red_{\mathrm{II}}$, $green = green_{\mathrm{I}} + green_{\mathrm{II}} + green_{\mathrm{III}}$.

A converse construction is used when the system is "too crude" and when we need to split some descriptors (this should be used specially when the value of a descriptor is a too big set of objects). We present the construction below.

As introduced, for each $i \in I$, $\{U(a) : a \in A_i\}$ is a decomposition of X. Let T_i be an equivalence relation (on X) corresponding to this decomposition.

Assume now, that for each $i \in I$ there is an equivalence W_i on X such that $W_i \prec T_i$. The family $\{W_i\}_{i \in I}$ generates an i.s.r. system $\mathcal{S}^W = \langle X, B, R_j, V \rangle$ as follows:

$$B = \bigcup_{i \in I} \{x/_{W_i} : i \in I\},$$

$$R_I = \{\langle x/_{W_i}, y/_{W_j} \rangle : i = j\},$$

$$V(x/_{W_i}) = \{y : yW_ix\}.$$

Theorem 4.19. *$\mathcal{S}$ is isomorphic to a certain quotient system of $\mathcal{S}^W$.*

Proof. It is enough to give the relation S such that $\mathcal{S}$ is isomorphic to $\mathcal{S}^W/_S$. Since for each $i \in I$, $W_i \prec T_i$, S_i is unique such that $T_i = S_i * W_i$. Put $S = \cup S_i$. We leave to the reader details of the proof that $\mathcal{S}^W/_S$ is isomorphic to $\mathcal{S}$. □

Similarly we have:

Theorem 4.20. *If $S \prec R_I$, then there is W such that $(\mathcal{S}/_S)^W$ is isomorphic to $\mathcal{S}$.*

5. Describable sets

Defini ion 5.1. Let $\mathcal{S} = \langle X, A, R_I U \rangle$ be an i.s.r. system, Let $Y \subseteq X$.

(a) Y is said to be describable within $\mathcal{S}$ iff there is a term t such that $\|t\|_{\mathcal{S}} = Y$.

(b) $\mathcal{B}(\mathcal{S})$ is the family of all subsets of X describable within $\mathcal{S}$.

Lemma 5.2. (a) *Describable subsets of X form a Boolean algebra.*

(b) *Moreover if Y is a describable subset of X then the subsets of Y which are describable in X also form a Boolean algebra.*

Proof. (a) follows directly from the choice of the axioms for our system.

(b) follows from the fact that if t is a description of Y in $\mathcal{S}$ (i.e., $\|t\|_{\mathcal{S}} = Y$), then the values of terms of the form $t \cap s$ (s ranging over $\mathcal{S}$) form a boolean algebra. □

Lemma 5.3. *If $\mathcal{S}$ is a finite selective i.s.r. system, then $\mathcal{B}(\mathcal{S}) = 2^X$ (recall that 2^X is the Boolean algebra of all subsets of X).*

Proof. Assume that t_x is a description of x in $\mathcal{S}$ (i.e., $\|t_x\|_{\mathcal{S}} = \{x\}$), then $t_Y = \sum_{y \in Y} t_y$ is a description of Y in $\mathcal{S}$. □

Remark. Here is a point in which a difference between finite and infinite i.s.r. system occurs. Indeed assuming the language $\mathcal{L}_A$ finitary (i.e. allowing only finitary conjunctions and disjunctions) with A infinite it is easy to produce infinite selective system with indescribable subset (by cardinality argument).

Theorem 5.4. If $\mathcal{S}$ is a finite te i.s.r. system then $\mathcal{S}$ is selective iff $\mathcal{B}(\mathcal{S}) = 2^X$

Proof. $\Rightarrow$ was proved in 5.3.
$\Leftarrow$ Since $\mathcal{B}(\mathcal{S}) = 2^X$ then in particular $\{x\} \in \mathcal{B}(\mathcal{S})$. We need to show that $\|t_x\|_{\mathcal{S}} = \{x\}$, (where t_x was introduced in 2.5). Let $\|t\| = \{x\}$. We may assume that t is in complete positive additive normal form. Thus $t = \sum t_r$ where each t_r is primitive term. if, for each t_r occuring in t, $\|t_r\|_{\mathcal{S}} \neq \{x\}$ then, since $\|t_r\|_{\mathcal{S}} \leqslant \|t\|_{\mathcal{S}}$ we have $\|t_r\| = \emptyset$ and so $\|t\| = \emptyset$. But this is not the case and so form some t_r, $\|t_r\|_{\mathcal{S}} = \{x\}$. Thus t_r is description of x. □

