

U N I W E R S Y T E T W A R S Z A W S K I

Wydział Matematyki i Mechaniki

Instytut Informatyki

Zygmunt M a z u r

Własności i zastosowania modelu systemu
wyszukiwania informacji opartego na
tezaurusie z wagami

Praca doktorska
wykonana pod kierunkiem
prof.dr hab.
Władysława M. TURSKIEGO

W A R S Z A W A 1977

S p i s t r e ś c i

Wstęp	1
Rozdział 1. Model systemu wyszukiwania informacji	
1.1. Opis modelu systemu wyszukiwania informacji	5
1.1.1. Tezaurus	5
1.1.2. Opisy dokumentów	6
1.1.3. Kwerendy i system wyszukiwania informacji	7
1.1.4. Podobieństwo kwerend i opisów dokumentów	7
1.2. Własności systemu wyszukiwania informacji	14
1.2.1. Własności odpowiedzi systemu	14
1.2.2. Inkluzyjność systemu	16
1.2.3. Dalsze własności systemu wyszukiwania informacji	17
1.3. Geometryczny model systemu wyszukiwania informacji	23
1.3.1. Oznaczenia i definicje	23
1.3.2. Konstrukcja przestrzeni dokumentów PD.	25
1.3.3. Własności przestrzeni dokumentów	30
1.4. Własności list inwersyjnych	32
1.5. Wnioski	43
Rozdział 2. Modyfikacja przestrzeni dokumentów	
2.1. Cele modyfikacji	45
2.2. Sposoby modyfikacji przestrzeni dokumentów	48
2.3. Wnioski i zalecenia	55

Rozdział 3.	Symulacja procesu wyszukiwania informacji	
3.1.	Realizacja maszynowa modelu systemu	59
3.2.	Symulacja ocen o relewantności dokumentów	66
3.3.	Uwagi o przyjętych parametrach modelu	69
Rozdział 4.	Analiza wyników symulacji pracy systemu i wyszukiwania informacji.	
4.1.	Sposoby obliczania współczynników korelacji	70
4.1.1.	Miara podobieństwa odpowiedzi	72
4.1.2.	Weryfikacja różnicy współczynników przed i po adaptacji	75
4.1.3.	Współczynnik korelacji Pearsona i cosinusowej	77
4.1.4.	Weryfikacja różnicy współczynników korelacji Pearsona i cosinusowej	86
4.2.	Zależność współczynników korelacji od liczby kwerend biorących udział w modyfikacji	88
4.3.	Czas wyszukiwania dokumentów	93
4.4.	Wnioski	99
Rozdział 5.	Opis metody zmniejszenia redundancji dokumentów	
5.1.	Zmniejszenie redundancji a efektywność systemu	102
5.2.	Własności rozkładu beta	105
5.3.	Algorytm wyboru wartości progowych	107
5.4.	Wnioski	111
Zakończenie		112
Literatura		113
Dodatek (treści programów)		115

W s t ę p

Systemy przechowywania i wyszukiwania informacji to dziedzina, której znaczenie stale wzrasta w dziale zastosowań maszyn cyfrowych. Jej istotą jest duża ilość danych a nie złożoność operacji, którym są poddawane dane (jak to ma miejsce w przypadku zastosowań numerycznych). Automatyzacja systemów informacyjnych pozwala usprawnić i przyspieszyć proces przepływu informacji równocześnie odciążając dokumentalistów od wielu pracochłonnych operacji.

W większości automatycznych systemów wyszukiwania stosuje się statyczną przestrzeń dokumentów w tym sensie, że dokumentom odpowiadają stałe pozycje w przestrzeni. Statyczne przestrzenie dokumentów indukują systemy klasyfikacji, które nie zawsze odpowiadają potrzebom użytkowników. Stopień zaspokojenia wymagań użytkowników przy posługiwaniu się takim systemem jest bezpośrednio zależny od tego, jak bardzo jego potrzeby są zgodne z kategoriami tematycznymi klasyfikacji.

Aby móc zaspokajać wymagania użytkowników będziemy stosować dynamiczną przestrzeń dokumentów, w której dokumenty będą mogły zmieniać swoje pozycje w przestrzeni zgodnie z potrzebami użytkowników.

W niniejszej pracy do badania systemów z dynamiczną przestrzenią dokumentów zastosowano model systemu wyszukiwania informacji oparty na tezaurusie z wagami, który jest uogólnieniem modeli W.M. Turskiego [7] i M. Dąbrowskiego [1].

Położenie dokumentu w przestrzeni określa jego opis, który otrzymujemy w wyniku klasyfikacji dokumentu. Klasyfikacja (opisywanie) dokumentu polega na wybraniu podzbioru deskryptorów

najdokładniej charakteryzujących klasyfikowany dokument. Wszystkie deskryptory w opisie dokumentu występują z wagami. Waga jest miernikiem ważności danego deskryptora w opisie. Wprowadzenie wag deskryptorów pozwala na precyzyjne wyspecyfikowanie treści dokumentu za pomocą niedużej liczby deskryptorów występujących w jego opisie.

Modelem systemu wyszukiwania informacji opartym na teza-
urysie z wagami nazywamy czwórką $S = (T, D, R, \delta)$, gdzie
 T jest tezauresem (zbiorem terminów używanych do klasyfi-
kowania dokumentów), D - zbiorem dokumentów, R - zbiorem
kwerend (pytań stawianych systemowi) a δ odwzorowaniem
 $R \times [0,1] \rightarrow 2^{D \times [0,1]}$, przyporządkowującym kwerendom z
zadanymi dokładnościami zbiór par postaci (d, v) , gdzie ele-
ment d należy do podzbioru zbioru dokumentów a v jest miarą
podobieństwa kwerendy do elementu d .

W rozdziale pierwszym pracy pokazano, że wprowadzenie
wag do opisów dokumentów nie zmienia podstawowych własności
systemu opisanego w pracy [7]. Podstawową własnością jest inklu-
zyjność systemu pozwalająca na tworzenie łańcuchów odpowiedzi
na łańcuchy kwerend, tj. na poszukiwanie odpowiedzi na "ściślejszą"
kwerendę wśród odpowiedzi uzyskanych na "mniej precyzyjną"
kwerendę. W dalszej części tego rozdziału zdefiniowane są listy
inwersyjne oraz zbadane ich własności. Zastosowane tutaj listy
inwersyjne różnią się od powszechnie znanych tym, że adresy
(numery) dokumentów występują w nich razem z wagami. W czasie
wyszukiwania dokumentów adresy uporządkowane ze względu na wagi
pozwalają na bardzo proste i szybkie wyszukiwanie dokumentów
należących do odpowiedzi z wymaganą dokładnością. Udowodniono

twierdzenie mówiące o tym, że wyszukiwanie przy pomocy list inwersyjnych zdefiniowanych w rozdziale 1.4 zachowuje własności modelu podane w rozdziałach 1.1 i 1.2.

Stała modyfikacja opisów dokumentów prowadzi do zwiększenia efektywności systemu, szczególnie z punktu widzenia użytkowników.

Metody modyfikacji opisów polegają na wprowadzaniu nowych pojęć do opisów, usuwaniu zbędnych pojęć oraz zmianie wag deskryptorów, mogą zapewnić stałą aktualizację słownictwa i wag w tych opisach.

Aktualizacja wag pojęć użytych do opisu dokumentu na podstawie subiektywnej opinii użytkowników o trafności odpowiedzi pozwala utworzyć opisy lepiej odzwierciedlające rzeczywistą treść dokumentów.

