b 1426/101
1890 1

PRACE CO PAN • CC PAS REPORTS

Zdzisław Pawlak

Mathematical foundations of information retrieval

101
1973
WARSZAWA

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

Zdzisław Pawlak

MATHEMATICAL FOUNDATIONS OF INFORMATION RETRIEVAL

(A new approach)

101

Warszawa 1973

Kolegium Redakcyjne

A. Blikle (przewodniczący), J. Lipski (sekretarz), J. Łoś,
L. Łukaszewicz, R. Marczyński, A. Mazurkiewicz, Z. Pawlak,
Z. Szoda (zastępca przewodniczącego), M. Warmus

Mailing address: Prof. Dr. Z. Pawlak
Warszawa
ul. Gamerskiego 3 m.5
POLAND

Na prawach rękopisu

Printed as a manuscript

Nakład 450 egz., ark. wyd. 0,25; ark. druk. 0,50.
Papier offset., kl. III, 70 g, 70 x 100. Oddano do
druku w styczniu 1973 r. W. D. N. Zam. nr 32/0/73

b 1426/101 18902

Abstract

This note contains a simple mathematical formulation of basic ideas concerning information retrieval and its computer implementation. The presented theory is based on the results given in [1], [2] and [3].

Содержание

Работа касается математической формулировки основных терминов связанных с проблемой поиска информации и применения этого аппарата в цифровых машинах. Представленная теория опирается на результаты даны в [1], [2] и [3].

Streszczenie

Praca dotyczy matematycznego sformułowania podstawowych pojęć związanych z problemem wyszukiwania informacji oraz zastosowania tego aparatu na maszynach cyfrowych. Przedstawiona teoria oparta jest na wynikach podanych w [1], [2] i [3].

This note contains a simple mathematical formulation of basic ideas concerning information retrieval and its computer implementation. The presented theory is based on the results given in [1], [2] and [3].

1. Descriptive systems

By a descriptive system we mean triplet $\mathcal{D} = \langle A_{\mathcal{D}}, X_{\mathcal{D}}, \delta_{\mathcal{D}} \rangle$ (or briefly $\mathcal{D} = \langle A, X, \delta \rangle$), where

A - is a (finite or infinite) set; elements of A are called objects of $\mathcal{D}$,

X - is a finite set of symbols; elements of X are referred to as elementary descriptors of $\mathcal{D}$,

$\delta \subseteq A \times X$ - is a binary relation, called description relation (or description) in $\mathcal{D}$.

Relation δ may be replaced by the function:

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

such that:

$$\Psi(x) = \{a \in A ; \delta(a,x)\} .$$

Let X^* denote the smallest set containing X and such that: if $x,y \in X^*$ than $x \wedge y$, $x \vee y \sim x$ are also in X , and let Ψ^* be defined as follows:

$$\Psi^*(x) = \Psi(x) \quad \text{if} \quad x \in X$$

$$\bigwedge_{x,y\in X^*} \Psi^*(x\vee y) = \Psi^*(x) \cup \Psi^*(y)$$

$$\bigwedge_{x,y\in X^*} \Psi^*(x\wedge y) = \Psi^*(x) \cap \Psi^*(y)$$

$$\bigwedge_{x\in X^*} \Psi^*(\sim x) = A - \Psi^*(x),$$

where Ψ^* is the extension of Ψ (in what follows, the asterisk (*) will be omitted.)

Descriptors $x,y \in X^*$ are said to be equal ($x \equiv y$) iff

$$\Psi(x) = \Psi(y)$$

If $x \equiv y$ does not hold, then we say that x is different from y. We assume that all elementary descriptors in $\mathcal{D}$ are always different.

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

2. Atomic descriptors

Every product of all elementary descriptors in $\mathcal{D}$ with or without negation

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

where $i_j = 0$ or 1, $k = \bar{\bar{X}}$,

and $x_j^0 = x_j$, $x_j^1 = \sim x_j$

will be called atomic descriptor in $\mathcal{D}$.

Of course for every $\mathcal{D}$ there are at most $2^{\bar{\bar{X}}}$ different atomic descriptors in $\mathcal{D}$.

If x is an atomic descriptor in $\mathcal{D}$ then $\Psi(x)$ is called atom in $\mathcal{D}$.

Descriptors $x, y \in X^*$ are said to be independent iff

$$\Psi(x) \cap \Psi(y) = \underline{\bar{\Phi}}.$$

Theorem 2. Every two different atomic descriptors in $\mathcal{D}$ are independent.

Theorem 3.

$$\bigcup_{x \in \bar{X}_{\mathcal{D}}} \Psi(x) = A_{\mathcal{D}}$$

where $\bar{X}_{\mathcal{D}}$ denotes the set of all atomic descriptors in $\mathcal{D}$.

Theorem 4. Every elementary descriptor $x \in X_{\mathcal{D}}$ may be represented as

$$x = x_1 \vee x_2 \vee \ldots \vee x_n, \quad x_i \in \bar{X}_{\mathcal{D}},$$

where $x_1, x_2, \ldots, x_n$ are all atomic descriptors in $\mathcal{D}$ containing x.

Theorem 5. Every descriptor $x \in X^*_{\mathcal{D}}$ may be represented as the sum of some atomic descriptors in $\mathcal{D}$.

By means of the theorem 5 we are able to represent descriptors in some standard (normal) form.

3. Remarks on implementation

Given some set of objects A (for example books, papers, documents, etc.) and a set of elementary descriptors X (for example, authors´ names, languages, key words, etc.) thus the relation δ is defined. Now we may ask about some sets of objects defined by any compound descriptor $x \in X^*_{\mathcal{D}}$.

The set $B \in A_{\mathcal{D}}$ will be called <u>descriptive</u> in $\mathcal{D}$ iff there exists $x \in X^*_{\mathcal{D}}$ such that

$$\Psi(x) = \mathcal{B}$$

Let $\mathcal{D}(A)$ denote the set of all descriptive sets in $\mathcal{D}$.

From theorem 1 follows that there is only a finite number of descriptive sets in any $\mathcal{D}$. So we are unable to "describe" (in the sense of this paper) all subsets of 2^A in $\mathcal{D}$. By theorem 5 it follows that only the sets which are sums of atoms are descriptive in $\mathcal{D}$.

This result leads to very simple computer implementations of any information retrieval system:

given any descriptor $x \in X^*_{\mathcal{D}}$ by means of theorem 4 we can represent it in a normal form, and then find out the corresponding atoms and form its sum.

This method of searching for sets of documents satisfying some conditions leads to very quick, efficient and simple computer algorithms. Moreover, in the case when the set of elementary descriptors should be extended, the whole system may be easily extended too without destroying and rebuilding already working system.

Many thanks due to Dr. A. Mazurkiewicz for valuable comments.

Literature

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

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

[3] Z. Semadeni: Logical kits, Manuscript, 1971.