Theorem 5.5. (a) *If $\mathcal{S}$ is in an i.s.r. system and $Y \subseteq X$ then there is $\mathcal{S}'$ such that $\mathcal{S} \subseteq \mathcal{S}'$ and $Y \in \mathcal{B}(\mathcal{S}')$*

(b) *If $\mathcal{S}$ is an i.s.r. system and $\mathcal{B}$ a Boolean algebra such that $\mathcal{B}(\mathcal{S}) \subseteq \mathcal{B} \subseteq 2^X$ then there is $\mathcal{S}'$ such that $\mathcal{S} \subseteq \mathcal{S}'$ and $\mathcal{B}(\mathcal{S}') = \mathcal{B}$*

Proof. (a) If Y is describable in $\mathcal{S}$ put $\mathcal{S}' = \mathcal{S}$. Assume Y not describable within $\mathcal{S}$. Add two new elements a and a' (both not in $A \cup I$) to the set A. Define R' on $A \cup \{a,a'\}$ as follows

$$R' = R \cup \{\langle a,a'\rangle, \langle a',a\rangle, \langle a,a\rangle, \langle a', a'\rangle\}$$

form $\mathcal{S}'$ as follows

$$\mathcal{S}' = \langle X, A \cup \{a, a\}, R', U'\rangle$$

where $U'(b) = U(b)$ whenever $b \in A$

$$U'(a) = Y$$

$$U'(a') = X - Y.$$

(b) Let us notice the following easy fact from the theory of Boolean algebras.

If $Y \in \mathcal{B}$, $\mathfrak{A} \subseteq \mathcal{B}$ ($\mathfrak{A}$,$\mathcal{B}$ Boolean algebras of sets) then the smallest Boolean algebra containing $\mathfrak{A}$ and Y, $[\mathfrak{A}, Y]$ is included in $\mathcal{B}$.

Now we proceed as follows. We order the elements of $\mathcal{B}-\mathcal{B}(\mathcal{S})$ into (possibly transfinite) sequence $\{Y_\alpha\}_{\alpha<\beta}$ and form an increasing family of i.s.r. systems $\{\mathcal{S}_\alpha\}_{\alpha<\beta}$ as follows: $\mathcal{S}_{\alpha+1}$ is $\mathcal{S}'_\alpha$ (Operation was described in the proof of part *a*) if $Y_\alpha \notin \mathcal{B}(\mathcal{S}_\alpha)$ or $\mathcal{S}_\alpha$ if $Y \in \mathcal{B}(\mathcal{S}_\alpha)$.

In the limit step λ we take a union of $\mathcal{S}_\beta$, $\beta \in \lambda$ Using the fact mentioned at the beginning of the proof we find that for all $\alpha < \beta$, $\mathcal{B}(\mathcal{S}_\alpha) \subseteq \mathcal{B}$ and since $Y_\alpha \in \mathcal{B}(\mathcal{S}_{\alpha+1})$, and $\mathcal{B}(\mathcal{S}) \subseteq \mathcal{B}(\mathcal{S}_\alpha)$ for all $\alpha < \beta$ we get

$$\mathcal{B} = \mathcal{B}(\mathcal{S}) \cup \bigcup_{\alpha<\beta} \{Y_\alpha\} \subseteq \bigcup_{\alpha<\beta} \mathcal{B}(\mathcal{S}_\alpha) \subseteq \mathcal{B}$$

thus $\bigcup_{\beta<\alpha} \mathcal{B}(\mathcal{S}_\beta) \subseteq \mathcal{B}$. But, by construction, the left hand side is $\mathcal{B}(\mathcal{S}_\beta)$ and

$$\mathcal{S} \subseteq \mathcal{S}_{\mathcal{B}}$$

The construction given in § 3, as we mentioned resembles that of components (cf. [2]). The selectiveness of the i.s.r. system means that each generalized component (i.e., value of primitive complete positive term) consists of at most one element. If each of the components is nonempty then the system is isomorphic with the universal system, constructed in Section 3. It is clear that every nonempty set of the set of components determines selective system and conversly. This allows us to calculate the cardinality of the family of all selective systems (up to isomorphism) over A and I.