W rozdziale drugim omówione są sposoby modyfikacji opisów dokumentów na podstawie ocen użytkowników o relewantności dokumentów należących do odpowiedzi na zadaną kwerendę. W wyniku modyfikacji opisy są "upodabniane" do kwerend. W czasie pracy systemu gromadzone są informacje od użytkowników w postaci modyfikacji opisów aby po pewnym czasie zreorganizować listy inwersyjne w celu przystosowania ich do aktualnych opisów dokumentów.

W rozdziale trzecim omówiono sposób symulacji pracy systemu wyszukiwania informacji na maszynie cyfrowej CDC 6400.

W rozdziale czwartym przedstawiono wyniki symulacji pracy systemu. Z otrzymanych wyników wypływa wniosek, że z bieżącej modyfikacji przestrzeni dokumentów korzyści czerpią nie tylko ci użytkownicy, którzy brali udział w modyfikowaniu opisów dokumentów. Nowi użytkownicy korzystają z "doświadczeń" zebranych

nych podczas wyszukiwań poprzednich.

W rozdziale piątym podany jest opis metody zmniejszającej redundancję opisów dokumentów. W metodzie tej wykorzystuje się własności list inwersyjnych i zbadane potrzeby użytkowników. W czasie pracy systemu gromadzone są informacje o wagach deskryptorów w kwerendach i wymaganych dokładnościach odpowiedzi. W czasie reorganizacji, na podstawie kryterium statystycznego usuwa się z list inwersyjnych adresy dokumentów z wagami mniejszymi od ustalonych wartości progowych.

Pragnę wyrazić moją wdzięczność Profesorowi
doktorowi Władysławowi M. Turskiemu za cenne uwagi oraz
stałą pomoc i zachętę w trakcie pisania pracy.

Rozdział 1. Model systemu wyszukiwania informacji.

1.1. Opis modelu systemu wyszukiwania informacji.

W rozdziale 1 podajemy opis modelu systemu wyszukiwania informacji oparty na tezaursie. Zostały rozszerzone definicje procesu klasyfikacji (opisu dokumentów i kwerend) podane w pracach [7], [9]. W opisach dokumentów i kwerendach wykorzystujemy deskryptory z wagami. Podstawowe własności systemu podanego w [7], [9] zostały zachowane w opisywanym systemie.

1.1.1. Tezaurus

Tezauresem nazywamy niepusty, skończony zbiór terminów T , dla którego spełnione są następujące warunki:

a) istnieje niepusty podzbiór $T_0 \subset T$, nazywany zbiorem deskryptorów,

b) określona jest symetryczna, przechodnia i zwrotna relacja $\mathcal{S} \subset T \times T$ taka, że

$$t_1 \neq t_2 \wedge t_1 \mathcal{S} t_2 \implies (t_1 \in T \setminus T_0) \vee (t_2 \in T \setminus T_0)$$

$$t_1 \in T \setminus T_0 \implies (\exists t \in T_0) (t \mathcal{S} t_1)$$

relację $\mathcal{S}$ nazywamy relacją synonimii, a terminy t_1, t_2 będące w tej relacji - terminami synonimicznymi.

c) określona jest zwrotna, przechodnia i antysymetryczna relacja $R \subset T_0 \times T_0$, którą nazywamy relacją uogólnienia. Jeśli dwa deskryptory t_1, t_2 są w relacji $t_1 R t_2$ to mówimy, że deskryptor t_1 jest ogólniejszy niż deskryptor t_2 , albo że deskryptor t_2 jest bardziej szczegółowy niż t_1 .

Elementy ze zbioru $T \setminus T_0$, mające znaczenie pomocnicze, nazywamy askryptorami. Z podanej definicji tezaursu wynika, że dla każdego askryptora istnieje dokładnie jeden synonimiczny deskryptor.

1.1.2. Opisy dokumentów

W wyniku klasyfikacji powstaje opis dokumentu

$$\mathcal{D} = \{(t_1, w_1), (t_2, w_2), \dots, (t_k, w_k)\} \quad (1.1.1)$$

spełniający warunki:

$$\begin{aligned} \text{a) } t' \neq t'' \wedge t' R t'' & \implies \\ & \implies (\forall w' \in [0, 1]) (t', w') \notin \mathcal{D} \vee (\forall w'' \in [0, 1]) (t'', w'') \notin \mathcal{D} \end{aligned} \quad (1.1.2)$$

$$\text{b) } \sum_{i=1}^k w_i = 1 \quad (1.1.3)$$

Wagi mówią o rozkładzie informacji na poszczególne deskryptory. Jeśli informację zawartą w opisie dokumentu uważać będziemy za jednostkę to waga w_i będzie mówić jaka część tej informacji przypada na deskryptor t_i .

Powiemy, że opis dokumentu $\mathcal{D}_1 = \{(t_{11}, w_{11}), (t_{12}, w_{12}), \dots, (t_{1l_1}, w_{1l_1})\}$ zawiera się w opisie innego dokumentu

$$\mathcal{D}_2 = \{(t_{21}, w_{21}), (t_{22}, w_{22}), \dots, (t_{2l_2}, w_{2l_2})\}$$

jeśli

$$(\forall i)(1 \leq i \leq l_1)(\exists j)(1 \leq j \leq l_2) (t_{1i} R t_{2j}) \quad (1.1.4)$$

Jeśli $\mathcal{D}_1$ i $\mathcal{D}_2$ spełniają warunek (1.1.4) to będziemy pisać $\mathcal{D}_1 \leq \mathcal{D}_2$. Relacja $\leq$ indukuje w zbiorze opisów dokumentów porządek częściowy.

1.1.3. Kwerendy i system wyszukiwania informacji.

Kwerendą nazywamy skończony ciąg deskryptorów z sumą wag równą jedności nie zawierający powtórzeń i uogólnień występujących deskryptorów. Kwerenda ma więc taką samą postać jak opisy dokumentów, tj. (1.1.1) z warunkami (1.1.2) i (1.1.3). Relacja $\leq$ jest zdefiniowana dla kwerend i par (opis, kwerenda) oraz (kwerenda, opis) tak samo jak dla opisów. Dla wygody będziemy czasem elementy kwerendy oznaczać literą r ze wskaźnikami.

Systemem wyszukiwania informacji opartym na tezaurusie nazwiemy czwórkę $S = (T, D, R, \sigma)$, gdzie T jest tezaurusem, D - zbiorem dokumentów, R zbiorem kwerend, σ odwzorowaniem $R \times [0,1] \rightarrow 2^{D \times [0,1]}$, przyporządkowującym kwerendom z zadanymi dokładnościami zbiór par postaci (d, v) , gdzie element d należy do podzbioru zbioru dokumentów a v jest miarą podobieństwa kwerendy do elementu d .

W systemie S będziemy używać miary podobieństwa pomiędzy kwerendą $\mathcal{P}$ a opisem dokumentu $\mathcal{B}$. Odpowiedzią systemu S na kwerendę $\mathcal{P} \in R$ nazywać będziemy $Q = \sigma(\mathcal{P}, c) \subset D \times [0,1]$.

1.1.4. Podobieństwo kwerend i opisów dokumentów.

Definicja 1.1.1.

