MATHEMATICAL INSTITUTE OF THE
SLOVAK ACADEMY OF SCIENCES

COMPUTING RESEARCH CENTRE
UNITED NATIONS D.P. BRATISLAVA

mathematical foundations of computer science

proceedings of symposium and summer school

HIGH TATRAS
SEPTEMBER 3-8, 1973

MATHEMATICAL FOUNDATION OF INFORMATION RETRIVAL

Zdislav Pawlak
Computation Centre
Polish Academy of Sciences, Warsaw

In this paper we present a simple mathematical model of information retrival systems - based on some ideas given in [1],[2] and [4]. The presented theory gives besides mathematical formulation of basic ideas concerning information retrival a new simple implementation method.

1. Descriptive systems and descriptive sets

By a descriptive system we shall mean a triplet $S = \langle A_S, X_S, \delta_S \rangle$, or briefly $S = \langle A, X, \delta \rangle$, where

A - is a set; elements of A are called objects.

X - is a set called a description language; elements of X are referred to as descriptors.

δ - is a binary relation called semantics of X ($\delta \subseteq A \times X$).

We assume that X is the smalest set containing finite set $\hat{X}$ (the set of elementary descriptors) and such that if $x, y \in X$, then $x \wedge y$, $x \vee y$, $\sim x$ are also in X.

We shall replace the semantic relation δ by the function

$$\psi : X \to 2^A$$

such that

$$\psi(x) = \{a \in A : \delta(a, x)\},$$

where ψ is defined as follows:

$$\psi(x) = \hat{\psi}(x) \ , \quad \text{if } x \in \hat{X} \quad ,$$

$$\bigwedge_{x,y \in X} \psi(x \vee y) = \psi(x) \cup \psi(y),$$

$$\bigwedge_{x,y \in X} \psi(x \wedge y) = \psi(x) \cap \psi(y),$$

$$\bigwedge_{x \in X} \psi(\sim x) = A - \psi(x),$$

and $\hat{\psi}$ is the semantics of elementary descriptors, that is

$$\hat{\psi} : \hat{X} \to 2^A.$$

Every set $B \subseteq A$ such that $B = \hat{\psi}(x)$ where $x \in \hat{X}$ will be called elementary. We shall say that a set $B \subseteq A$ is descriptive in S - or if S is fixed - descriptive, if $B = \psi(x)$ for some $x \in X$. The class of all descriptive sets in S will be denoted by D_S. Of course D_S is the smalest class containing elementary sets in S and closed under set theoretical operations $\cup$, $\cap$, $\sim$.

Descriptors $x, y \in X$ are said to be equal if

$$\psi(x) = \psi(y).$$

Otherwise descriptors x,y are called different. We assume that all elementary descriptors in S are always different.

Theorem 1. For every descriptive system $S = \langle A, X, \delta \rangle$ the number of different descriptors is finite and is not greater than $2^{2^{\overline{\overline{X}}}}$.

2. Atomic descriptors and atomic sets

Every product of all elementary descriptors in S with or witout negation

$$x_1^{i_1} \wedge x_2^{i_2} \wedge \dots \wedge x_k^{i_k} , \qquad x_j^{i_j} \in X ,$$

where $i_j = 0$ or 1, $k = \bar{\bar{X}}$, and $x_j^0 = x_j$, $x_j^1 = x_j$ will be called atomic descriptor in S.

Of course for every S there are at most $2^{\bar{\bar{X}}}$ different atomic descriptors in S. If x is an atomic descriptor in S then $\psi(x)$ is called atom (or atomic set) in S. Descriptors $x,y \in X$ are said to be independent iff

$$\psi(x) \cap \psi(y) = \emptyset$$

Theorem 2. Every two different atomic descriptors in S are independent.

Theorem 3.

$$\bigcup_{x \in \bar{X}_S} \psi(x) = A_S$$

where $\bar{X}_S$ denotes the set of all atomic descriptors in S.

Theorem 4. Every elementary descriptor $x \in X_S$ may be represented as

$$x = x_1 \vee x_2 \vee \dots \vee x_n , \quad x_i \in \bar{X}_S,$$

where $x_1, x_2, \dots, x_n$ are all atomic descriptors in S containing x.

Theorem 5. Every descriptor $x \in X_S$ may be represented as the sum of some atomic descriptors in S. By means of theorem 5 we are able to represent descriptors in some standard (normal) form.

Theorem 6. $B \in D_S$ iff B is the sum of some atoms in S.

3. Remarks on implementation

Given some set of objects A (for example books, papers, documents, etc.) and the set of elementary descriptors X (for example authors names, languages, key words, etc.) thus the function $\hat{\psi}$ is defined. From above theorems it follows that we are unable to describe in X all possible sets of documents belonging to A - we can describe only sets which are sums of atoms in S (Theorem 6). This leads to a very simple computer implementation of information retrival systems: given any descriptor $x \in X_S$ by means of theorems 4 and 5 we can represent it in normal form and then according to the obtained formula find out corresponding atoms in the memory of the computer. So we do not need access to all elements of the set A in the memory but only access to the whole atoms, which simplifies the implementation considerably.

REFERENCES

[1] A. Mostowski, K.Kuratowski: "Teoria mnogości", PWN,1966

[2] Z. Pawlak: "About the meaning of personal pronouns", Cahiers de Linguistique Théoretique et Appliquée,Vol.X,1973,No 1

[3] Z. Pawlak: "Mathematical foundation of information retrival", CC PAS Reports, 101,1973,preprint

[4] Z. Semadeni: "Logical kits", Manuscript,1971