Indeed let $I = \{0...k\}, \bar{A}_c = n_i$. Then we have

Theorem 5.6. *There is exactly* $2^{\prod_{i=1}^{k} n_i} - 1$ of *nonempty selective systems over A and* R_I.

Producing an isomorphic copy each $\mathcal{S}_j$, $0 \leqslant j < 2^{\prod_{i=0}^{k} n_i} - 1$ we are able to produce a finite system (in extended language) such that each $\mathcal{S}_j$ is isomorphic with certain subsystem of $\mathcal{S}$. (In fact it is simply $\bigoplus_j \mathcal{S}_j$).

One may even produce an infinite system $\mathcal{S}'$ universal in the above sense for all finite (even non selective) systems over A and I.

Let us remark that $\mathcal{B}(\mathcal{S})$ is a Boolean algebra of subsets of X generated by $\|t\|_{\mathcal{S}}$ where t is primtive complete, positive normal term. This fact has deep implementational consequences.

While performing $\overset{*}{\subseteq}$ operation the notion of the generalized components change; "old" generalized components are unions of "new" generalized components.

While performing $\underset{*}{\subseteq}$ operation generalized components do not change in the sense that the trace of a generalized component in new system on the carrier of old system is again a generalized component (in old system). However if $\mathcal{S}_1 \underset{*}{\subseteq} \mathcal{S}_2$ then there may be some generalized components which are empty is $\mathcal{S}_1$ but not in $\mathcal{S}_2$. However, if a component is nonempty in $\mathcal{S}_1$ then it is also nonempty in $\mathcal{S}_2$. In the hierarchical operations $\mathcal{S}/_S$ and $\mathcal{S}^W$ the components are glued together (in the first case) and split in parts (in the second).

6. Implementation, combinatorial problems

Our syntactical approach suggests the following implementational proposal: We store in the memory documents as follows; documents belonging to a generalized component are stored "together". Then, any query is transformed into the alternative of the description of generalized components; thus we need only to find the generalized components. A quasi-practical suggestion is the following: In the linearly ordered memory, the documents are stored such that the generalized components form segments in the ordering. Then, each component is determined by the address of its beginning and the end. Thus, while the query is received we transform it into the normal, positive complete form and find the addresses corresponding to the primitive components of the term obtained. Similarly the question in the form of statement about our system is reduced as in the proof of completeness theorem the conjunction of alternatives of terms of the form $t_i = F$ or $t_j \neq F$ where t_i and t_j are primitive complete positive terms and thus checked.

Definition 6.1. Let $\langle T. \leqslant \rangle$ be a linearly ordered set and let $\mathcal{S} = \langle X, A, R_I, U \rangle$ be an i.s.r. system.

(a) A function $\varphi : T \xrightarrow{\text{onto}} X$ is called enumeration of $\mathcal{S}$.

(b) A function $\varphi: T \xrightarrow[1-1]{\text{onto}} X$ is called one-one enumeration of $\mathcal{S}$.

Roughly speaking enumeration is a listing of element of X in certain order possibly with repetitions.

Definition 6.2. A term t is called segmential in the enumeration φ iff there is a segment $W \subseteq T$ such that the image of W, $\varphi * W$ is $\|t\|_{\mathcal{S}}$.

It is obvious that segmential terms are particularly useful in the i.s.r. processes. We need therefore some criterion to determine whether we may find an enumeration in which given term is segmential.

Lemma 6.3. There always is a linearly ordered set $\langle T, \leqslant \rangle$ and enumeration φ such that all terms $t \in \mathcal{S}$ are segmential.

Proof. List all terms t ($t \in \mathcal{T}$) and consecutively order $\|t\|_{\mathcal{S}}$.

However this enumeration can be useful only in case of very simple i.s.r.systems. In fact there will be a lot of repetitions, and so the memory will be used competely uneconomically. The most important case is when the enumeration used is one-one.

Definition 6.4. A family of terms H is linear over $\mathcal{S}$ if there is a set $\langle T, \subseteq \rangle$ and one-one enumeration φ of $\mathcal{S}$ in T such that for all $t \in H$, t is segmential in φ.

Theorem 6.5. *If for all* $t_1, t_2 \in H$, $t_1 \neq t_2 \rightarrow \|t_1\|_{\mathcal{S}} \cap \|t_2\|_{\mathcal{S}} = 0$ *then* H *is linear over* $\mathcal{S}$.