Niech $\mathcal{B}_1, \mathcal{B}_2$ będą opisami postaci:

$$\mathcal{B}_1 = \{(t_{11}, w_{11}), (t_{12}, w_{12}), \dots, (t_{1n_1}, w_{1n_1})\} \quad (1.1.5)$$

$$\mathcal{B}_2 = \{(t_{21}, w_{21}), (t_{22}, w_{22}), \dots, (t_{2n_2}, w_{2n_2})\} \quad (1.1.6)$$

Dla dowolnych opisów dokumentów $\mathcal{B}_1, \mathcal{B}_2$ wprowadzamy miarę

podobieństwa określoną w następujący sposób:

$$\begin{aligned} \mu(t_1, t_2) = & \sum_{t_{1i} \in L_1} \min(w_{1i}, \sum_{t_{2j} \in L_1^*(t_{1i})} w_{2j}) \\ & + \sum_{t_{2k} \in L_2} \min(w_{2k}, \sum_{t_{1l} \in L_2^*(t_{2k})} w_{1l}) \quad (1.1.7) \end{aligned}$$

gdzie:

$$L_1 = \{t_{1i} : (\exists t_{2j}) t_{1i} R t_{2j} \wedge 1 \leq i \leq n_1 \wedge 1 \leq j \leq n_2\}$$

$$L_2 = \{t_{2k} : (\exists t_{1l}) t_{2k} R t_{1l} \wedge t_{2k} \neq t_{1l} \wedge 1 \leq k \leq n_2 \wedge 1 \leq l \leq n_1\}$$

$$L_1^*(t_{1i}) = \{t_{2j} : t_{1i} R t_{2j} \wedge 1 \leq j \leq n_2\} \quad \text{dla } 1 \leq i \leq n_1$$

$$L_2^*(t_{2k}) = \{t_{1l} : t_{2k} R t_{1l} \wedge 1 \leq l \leq n_1\} \quad \text{dla } 1 \leq k \leq n_2$$

Zbiór $L_1(L_2)$ jest zbiorem deskryptorów z opisu $t_1(t_2)$, które posiadają deskryptory bardziej szczegółowe w opisie $t_2(t_1)$ i różne od deskryptorów z opisu t_2 .

Dla danego deskryptora ze zbioru $L_1(L_2)$ zbiór $L_1^*(L_2^*)$ jest zbiorem wszystkich deskryptorów ściślejszych od tego deskryptora występujących w opisie $t_2(t_1)$.

Przykład 1.1.1.

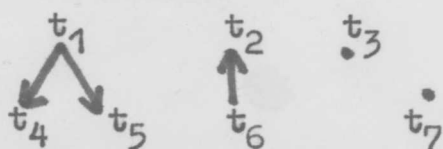
Niech t_1, t_2 będą opisami

$$t_1 = \{(t_1, 0.5), (t_2, 0.4), (t_3, 0.1)\}$$

$$t_2 = \{(t_4, 0.2), (t_5, 0.1), (t_6, 0.3), (t_7, 0.4)\}$$

oraz $t_1 R t_4, t_1 R t_5, t_6 R t_2$.

Relację R pomiędzy deskryptorami można graficznie przedstawić następująco :



$$L_1 = \{t_1\}, \quad L_1^*(t_1) = \{t_4, t_5\}, \quad L_2 = \{t_6\}, \quad L_2^*(t_6) = \{t_2\}$$

$$\mu(t_1, t_2) = \min(0.5, 0.2 + 0.1) + \min(0.3, 0.4) = \\ = 0.3 + 0.3 = 0.6$$

Twierdzenie 1.1.1.

Jeśli $t_{1i} \in L_1 \wedge t_{2j} \in L_1^*(t_{1i})$ to $t_{2j} \notin L_2$

Dowód

Weźmy dowolny element $t_{2j} \in L_1^*(t_{1i})$ i założmy niewprost, że $t_{2j} \in L_2$. Na mocy definicji zbioru L_2 otrzymujemy, że $(\exists t_{1l}) t_{2j} R t_{1l}$. Korzystając z przechodniości relacji R , $t_{1i} R t_{2j} \wedge t_{2j} R t_{1l} \implies t_{1i} R t_{1l}$, znaczy to, że w opisie dokumentu t_1 deskryptor t_{1i} jest uogólnieniem deskryptora t_{1l} co daje sprzeczność z warunkiem (1.1.2).

Twierdzenie 1.1.1 można sformułować w innej postaci.

Twierdzenie 1.1.2.

Dla dowolnych opisów dokumentów t_1, t_2 postaci (1.1.5), (1.1.6) spełniony jest warunek:

$$\sim \exists (t_2, w) \in t_2 \quad (\exists (t_{1i}, w') \in t_1 \quad \exists (t_{2j}, w'') \in t_2 \quad t_{1i} R t_{2j} \wedge t_{2j} R t_{1j})$$

Dowód wynika z warunku (1.1.2).

Twierdzenie 1.1.3.

Miara podobieństwa opisów dokumentów spełnia warunek

$$0 \leq \mu(t_1, t_2) \leq 1$$

Dowód

Korzystając z nierówności:

$$\min(w_{1i}, \sum_{t_{2j} \in L_1^*(t_{1i})} w_{2j}) \leq w_{1i}$$

$$\min(w_{2k}, \sum_{t_{1l} \in L_2^*(t_{2k})} w_{1l}) \leq \sum_{t_{1l} \in L_2^*(t_{2k})} w_{1l}$$

oraz ze wzoru (1.1.7) otrzymujemy

$$\mu(t_1, t_2) \leq \sum_{t_{1i} \in L_1} w_{1i} + \sum_{t_{2k} \in L_2} \sum_{t_{1l} \in L_2^*(t_{2k})} w_{1l}$$

Ponieważ wagi w_{1i} i w_{1l} w opisie t_1 wstępują z różnymi deskryptorami (wynika to z twierdzenia 1.1.1), możemy napisać, że prawa część ostatniej nierówności jest nie większa od jedności. Stąd otrzymujemy, że $\mu(t_1, t_2) \leq 1$. Ponieważ wagi w opisach są nieujemne otrzymujemy, że $\mu(t_1, t_2) \geq 0$.

Z twierdzenia 1.1.3 i definicji 1.1.1 wynika, że miara podobieństwa pomiędzy opisami t_1, t_2 jest równa jedności jeśli każdy deskryptor z opisu t_1 jest uogólnieniem lub uściśleniem dowolnego deskryptora z t_2 i na odwrót oraz we wzorze (1.1.7) argumenty funkcji minimum są równe.

Przykład 1.1.2.

$$\text{Niech } t_1 = \{(t_1, 0.6), (t_2, 0.4)\}$$

$$t_2 = \{(t_3, 0.6), (t_4, 0.4)\}$$

$$t'_1 = \{(t_1, 0.9), (t_2, 0.1)\}$$

$$t'_2 = \{(t_3, 0.9), (t_2, 0.1)\}$$

$$\text{oraz } t_1 R t_3 \quad \text{i} \quad t_2 R t_4$$

$$\text{wtedy } \mu(t_1, t_2) = 1 \quad \text{i} \quad \mu(t'_1, t'_2) = 1.$$

Miara podobieństwa μ jest zdefiniowana dla kwerend i par (opis, kwerend) oraz (kwerenda, opis) tak samo jak dla opisów.

Twierdzenie 1.1.4.

Jeśli $\mathbb{t}_1 \leq \mathbb{t}_2$ to $L_2 = \emptyset$

Dowód

Oznaczmy składowe opisu $\mathbb{t}_1$ przez (t_{1i}, w_{1i}) zaś składowe opisu $\mathbb{t}_2$ przez (t_{2j}, w_{2j}) .

Z warunków (1.1.2), (1.1.4) otrzymujemy, że

$(\sim \exists j) t_{2j} R t_{1i}$, a więc $L_2 = \emptyset$

Korzystając z twierdzenia 1.1.4 wzór (1.1.7) dla opisów spełniających $\mathbb{t}_1 \leq \mathbb{t}_2$ można zapisać w postaci:

$$\mu(\mathbb{t}_1, \mathbb{t}_2) = \sum_{t_{1i} \in L_1} \min(w_{1i}, \sum_{t_{2j} \in L_1^*(t_{1i})} w_{2j}) \quad (1.1.8)$$

Definicja 1.1.2.

