POLSKA AKADEMIA NAUK
CENTRUM OBLICZENIOWE

SYSTEMY WYSZUKIWANIA INFORMACJI

Materiały z I Krajowej Konferencji
Jadwisin 14–19 maja 1973 r.

WARSZAWA 1974
PAŃSTWOWE WYDAWNICTWO NAUKOWE

REDAKTOR NAUKOWY
CENTRUM OBLICZENIOWEGO PAN
Zdzisław Pawlak

REDAKTOR
Mirosław Dąbrowski

Okładkę projektowała
Alicja Szubert-Olszewska

REDAKTOR WYDAWNICZY
CENTRUM OBLICZENIOWEGO PAN
Jan Lipski

Printed in Poland

Państwowe Wydawnictwo Naukowe
Oddział w Łodzi 1974

Wydanie I. Nakład 1110+90 egz. Ark. wyd. 13.00. Ark. druk. 12.75
Papier offsetowy, kl. III. 80 g. 70×100. Oddano do składania w lutym
1974 r. Podpisano do druku w lipcu 1974 r. Druk ukończono w lipcu
1974 r. Zam. 60/74. Cena zł 39,— C-13.

Zakład Graficzny Wydawnictw Naukowych
Łódź, ul. Żwirki 2

Zdzisław PAWLAK
Centrum Obliczeniowe PAN

MATEMATYCZNE PODSTAWY SYSTEMÓW WYSZUKIWANIA INFORMACJI

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, 3].

1. SYSTEM DESKRYPTOROWY

Przez system deskryptorowy rozumieć będziemy trójkę

$$D = \langle A_D, X_D, \delta_D \rangle$$

(lub krócej $D = \langle A, X, \delta \rangle$), gdzie

A - jest skończonym (lub nieskończonym) zbiorem; elementy zbioru A nazywać będziemy - obiektami systemu D;

X - jest skończonym zbiorem symboli; elementy zbioru X są deskryptorami elementarnymi systemu D;

$\delta \subseteq A \times X$ - jest relacją binarną zwaną opisem relacyjnym (lub opisem) w systemie D.

Relacja δ może być zastąpiona przez funkcję

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

taką, że:

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

Niech X^* oznacza najmniejszy zbiór zawierający X i taki, że jeżeli $x, y \in X^*$, to $x \wedge y, x \vee y, \sim x$ również należą do X, i niech ψ^* będzie funkcją zdefiniowaną następująco:

$$\psi^*(x) = \psi(x) \quad \text{jeżeli} \quad x \in X$$

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

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

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

gdzie Ψ^* jest rozszerzeniem funkcji Ψ (znak gwiazdki będzie w dalszej części artykułu pomijany).

Niech $x,y \in X^*$. Będziemy mówić, że deskryptor x jest równy deskryptorowi y ($x \equiv y$) wtedy i tylko wtedy, gdy

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

w przeciwnym przypadku deskryptory x i y są różne. Jeżeli $x \equiv y$ nie zachodzi, to będziemy mówić, że x jest różne od y. Zakładamy, że wszystkie deskryptory elementarne w systemie D są różne.

Twierdzenie 1

Dla każdego systemu deskryptorowego $D = \langle A, X\delta \rangle$, liczba różnych deskryptorów jest skończona i nie większa niż $2^{2^{\overline{\overline{X}}}}$.

2. DESKRYPTORY ATOMOWE

Iloczyn logiczny wszystkich deskryptorów elementarnych (zaprzeczonych lub nie zaprzeczonych) postaci:

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

gdzie $i_j = o$ lub 1, $k = \overline{\overline{X}}$ i $x_j^0 = x_j$, $x_j^1 = \sim x_j$, nazywać będziemy <u>deskryptorem atomowym</u> w systemie D. Oczywiście każdy system deskryptorowy D zawiera nie więcej niż $2^{\overline{\overline{X}}}$ różnych deskryptorów atomowych.

Dla deskryptora atomowego x systemu D wartość $\psi(x)$ nazywać będziemy <u>atomem</u> w D. Mówimy, że deskryptory $x,y \in X^*$ są niezależne wtedy i tylko wtedy, gdy

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

Twierdzenie 2

Każde dwa różne deskryptory atomowe w systemie D są niezależne.

Twierdzenie 3

$$\bigcup_{x \in \overline{X}_D} \psi(x) = A_D,$$

gdzie $\overline{X}_D$ oznacza zbiór wszystkich deskryptorów atomowych w systemie D.

Twierdzenie 4

Każdy deskryptor elementarny x $X \in X_D$ może być przedstawiony jako:

$$x = x_1 \vee x_2 \vee \ldots \vee x_n, \qquad x_i \in \overline{X}_D,$$

gdzie $x_1, x_2, \ldots, x_n$ są to wszystkie deskryptory atomowe w systemie D zawierające deskryptor x.

Twierdzenie 5

Każdy deskryptor $x \in X_D^*$ może być przedstawiony jako su
rów atomowych w D.

Przy pomocy twierdzenia 5 możemy deskryptory przedsta
standardowej (normalnej) postaci.

3. UWAGI O REALIZACJI MASZYNOWEJ

Dany jest pewien zbiór obiektów A (np. książki, artykuł
etc.) i zbiór deskryptorów elementarnych X (np. nazwiska a
publikacji, słowa kluczowe etc.) oraz relacja δ. Teraz może
ne podzbiory obiektów zdefiniowane za pomocą deskryptora x

O zbiorze $B \in A_D$ będziemy mówili, że jest opisywalny w
ko wtedy, gdy istnieje deskryptor $x \in X_D^*$ taki, że

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

Niech D(A) oznacza zbiór wszystkich opisywalnych w D z
dzenia 1 wynika, że w każdym systemie D istnieje skończon
rów opisywalnych. Tak więc nie jesteśmy w stanie "opisać"
(w sensie tego artykułu) wszystkich podzbiorów zbioru 2^A.

Z twierdzenia 5 wynika, że tylko zbiory będące sumą ato
walne w systemie D. To prowadzi do bardzo prostej maszyno
systemu wyszukiwania:

Dowolny deskryptor $x \in X_D^*$ można, na mocy twierdzenia
w postaci normalnej, a następnie wyznaczyć odpowiednie atom

Opisana metoda podaje szybki, prosty i efektywny algoryt
nia informacji. Ponadto system wyszukiwania oparty o tę metc
adaptować w przypadku, gdy zbiór deskryptorów elementarnyc

LITERATURA

[1] MOSTOWSKI A., KURATOWSKI K.: Teoria mnogości. PWN 1966.

[2] PAWLAK Z.: About the meaning of personal pronouns (to app
de Linguistique Theoretique et Appliquée, Vol. X, 1973, N

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