Proof. Let us list all elements of H and order them consecutively, the elements of $X - \bigcup_{t \in H} \|t\|_{\mathcal{G}}$ are listed at the end.

Corollary 6.6. The family of primitive complete, positive normal terms is linear over every i.s.r. system $\mathcal{G}$.

Proof. They satisfy assumptions of 6.5.

Definition 6.7. Let H be a family of terms. $\mathrm{Sub}_{\mathcal{G}}(H)$ is a family of all primitive, normal, complete positive terms which are implicants of elements of H i.e. $t_1 \in \mathrm{Sub}_{\mathcal{G}}(H)$ iff t_1 is primitive normal, complete positive and there is $t \in H$ such that $\|t_1\|_{\mathcal{G}} \subseteq \|t\|_{\mathcal{G}}$.

Theorem 6.8. *If H is linear then also $H \cup sub_{\mathcal{G}}(H)$ is linear.*

Proof. Let φ be one-one enumeration of X in which all elements od H were segmential. We show how to change φ in such a way that all generalized components of terms occurring in H become segments. Indeed the component may be split into the segments; then we fix one of them and push it up to contain all other parts. This operation, when consecutively applied to all generalized components, gives the result.

Therefore we may conclude that in order to know whether or not given family H of terms is linear over $\mathcal{G}$ it is enough to check whether or not this family linear over unique (up to isomoprnism) selective system with the same theory.

The question whether or not given family H of terms is linear over $\mathcal{G}$ may be reduced to the problem of so called interval graphs (cf. [1]).

In that paper there is a condition under which a graph is an isomoprhic to the incidence graph on the real line. Thus considering a family $H \cup \mathrm{Sub}_{\mathcal{G}}(H)$ we are able to find whether it is linear or not. The method given there, together with Theorem 6.8. allows to check linearity of H. We shall not pursue the matter in this paper. If however H is not linear, we run into the problem of chosing of an enumeration (which is not one-one then) optimal (for instance with respect to the power of T). This problem is treated by Lipski and Marek [4].

7. Dynamical treatment

One can remark that our approach allows to see an i.s.r. system in "microscopic", i.e., static situation. Yet in a "real" situation, we have to modify our system according to requirements which may consist of:

(a) Changing of the set of documents — increasing or decreasing
(b) Adding or deleting of attributes (and so descriptors too)
(c) Changing a descriptors within the attributes.

Let us note that relations $\overset{*}{\subseteq}$, $\overset{*}{\supseteq}$, $\underset{*}{\subseteq}$, $\underset{*}{\supseteq}$ serve to enable us to speak about first two problems; the third one is treated as follows; With the help of hierarchical relationship we are able to make the attributes "more crude" and with the help of division relationship "more fine". Thus we want to express the following "meta--theorem". Relations: $\overset{*}{\subseteq}$, $\overset{*}{\supseteq}$, $\underset{*}{\subseteq}$, $\underset{*}{\supseteq}$, $\mathcal{G}/_S$, $\mathcal{G}^W$ are sufficient to describe what happens in real time while the i.s.r. system is subjected to accommodation changes.

8. Problems

The following general question seems to be of great importance:

$Q1$. How should be the memory of a computer organized to simplify the implementation of i.s.r. system?

For the important results in this direction we refer the reader to Lipski [3].

Another problem which seems to be of great practical importance is the following:

In the axioms of i.s.r. systems we assume that the classification is complete i.e. every element of our i.s.r. has full description. Thus:

$Q2$. What properties of our theory are preserved if we admit that some elements are not fully classified? Again some results were obtained in this direction by Lipski and the first author.

Appendix

While the i.s.r. system is implemented, it is necessary to enumerate the generalized components. If $A_i = n_i$ $(i \in I)$ then there is exactly $\prod_{i \in I} n_i$ of generalized components. We may assume that $I = \{1,...,k\}$ (i.e., $i = k$). Thus generalized components may be viewed as sequences

$$\langle b_1 ... b_k \rangle \text{ where } 0 \leqslant b_1 \leqslant n_1 - 1 ... 0 \leqslant b_k \leqslant n_k - 1$$

We know that set of all these sequence has power $\prod_{i=1}^{k} n_i$ and so is equipollent with the set $\{0,...,n_1,...,n_k,-1\}$ However we should be able to decode in some simple way form the number $0 \leqslant a \leqslant n_1 \cdot n_2 ... n_k - 1$ the sequence $\langle b_1,...,b_k \rangle$ it codes.