Niech $\mathbb{t}_1, \mathbb{t}_2$ będą postaci (1.1.5), (1.1.6) i

$\mathbb{t}_1 \leq \mathbb{t}_2$ wtedy

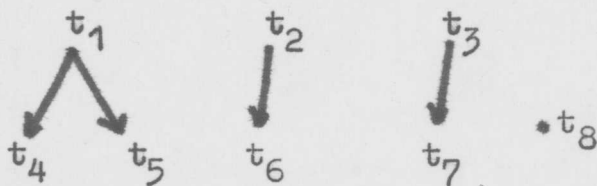
$$\bigvee_{0 \leq c \leq 1} \mathbb{t}_1 \triangleleft_c \mathbb{t}_2 \equiv \bigvee_{\substack{(t_{1i}, w_{1i}) \in \mathbb{t}_1 \\ 1 \leq i \leq n_1}} w_{1i} \cdot c \leq \sum_{t_{2j} \in L_1^*(t_{1i})} w_{2j} \quad (1.1.9)$$

Relację $\triangleleft_c$ nazywać będziemy relacją podobieństwa, gdzie c jest dokładnością relacji podobieństwa. Relacja podobieństwa $\triangleleft_c$ jest zwrotna, antysymetryczna i quasi-przechodnia. Zwrotność i antysymetryczność relacji $\triangleleft_c$ wy-

nika bezpośrednio z definicji relacji $\triangleleft_c$. Quasi - przechodniość relacji zostanie pokazana w twierdzeniu 1.1.6.

Przykład 1.1.3.

Niech $t_1 R t_4$, $t_1 R t_5$, $t_2 R t_6$, $t_3 R t_7$



$$\mathcal{t}_1 = \{(t_1, 0.3), (t_2, 0.4), (t_3, 0.3)\}$$

$$\mathcal{t}_2 = \{(t_4, 0.1), (t_5, 0.1), (t_6, 0.3), (t_7, 0.2), (t_8, 0.3)\}$$

Relacja $\mathcal{t}_1 \triangleleft_{0.5} \mathcal{t}_2$ jest spełniona ponieważ każda nierówność w układzie (1.1.10) jest prawdziwa.

$$\left. \begin{array}{l} 0.3 \times 0.5 \leq 0.1 + 0.1 \\ 0.4 \times 0.5 \leq 0.3 \\ 0.3 \times 0.5 \leq 0.2 \end{array} \right\} \quad (1.1.10)$$

natomiast relacja $\mathcal{t}_1 \triangleleft_{0.8} \mathcal{t}_2$ nie jest spełniona ponieważ poniższe nierówności są fałszywe

$$0.3 \times 0.8 \leq 0.1 + 0.1$$

$$0.4 \times 0.8 \leq 0.3$$

$$0.3 \times 0.8 \leq 0.2$$

Twierdzenie 1.1.5.

Jeśli $\mathcal{t}_1 \triangleleft_{c_1} \mathcal{t}_2 \wedge c_1 \geq c_2$ to $\mathcal{t}_1 \triangleleft_{c_2} \mathcal{t}_2$

Dowód

Z definicji 1.1.2 wynika

$$t_1 \triangleleft_{c_1} t_2 \Rightarrow t_1 \leq t_2 \wedge \bigvee_{(t_{1i}, w_{1i}) \in t_1} w_{1i} \cdot c_1 \leq \sum_{t_{2j} \in L_1^*(t_{1i})} w_{2j}$$

Ponieważ $c_1 \geq c_2 \Rightarrow w_{1i} \cdot c_2 \leq w_{1i} \cdot c_1$, stąd

$$w_{1i} \cdot c_2 \leq \sum_{t_{2j} \in L_1^*(t_{1i})} w_{2j}$$

Twierdzenie 1.1.6.

Jeśli $t_1 \triangleleft_{c_1} t_2 \wedge t_2 \triangleleft_{c_2} t_3$ to $t_1 \triangleleft_c t_3$

i $c \leq c_1 \cdot c_2$

Dowód

Korzystając z definicji 1.1.2 otrzymujemy,

$$t_1 \triangleleft_{c_1} t_2 \Rightarrow t_1 \leq t_2 \wedge \bigvee_{(t_{1i}, w_{1i}) \in t_1} w_{1i} \cdot c_1 \leq \sum_{t_{2j} \in L_1^*(t_{1i})} w_{2j}$$

$$t_2 \triangleleft_{c_2} t_3 \Rightarrow t_2 \leq t_3 \wedge \bigvee_{(t_{2j}, w_{2j}) \in t_2} w_{2j} \cdot c_2 \leq \sum_{t_{3l} \in L_1^*(t_{2j})} w_{3l}$$

z przechodniości relacji $\leq$ wynika $t_1 \leq t_3$,

z powyższych nierówności otrzymujemy

$$w_{1i} \cdot c_1 \leq \sum_{t_{2j} \in L_1^*(t_{1i})} w_{2j} \leq \sum_{t_{2j} \in L_1^*(t_{1i})} \left(\frac{1}{c_2} \sum_{t_{3l} \in L_1^*(t_{2j})} w_{3l} \right) =$$

$$= \frac{1}{c_2} \sum_{t_{3l} \in L_1^*(t_{1i})} w_{3l}$$

Ostatnia równość wynika ze związku

$$\bigcup_{t_{2j} \in L_1^*(t_{1i})} \{ t_{3l} : t_{3l} \in L_1^*(t_{2j}) \} = \{ t_{3l} : t_{3l} \in L_1^*(t_{1i}) \}$$

W ten sposób otrzymaliśmy, że

$$\bigvee_{(t_{1i}, w_{1i}) \in \mathcal{L}_1} w_{1i} \cdot c_1 \cdot c_2 \leq \sum_{t_{31} \in L_1^*(t_{1i})} w_{31}$$

Na podstawie twierdzenia 1.1.5 iloczyn $c_1 \cdot c_2$ w ostatniej nierówności możemy zastąpić stałą $c \leq c_1 \cdot c_2$

W twierdzeniu 1.1.6 udowodniliśmy quasi - przechodniość relacji podobieństwa $\triangleleft_c$ oraz pokazaliśmy, że jeśli $\mathcal{L}_1 \triangleleft_{c_1} \mathcal{L}_2 \triangleleft_{c_2} \mathcal{L}_3$ to dokładność w relacji $\mathcal{L}_1 \triangleleft_c \mathcal{L}_3$ równa jest iloczynowi $c_1 \cdot c_2$. Relacja podobieństwa $\triangleleft_c$ jest zdefiniowana dla kwerend i par (opis, kwerenda) oraz (kwerenda, opis) tak samo jak dla opisów. Jeżeli relacja podobieństwa $\triangleleft_c$ zachodzi pomiędzy $\mathcal{Q}$ i $\mathcal{T}$ to powiemy, że kwerenda $\mathcal{Q}$ jest podobna do opisu $\mathcal{T}$ z dokładnością c .

Liczbę $v = \mu(\mathcal{Q}, \mathcal{T})$ będziemy nazywać stopniem podobieństwa kwerendy $\mathcal{Q}$ do opisu $\mathcal{T}$.

1.2. Własności systemu wyszukiwania informacji.

1.2.1. Własności odpowiedzi systemu.

Oznaczając przez $\mathcal{T}(d)$ opis dokumentu $d \in D$ uzyskiwany w wyniku klasyfikacji, odpowiedzią systemu na kwerendę $\mathcal{Q} \in \mathcal{R}$ z dokładnością c będziemy nazywać :

$$Q = \mathcal{J}(\mathcal{Q}, c) = \{ (d, v) : d \in D \wedge \mathcal{Q} \triangleleft_c \mathcal{T}(d) \wedge v = \mu(\mathcal{Q}, \mathcal{T}) \} \quad (1.2.1)$$

Podstawową własność odpowiedzi systemu, a mianowicie zależność zachodzącą pomiędzy dokładnością c a stopniem podobieństwa

stwa $\mathcal{V}$ kwerendy do opisu dokumentu należącego do odpowiedzi można wyrazić w postaci następującego twierdzenia

Twierdzenie 1.2.1,

Jeśli $(d, v) \in \delta(\mathcal{V}, c)$ to $v \geq c$

Dowód wynika bezpośrednio z definicji 1.1.2 relacji podobieństwa $\triangleleft_c$ oraz ze wzoru (1.1.8)

Twierdzenie odwrotne nie jest prawdziwe, tzn. jeśli $v \geq c$ to nie zawsze $(d, v) \in \delta(\mathcal{V}, c)$

Przykład 1.2.1.

Niech $r_1 R t_1$, $r_2 R t_2$, $c = 0.5$

oraz $\mathcal{V} = \{(r_1, 0.4), (r_2, 0.6)\}$

$\mathcal{U} = \{(t_1, 0.5), (t_2, 0.2), (t_3, 0.3)\}$,

stopień podobieństwa $\mathcal{V}$ do opisu $\mathcal{U}$

$v = \mu(\mathcal{V}, \mathcal{U}) = \min(0.4, 0.5) + \min(0.6, 0.2) = 0.6$

Mimo, że $v > c$, opis dokumentu $\mathcal{U}$ nie należy do odpowiedzi na kwerendę $\mathcal{V}$ ponieważ nie jest spełniona nierówność $0.6 \times 0.5 \leq 0.2$ wynikająca z definicji 1.1.2.

Twierdzenie 1.2.2.

Jeśli $c_1 \geq c_2$ to $\delta(\mathcal{V}, c_1) \subset \delta(\mathcal{V}, c_2)$

Dowód

Niech $(d, v) \in \delta(\mathcal{V}, c_1)$, stąd $\mathcal{V} \triangleleft_{c_1} \mathcal{U}(d)$

Korzystając z twierdzenia 1.1.5 otrzymujemy, że

$\mathcal{V} \triangleleft_{c_2} \mathcal{U}(d)$ czyli $(d, v) \in \delta(\mathcal{V}, c_2)$

1.2.2. Inkluzyjność systemu.

Inkluzyjnością systemu nazywamy własność pozwalającą na tworzenie łańcuchów odpowiedzi na łańcuchy kwerend, tj. na poszukiwanie odpowiedzi na "ściślejszą" kwerendę wśród odpowiedzi uzyskanych na "mniej precyzyjną" kwerendę. W rozaptrywanym systemie, warunek inkluzyjności systemu można sformułować w postaci twierdzenia.

Twierdzenie 1.2.3.

Jeśli odpowiedzi systemu na kwerendy $\mathcal{V}$ są postaci (1.2.1) to system jest inkluzyjny, czyli

$$\{\mathcal{V}_1, \mathcal{V}_2\} \subset \mathbb{R} \wedge \mathcal{V}_1 \triangleleft_c \mathcal{V}_2 \wedge c_1 \leq c_2 \cdot c \implies \implies \mathcal{J}(\mathcal{V}_1, c_1) \supset \mathcal{J}(\mathcal{V}_2, c_2) \quad (1.2.2)$$

Dowód.

Wykażemy najpierw, że dla dowolnych dwu kwerend

$\mathcal{V}_1 \triangleleft_c \mathcal{V}_2$ zachodzi $\mathcal{Z}(\mathcal{V}_1, c_1) \supset \mathcal{Z}(\mathcal{V}_2, c_2)$, gdzie $\mathcal{Z}(\mathcal{V}, c') = \{t : \exists_{d \in D} \mathcal{V} \triangleleft_{c'} t(d)\}$ jest zbiorem tych opisów, które zawierają kwerendę z dokładnością c' . Załóżmy, że $t \in \mathcal{Z}(\mathcal{V}_2, c_2)$, tj. $\mathcal{V}_2 \triangleleft_{c_2} t$; ponieważ $\mathcal{V}_1 \triangleleft_c \mathcal{V}_2$ i relacja $\triangleleft$ jest ^{quasi}przechodnia oraz $c_1 \leq c_2 \cdot c$ to $\mathcal{V}_1 \triangleleft_{c_1} t$, a więc $t \in \mathcal{Z}(\mathcal{V}_1, c_1)$. Wynika stąd, że $t \in \mathcal{Z}(\mathcal{V}_2, c_2) \implies t \in \mathcal{Z}(\mathcal{V}_1, c_1)$, czyli $\mathcal{Z}(\mathcal{V}_2, c_2) \subset \mathcal{Z}(\mathcal{V}_1, c_1)$.

Pozostaje do udowodnienia, że $\mathcal{Z}(\mathcal{V}_1, c_1) \supset \mathcal{Z}(\mathcal{V}_2, c_2) \implies \implies \mathcal{J}(\mathcal{V}_1, c_1) \supset \mathcal{J}(\mathcal{V}_2, c_2)$. Zgodnie z założeniem

$\delta(\mathcal{P}, c') = \{(d, v) : d \in D \wedge \mathcal{P} \triangleleft_{c'} \mathcal{B}(d) \wedge v =$
 $= \mu(\mathcal{P}, \mathcal{B})\} = \{(d, v) : d \in D \wedge \mathcal{B}(d) \in \tau(\mathcal{P}, c') \wedge v =$
 $= \mu(\mathcal{P}, \mathcal{B})\}.$ Niech $(d', v_2) \in \delta(\mathcal{P}_2, c_2)$, tj. $d' \in D \wedge$
 $\wedge \mathcal{B}(d') \in \tau(\mathcal{P}_2, c_2) \wedge v_2 = \mu(\mathcal{P}_2, \mathcal{B}(d'))$, ale ponieważ
 $\tau(\mathcal{P}_2, c_2) \subset \tau(\mathcal{P}_1, c_1)$ to $\mathcal{B}(d') \in \tau(\mathcal{P}_1, c_1)$ i $(d', v_1) \in$
 $\in \delta(\mathcal{P}_1, c_1)$, gdzie $v_1 = \mu(\mathcal{P}_1, \mathcal{B}(d'))$

Z udowodnionego twierdzenia wypływa ważny wniosek :
 w systemie wyszukiwania informacji opartym na tezaurusie i
 z wagami deskryptorów, monotonicznemu ciągowi kwerend odpowia-
 da monotoniczny ciąg odpowiedzi, tj. jeśli np.

$$\mathcal{P}_1 \triangleleft_{c_1} \mathcal{P}_2 \triangleleft_{c_2} \dots \triangleleft_{c_{n-1}} \mathcal{P}_n \quad \text{to}$$

$$\delta(\mathcal{P}_1, c) \supset \delta(\mathcal{P}_2, \frac{c}{c_1}) \supset \dots \supset \delta(\mathcal{P}_n, \frac{c}{c_1 \cdot c_2 \cdot \dots \cdot c_{n-1}})$$

Wniosek ten ma poważne znaczenie praktyczne: pozwala on ogra-
 niczyć poszukiwanie odpowiedzi na coraz precyzyjniej sformuło-
 wane kwerendy do odpowiedzi kwerend mniej precyzyjnych. użyt-
 kownik w trakcie konwersacji może precyzować swoje kwerendy
 bądź to przez zastępowanie deskryptorów ogólniejszych przez
 deskryptory ściślejsze, bądź to przez dodawanie nowych deskryp-
 torów, bądź przez zwiększenie dokładności odpowiedzi.