We may assume that each $n_i \neq 1$ since if $n_i = 1$ then in the representing sequence there will be always 0 at i-th position.

We define

$$\begin{aligned} u_0 &= n_1 \cdot \ldots n_k, \\ u_1 &= n_2 \cdot \ldots \cdot n_k, \\ u_{k-1} &= n_k, \\ u_k &= 1. \end{aligned}$$

By our assumption $u_0 > u_1 > u_2 > \ldots > u_k$

Theorem A1. *For every integer* $0 \leqslant a \leqslant n_1 \ldots n_k - 1$ *there is exactly one sequence* $\langle b_1, \ldots, b_k \rangle$ *such that*

(a) $0 \leqslant b_i \leqslant n_i - 1$

(b) $a = \sum_{i=1}^{k} b_i u_i$ (note the sum is taken from $i = 1$)

Proof. *Existence* Define b_i as follows

$$b_1 = E\left(\frac{a}{u_1}\right)$$

$$b_{n+1} = E\left(\frac{a - \sum_{j=1}^{n} b_j u_j}{u_{n+1}}\right)$$

(where E is "entier" function).

We prove first $0 \leqslant b_i \leqslant n_i - 1$.

This we show by simultaneous induction together with

(*) $$0 \leqslant a - \sum_{j=1}^{m-1} b_j u_j \leqslant n_m \cdot \ldots \cdot n_k - 1 \quad (\text{i.e. } u_{m-1} - 1)$$

Indeed, for $n = 1$

$$0 \leqslant a \leqslant n_1 \ldots n_k - 1$$

Then $0 \leqslant E\left(\frac{a}{u_1}\right) = E\left(\frac{a}{n_1 \ldots n_k}\right) \leqslant E\left(\frac{n_1 \ldots n_k - 1}{n_2 \ldots n_k}\right) = n_1 - 1$ i.e. $0 \leqslant b_1 \leqslant n_1 - 1$

Let us assume now

$$0 \leqslant b_{r-1} \leqslant n_{r-1} - 1$$

$$0 \leqslant a - \sum_{j=1}^{r-1} b_j u_j \leqslant u_{r-1} - 1$$

Then

$$0 \leqslant b_r = E\left(\frac{a - \sum_{j=1}^{r-1} b_j u_j}{u_r}\right) \leqslant E\left(\frac{u_{r-1} - 1}{u_r}\right) = E\left(\frac{n_r \cdot \ldots \cdot n_k - 1}{n_{r+1} \cdot \ldots \cdot n_k}\right) = n_r - 1$$

thus

$$0 \leqslant b_r \leqslant n_r - 1.$$

On the other hand

$$b_r = E\left(\frac{a - \sum_{j=1}^{r-1} b_j u_j}{u_r}\right) \text{ means}$$

$$0 \leqslant \frac{a - \sum_{j=1}^{r-1} b_j u_j}{u_r} - b_r \leqslant 1$$

thus

$$0 \leqslant a - \sum_{j=1}^{r-1} b_j u_j - b_r u_r < u_r$$

so

$$0 \leqslant a - \sum_{j=1}^{r} b_j u_j \leqslant u_r - 1$$

Now we prove $a = \sum_{j=1}^{k} b_j u_j$.

Notice that according to the definition

$$b_k = E\left(\frac{a - \sum_{j=1}^{k-1} b_j u_j}{u_k}\right)$$

since however $u_k = 1$ we have

$$b_k = a - \sum_{j=1}^{k-1} b_j u_j$$

thus

$$a = \sum_{j=1}^{k-1} b_j u_j + b_k = \sum_{j=1}^{k-1} b_j u_j + b_k u_k = \sum_{j=1}^{k} b_j u_j.$$

Uniqueness. Instead of showing this directly which is also possible (by the method we employ later) we notice that denoting

$$B = \{0, ..., n_1 \cdot ... \cdot n_k - 1\}$$

$$A = \{0, ..., n_1 - 1\} \times ... \times \{0, ..., n_k - 1\}.$$

We have $\overline{\overline{A}} = \overline{\overline{B}}$.

Our proof of existence exhibits 1—1 function of B into A. Since they have the same power it has to be onto, which shws uniqueness.