1.2.3. Dalsze własności systemu wyszukiwania informacji.

Niech

$$\mathcal{P}_1 = \{(r_{11}, w_{11}), (r_{12}, w_{12}), \dots, (r_{1m}, w_{1m})\} \quad (1.2.3)$$

$$\mathcal{P}_2 = \{(r_{21}, w_{21}), (r_{22}, w_{22}), \dots, (r_{2n}, w_{2n})\} \quad (1.2.4)$$

będą kwerendami. Oznaczmy przez $\mathcal{P}_0$ zbiór deskryptorów z wagami, do którego zaliczymy:

a) deskryptory występujące jednocześnie w $\mathcal{P}_1$ i $\mathcal{P}_2$ z wagami równymi maksimum wag w $\mathcal{P}_1$ i $\mathcal{P}_2$ tj.

$$\begin{aligned} \exists \\ 1 \leq i \leq m \\ 1 \leq j \leq n \end{aligned} (r_{1i} = r_{2j}) \Rightarrow (r_{1i}, \max(w_{1i}, w_{2j})) \in \mathcal{P}_0$$

b) deskryptory występujące w jednej kwerendzie i będące uogólnieniem deskryptorów drugiej kwerendy z wagami równymi maksimum wag w kwerendach $\mathcal{P}_1$ i $\mathcal{P}_2$, tj.

$$\begin{aligned} \exists \\ 1 \leq i \leq m \\ 1 \leq j \leq n \end{aligned} (r_{1i} R r_{2j}) \Rightarrow (r_{1i}, \max(w_{1i}, \sum_{r_{2j} \in L_1^*(r_{1i})} w_{2j})) \in \mathcal{P}_0$$

oraz

$$\begin{aligned} \exists \\ 1 \leq i \leq m \\ 1 \leq j \leq n \end{aligned} (r_{2j} R r_{1i}) \Rightarrow (r_{2j}, \max(w_{2j}, \sum_{r_{1i} \in L_1^*(r_{2j})} w_{1i})) \in \mathcal{P}_0$$

(Definicja zbiorów $L_1^*(r_{1i})$, $L_2^*(r_{2j})$ została podana w rozdziale 1.1.4)

Do zbioru

$$\mathcal{P}_0 = \{(r_{01}, w_{01}), (r_{02}, w_{02}), \dots, (r_{0k}, w_{0m_1})\}$$

nie wchodzi żadne inne pary (deskryptor, waga).

Przez $\mathcal{P}_1 \cdot \mathcal{P}_2$ oznaczać będziemy zbiór deskryptorów z wagami otrzymany ze zbioru $\mathcal{P}_0$, w którym wszystkie wagi

$$w_{01} \text{ zostały zastąpione przez } w_{01} / \sum_{i=1}^{m_1} w_{0i}.$$

Oznaczmy przez $\mathcal{V}_0$ zbiór deskryptorów z wagami, do którego zaliczymy:

a) deskryptory występujące jednocześnie w $\mathcal{V}_1, \mathcal{V}_2$ z wagami równymi minimum wag w $\mathcal{V}_1$ i $\mathcal{V}_2$, tj.

$$\exists (r_{1i} = r_{2j}) \Rightarrow (r_{1i}, \min(w_{1i}, w_{2j})) \in \mathcal{V}_0$$

$$1 \leq i \leq m$$

$$1 \leq j \leq n$$

b) deskryptory występujące w jednej kwerendzie i będące uściśleniami deskryptora występującego w drugiej kwerendzie z wagami równymi minimum wag w kwerendach $\mathcal{V}_1, \mathcal{V}_2$ mnożonymi przez czynnik normalizujący będący ilorazem odpowiednich wag, tj.

$$\exists (r_{1i} R r_{2j}) \Rightarrow (r_{2j}, \frac{w_{2j}}{v_2} \min(w_{1i}, v_2)) \in \mathcal{V}_0$$

$$1 \leq i \leq m$$

$$1 \leq j \leq n$$

$$\exists (r_{2j} R r_{1i}) \Rightarrow (r_{1i}, \frac{w_{1i}}{v_1} \min(w_{2j}, v_1)) \in \mathcal{V}_0$$

$$1 \leq i \leq m$$

$$1 \leq j \leq n$$

gdzie :

$$v_2 = \sum_{r_{2j} \in L_1^*(r_{1i})} w_{2j}$$

$$v_1 = \sum_{r_{1i} \in L_1^*(r_{2j})} w_{1i}$$

c) pozostałe deskryptory z wagami ze zbioru $\mathcal{V}_1, \mathcal{V}_2$

$$\exists r_{1i} \neq r_{2j} \wedge r_{1i} \sim R r_{2j} \wedge r_{2j} \sim R r_{1i} \Rightarrow$$

$$1 \leq i \leq m$$

$$1 \leq j \leq n$$

$$\implies (r_{1i}, w_{1i}) \in \mathcal{V}_0' \wedge (r_{2j}, w_{2j}) \in \mathcal{V}_0'$$

Zbiór $\mathcal{V}_0'$, w którym wagi zostały podzielone przez sumę wszystkich wag występujących w zbiorze $\mathcal{V}_0'$ nazwiemy zbiorem $\mathcal{V}_1 + \mathcal{V}_2$.

Twierdzenie 1.2.4.

Niech $\mathcal{V}_1, \mathcal{V}_2$ będą kwerendami postaci (1.2.3),

$$(1.2.4) \text{ oraz } \mathcal{V}_1 \cdot \mathcal{V}_2 = \{(r_{01}, w_{01}), (r_{02}, w_{02}), \dots, (r_{0m_1}, w_{0m_1})\},$$

$$\mathcal{V}_1 + \mathcal{V}_2 = \{(r'_{01}, w'_{01}), (r'_{02}, w'_{02}), \dots, (r'_{0m_1}, w'_{0m_1})\}.$$

Jeśli

$$c_1 = \min_{1 \leq k \leq m_1} \left(\sum_{r_{1i} \in L_1^*(r_{0k})} w_{1i} \right) / w_{0k} \quad (1.2.5)$$

$$c_2 = \min_{1 \leq k \leq m_1} \left(\sum_{r_{2j} \in L_1^*(r_{0k})} w_{2j} \right) / w_{0k} \quad (1.2.6)$$

$$c'_1 = \min_{1 \leq i \leq m} \left(\sum_{r'_{01} \in L_1^*(r_{1i})} w'_{01} \right) / w_{1i} \quad (1.2.7)$$

$$c'_2 = \min_{1 \leq j \leq n} \left(\sum_{r'_{01} \in L_1^*(r_{2j})} w'_{01} \right) / w_{2j} \quad (1.2.8)$$

to

$$\mathcal{V}_1 \cdot \mathcal{V}_2 \triangleleft_{c_1} \mathcal{V}_1 \quad (1.2.9)$$

$$\mathcal{V}_1 \cdot \mathcal{V}_2 \triangleleft_{c_2} \mathcal{V}_2 \quad (1.2.10)$$

oraz

$$\mathcal{V}_1 \triangleleft_{c'_1} \mathcal{V}_1 + \mathcal{V}_2 \quad (1.2.11)$$

$$\mathcal{V}_2 \triangleleft_{c'_2} \mathcal{V}_1 + \mathcal{V}_2 \quad (1.2.12)$$

Dowód

Aby udowodnić związek (1.2.9) należy sprawdzić nierówność

$$\forall (r_{0,1}, w_{01}) \in \mathcal{P}_1 \cdot \mathcal{P}_2 \quad w_{01} \cdot c_1 \leq \sum_{r_{1i} \in L_1^*(r_{01})} w_{1i} \quad (1.2.13)$$

$$1 \leq i \leq m_1$$

którą można zapisać w postaci

$$\forall (r_{01}, w_{01}) \in \mathcal{P}_1 \cdot \mathcal{P}_2 \quad c_1 \leq \sum_{r_{1i} \in L_1^*(r_{01})} w_{1i} / w_{01}$$

$$1 \leq i \leq m_1$$

Korzystając ze wzoru (1.2.5) otrzymujemy nierówność

$$\forall (r_{01}, w_{01}) \in \mathcal{P}_1 \cdot \mathcal{P}_2 \quad \min_{1 \leq k \leq m_1} \left(\sum_{r_{1i} \in L_1^*(r_{0k})} w_{1i} / w_{0k} \right) \leq$$

$$\leq \sum_{r_{1i} \in L_1^*(r_{01})} w_{1i} / w_{01}$$

Która jest prawdziwa dla każdego l .

Korzystając z równań (1.2.6) - (1.2.8) podobnie dowodzi się związki (1.2.10) - (1.2.12).

Jeśli $\mathcal{P}_1 \cdot \mathcal{P}_2 \triangleleft_{c_1} \mathcal{P}_2$ i c_1^* jest dokładnością odpowiedzi $\delta(\mathcal{P}_1, c_1^*)$ a dokładność c_3 odpowiedzi

$\delta(\mathcal{P}_1 \cdot \mathcal{P}_2, c_3)$ spełnia warunek $c_3 \leq c_1 \cdot c_1^*$ to na mocy twierdzenia 1.2.2 o inkluzyjności systemu otrzymujemy

$$\delta(\mathcal{V}_1, c_1^*) \subset \delta(\mathcal{V}_1 \cdot \mathcal{V}_2, c_3).$$

Podobnie gdy $\mathcal{V}_1 \cdot \mathcal{V}_2 \triangleleft_{c_2} \mathcal{V}_2$ i c_2^* jest
jest dokładnością odpowiedzi $\delta(\mathcal{V}_2, c_2^*)$ oraz dokładność

c_4 odpowiedzi $\delta(\mathcal{V}_1 \cdot \mathcal{V}_2, c_4)$ spełnia warunek $c_4 \leq c_2 \cdot c_2^*$

to

$$\delta(\mathcal{V}_2, c_2^*) \subset \delta(\mathcal{V}_1 \cdot \mathcal{V}_2, c_4)$$

Dla $c \leq \min(c_1 \cdot c_1^*, c_2 \cdot c_2^*)$ zachodzi

$$\delta(\mathcal{V}_1, c_1^*) \cup \delta(\mathcal{V}_2, c_2^*) \subset \delta(\mathcal{V}_1 \cdot \mathcal{V}_2, c)$$

Jeśli $\mathcal{V}_1 \triangleleft_{c'_1} \mathcal{V}_1 + \mathcal{V}_2$ a c'_3 jest dokładnością
odpowiedzi $\delta(\mathcal{V}_1 + \mathcal{V}_2, c'_3)$ spełniają^{ca} warunek $c'_3 \geq$

$\geq c_1^* / c'_1$ to na mocy twierdzenia 1.2.2 otrzymujemy

$$\delta(\mathcal{V}_1 + \mathcal{V}_2, c'_3) \subset \delta(\mathcal{V}_1, c_1^*)$$

Podobnie gdy $\mathcal{V}_2 \triangleleft_{c'_2} \mathcal{V}_1 + \mathcal{V}_2$ oraz dokładność
 c'_4 odpowiedzi $\delta(\mathcal{V}_1 + \mathcal{V}_2, c'_4)$ spełnia warunek

$$c'_4 \geq c_2^* / c'_2 \text{ to } \delta(\mathcal{V}_1 + \mathcal{V}_2, c'_4) \subset \delta(\mathcal{V}_1, c_2^*)$$

Dla $c' \geq \max(c_1^* / c'_1, c_2^* / c'_2)$ zachodzi

$$\delta(\mathcal{V}_1 + \mathcal{V}_2, c') \subset \delta(\mathcal{V}_1, c_1^*) \cap \delta(\mathcal{V}_2, c_2^*)$$

Wniosek.

W systemie wyszukiwania informacji opartym na tezaurysie
z wagami deskryptorów odpowiedź na kwerendę $\mathcal{V}_1 + \mathcal{V}_2$ sta-
nowi podzbiór przecięcia odpowiedzi kwerend $\mathcal{V}_1$ i $\mathcal{V}_2$, a
kwerenda $\mathcal{V}_1 \cdot \mathcal{V}_2$ zawiera w sobie przecięcie odpowiedzi
kwerend $\mathcal{V}_1$ i $\mathcal{V}_2$. Dokładności tych odpowiedzi zostały
podane wyżej.

1.3. Geometryczny model systemu wyszukiwania informacji

1.3.1. Oznaczenia i definicje

Rozpatrzmy zbiór deskryptorów T_0 , skończony, uporządkowany częściowo przez relację R . Niech $T^* \subset T_0$ będzie zbiorem wszystkich deskryptorów parami nieporównywalnymi w sensie relacji R , tj. dla $t' \neq t''$

$$\{t', t''\} \subset T^* \iff \{t', t''\} \subset T_0 \times T_0 \setminus R$$

Ponieważ T_0 jest zbiorem skończonym, skończony jest także zbiór T^* . Oznaczając przez s moc zbioru T^* , na mocy twierdzenia Dilwortha ([3], p.72) możemy utworzyć s ciągów

$$T_i = \{t_{i1}, t_{i2}, \dots, t_{im_i}\}, \quad i = 1, 2, \dots, s \quad (1.3.1)$$

spełniając warunki

$$t_{im_i} R t_{im_i-1} R \dots R t_{i1}, \quad i = 1, 2, \dots, s \quad (1.3.2)$$

$$\forall \begin{matrix} t \in T_0 \\ 1 \leq i \leq s \\ 1 \leq j \leq m_i \end{matrix} \quad \exists \quad t = t_{ij} \quad (1.3.3)$$

$$T_i \cap T_j = \emptyset, \quad i \neq j, \quad 1 \leq i, j \leq s \quad (1.3.4)$$

W każdym zbiorze T_i relacja R indukuje porządek liniowy.

Do każdego T_i dołączony jest pseudodeskryptor $\star$ taki, że $\star \in T_0$ i $t R \star$ dla dowolnego $t \in T$.

Z warunku (1.1.2) wynika, że w każdym opisie dokumentu liczba deskryptorów $k \leq s$. Na mocy warunku (1.1.2) możemy twierdzić, że poszczególne deskryptory występujące w opisie

dokumentu należą do różnych zbiorów T_i , tj. przy ustalonym

$$\forall_{t', t''} \forall_{\{ (t', w'), (t'', w'') \} \in \mathcal{B}} \Rightarrow \exists_{1 \leq i \leq s} \exists_{1 \leq j \leq s} (i \neq j \wedge t' \in T_i \wedge t'' \in T_j) \quad (1.3.5)$$

Niech $\mathcal{T}$ oznacza zbiór opisów dokumentów, $\mathcal{R}$ zbiór kwerend. Dla $\mathcal{B} \in \mathcal{T}$ określimy zbiór

$$I(\mathcal{B}) = \{ i : \exists_{(t, w) \in \mathcal{B}} t \in T_i \}.$$

Ponieważ dla każdego $i \in \{1, 2, \dots, s\}$ istnieje co najwyżej deskryptor z wagą $(t, w) \in \mathcal{B} \wedge t \in T_i$, dla $i \in I(\mathcal{B})$ można jednoznacznie ustalić deskryptor $t(i) : (t(i), w) \in \mathcal{B} \wedge t(i) \in T_i$. Podobnie określamy zbiór $I(\mathcal{R})$ oraz element kwerendy, który wchodzi do ciągu T_i , $i \in I(\mathcal{R})$. Dla elementu kwerendy r , określimy ponadto zbiór $J(r) = \{ j :$

$$\exists_{k \in \{1, 2, \dots, m_j\}} r R t_{jk} \}. \text{ Dla } j \in J(r) \text{ oznaczmy przez}$$

$k(r, j)$ największe k , przy którym zachodzi jeszcze $r R t_{jk}$. Przekształcenie opisów dokumentów w ciągi s -elementowe par (deskryptor, waga) zdefiniujemy następująco:

$$a(\mathcal{B})_i = \begin{cases} (t(i), w) & i \in I(\mathcal{B}) \\ (*, 0) & i \notin I(\mathcal{B}) \end{cases} \quad (1.3.6)$$

gdzie przez $a(\mathcal{B})_i$ rozumiemy i -ty element ciągu, w jest wagą deskryptora $t(i)$ występującego w opisie dokumentu. Ciągi te nazywać będziemy współrzędnymi deskryptorowymi - opisów dokumentów w przestrzeni dokumentów.