Let us notice that in the proof of existence we exhibited offective, iterative procedure which for every $0 \leqslant a \leqslant n_1 \cdot ... \cdot n_k - 1$ gives the sequence $\langle b_1, ..., b_k \rangle$. Let us write

$$\nu^{-1}(a) = \langle b_1, ..., b_k \rangle \text{ and}$$

$$\nu(b_1 ... b_k) = a$$

when $a = \sum_{j=1}^{k} b_j u_j$.

Question. What is the relation $\leqslant$ defined as follows?

$$\langle b_1, ..., b_k \rangle \leqslant \langle b'_1, ..., b'_k \rangle \leftrightarrow \nu(b_1, ..., b_k) \leqslant \nu(b'_1, ..., b'_k)$$

Theorem A2. $\leqslant$ *is lexicographic ordering of A.*

Proof. Since $\leqslant$ is connected therefore it is enough to show that $\langle b_1, ..., b_k\rangle \leqslant$ $\leqslant_{lex} \langle b'_1, ..., b_k\rangle \to \langle b_1, ..., b_k\rangle \leqslant \langle b'_1, ..., b'_k\rangle$ (where $\leqslant_{lex}$ is lexicographic ordering on A). I.e.

$$\langle b_1, ..., b_k\rangle, \leqslant_{lex} \langle b'_1, ..., b'_k\rangle \to \nu(b_1, ..., b_k) \leqslant \nu(b'_1, ..., b'_k),$$

Clearly it is enough to show

$$\langle b_1, ..., b_k\rangle \prec_{lex} \langle b'_1, ..., b'_k\rangle \to \nu(b_1, ..., b_k) < \nu(b'_1, ..., b'_k)$$

Thus we need show that under our assumption $\sum_{j=1}^{k} b_j u_j < \sum_{j=1}^{k} b'_j u_j$ holds.

We have $b_1 = b'_1, ..., b_{r-1} = b'_{v-1}$, $b_r < b'_r$

Thus we have to show $\sum_{j=r}^{k} b_j u_j < \sum_{j=r}^{k} b'_j u_j$

let us consider $\sum_{j=r+1}^{k} b_j u_j$

$$\sum_{j=r+1}^{k} b_j u_j = \nu(b_1, ..., b_k) - \sum_{j=1}^{r} b_j u_j \leqslant u_r - 1$$

(last inequality holds by (*))

Thus $\sum_{j=r}^{k} b_j u_j = b_r u_r + \sum_{j=r+1}^{k} b_j u_j \leqslant b_r u_r + u_r - 1 = (b_r + 1) u_r - 1$

but $(b_r + 1) u_r - 1 \leqslant b'_r u_r - 1 < b'_k u_r$

therefore we have

$$\sum_{j=r}^{k} b_j u_j < b'_r u_r \leqslant \sum_{j=r}^{k} b'_j u_j$$

thus

$$\sum_{j=r}^{k} b_j u_j < \sum_{j=1}^{k} b'_j u_j$$

As is clear from our construction, $\nu(b_1, ..., b_k) = \sum_{j=1}^{k} b_j u_j$. Since $b_j \leqslant n_j - 1$ therefore

$$\nu(b_1 ... b_k) \leqslant \nu(n_1 - 1, ..., n_k - 1) = \sum_{j=1}^{k} (n_j - 1) u_j = n \cdot ... \cdot n_k - 1$$

the last equality is easly provable by induction and we leave it to reader.

Let us finally note, that lexicographical ordering is specially convenient when, while extending the language we increase the number of attributes.

References

[1] J. Ch. Boland and C.C. Lekkerkerker, Representation of a finite graph by a set of intervals on the real line, Fund. Math. 51 (1962) 45–64.
[2] K. Kuratowski and A. Mostowski, Set Theory (North-Holland, Amsterdam, 1967).
[3] W. Lipski, Jr., Information storage and retrieval systems, mathematical foundations II, CC PAS Reports 153.
[4] W. Lipski Jr. and W. Marek, File organization, an application of graph theory, Springer Lecture Notes in Computer Science No. 14 (Springer, Berlin, 1974).
[5] R. Lyndon, Notes on Logic (Van Nostrand, Princeton, N. J., 1966).
[6] W. Marek and Z. Pawlak, Mathematical foundations of information storage and retrieval I, II, III, CC PAS Reports, 135, 6, 7.
[7] Z. Pawlak, Mathematical foundations of information retrieval, CC PAS Reports 101.
[8] E. Wong and T.C. Chiang, Canonical structure in attribute based file organization. Comm. ACM 14 (1970) 593–597.