Podobnie przekształcamy opisy kwerend w ciągi s -elementowe par (deskryptor, waga)

$$b(\mathcal{V})_i = \begin{cases} (r(i), w) & i \in I(\mathcal{V}) \\ (*, 0) & i \notin I(\mathcal{V}) \end{cases} \quad (1.3.7)$$

gdzie $b(\mathcal{V})_i$ oznacza i -ty element ciągu, w jest wagą deskryptora $r(i)$ występującego w kwerendzie $\mathcal{V}$. Ciągi te nazywać będziemy współrzędnymi deskryptorowymi kwerend w przestrzeni PD.

Dla danej kwerendy $\mathcal{V}$ określamy ciąg s -elementowy deskryptorów postaci:

$$h(\mathcal{V})_j = \begin{cases} r(i) & i = j \wedge i \in I(\mathcal{V}) \\ t(j)_i & \exists i \neq j \wedge j \in J(r(i)) \wedge l=k(r, j) \\ & i \in I(\mathcal{V}) \\ t(j)_{m_j} & \text{w p.p.} \end{cases}$$

gdzie $h(\mathcal{V})_j$ jest j -tym elementem ciągu $h(\mathcal{V})$, $t(j)_i$ oznacza i -ty element w zbiorze deskryptorów T_j .

1.3.2. Konstrukcja przestrzeni dokumentów PD.

Przestrzeń opisów dokumentów jest przestrzenią s -wymiarową. Pseudodeskryptor $*$ (deskryptor zerowy) jest początkiem układu współrzędnych. Zbiory T_i wyznaczają poszczególne osie. Na osiach T_i leżą deskryptory ze zbioru T_i . Deskryptory szczegółowe leżą bliżej początku układu współrzędnych, ogólniejsze dalej. Odległość pomiędzy deskryptorami na poszczególnych osiach układu wynosi 1. Deskryptory na osiach T_i ($i=1, 2, \dots, s$) wyznaczają przestrzeń składającą się z

$$m = \prod_{i=1}^s m_i \quad (m_i = |T_i|) \quad \text{jednostkowych kostek } s - \text{wymia-}$$

rowych. Opisy dokumentów traktować będziemy jako punkty leżące wewnątrz kostek. Odcinki $(*, t_{11}]$, $(t_{11}, t_{12}]$, $(t_{12}, t_{13}] \dots$ są długości 1. Każdy odcinek utożsamiać będziemy z jego prawym końcem. Waga deskryptora t_{ij} w opisie dokumentu określa położenie punktu względem prawego końca odcinka $(t_{ij-1}, t_{ij}]$. Deskryptory opisu dokumentu - punktu w przestrzeni PD określają kostkę, w której znajduje się punkt a wagi określają położenie punktu w kostce. Ponieważ w każdym opisie suma wag jest równa 1, punkty będą leżeć na przekątnych kostek. Deskryptory z większymi wagami będą znajdować się bliżej wierzchołków, którym odpowiadają dane deskryptory. Nie w każdej kostce punkty będą leżeć na przekątnej kostki, punkty mogą znajdować się tylko na przekątnych ścian bocznych lub na wierzchołkach kostek, ponieważ w opisach dokumentów występują tylko deskryptory spełniające warunek (1.1.2).

Przykład 1.3.1.

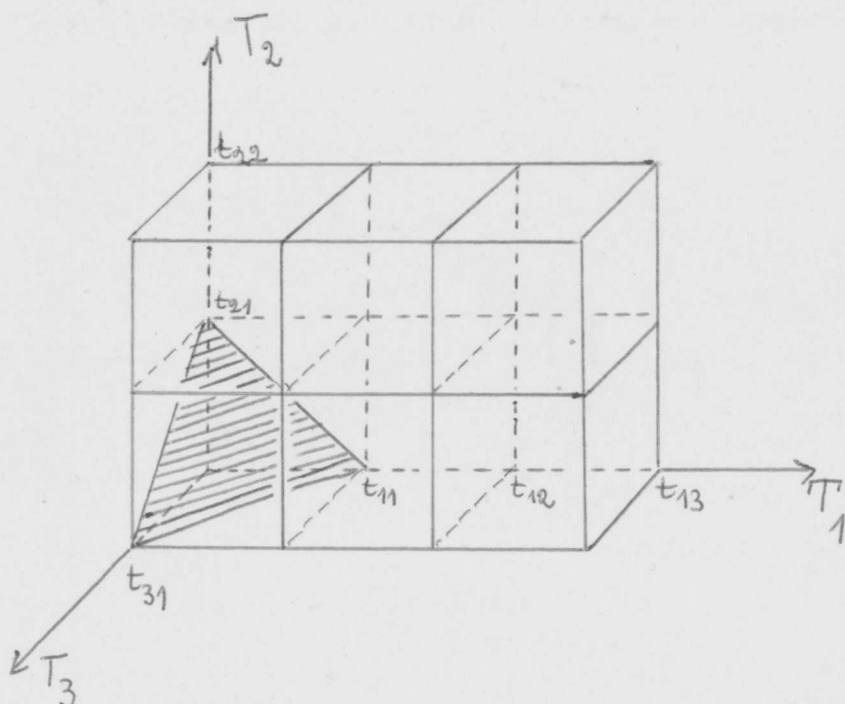
$$\text{Niech } T_1 = \{t_{11}, t_{12}, t_{13}\}$$

$$T_2 = \{t_{21}, t_{22}\}$$

$$T_3 = \{t_{31}\}$$

będą zbiorami deskryptorów.

Przestrzeń rozpięta na osiach T_1, T_2, T_3 ma postać



Zakreskowana przekątna kostki jest miejscem geometrycznym punktów, które w swoich opisach zawierają deskryptory t_{31}, t_{21}, t_{11}

Rzutem dowolnego punktu na jedną z osi T_i nazywać będziemy parę (deskryptor, waga). W celu łatwiejszego oznaczania kostek przestrzeni PD ciągi s-elementowe par (deskryptor, waga) zdefiniowane przy pomocy (1.3.6), (1.3.7) można zastąpić ciągami s-elementowymi w postaci par (numer deskryptora, waga)

$$A(\mathcal{E})_i = \begin{cases} (n(t,i), w) & i \in I(\mathcal{E}) \\ (0, 0) & i \notin I(\mathcal{E}) \end{cases} \quad (1.3.9)$$

gdzie $A(\mathcal{E})_i$ oznacza i-ty element ciągu, w jest wagą deskryptora $t(i)$ występującego w $\mathcal{E}$, $n(t,i)$ jest

numerem deskryptora $t(i)$ w zbiorze T_i . Ciągi te nazywać będziemy współrzędnymi liczbowymi punktu przestrzeni.

Współrzędne liczbowe punktów - kwerend mają postać:

$$B(\mathcal{V})_i = \begin{cases} (n(r,i), w) & i \in I(\mathcal{V}) \\ (0, 0) & i \notin I(\mathcal{V}) \end{cases} \quad (1.3.10)$$

gdzie $B(\mathcal{V})_i$ - oznacza i -tą współrzędną punktu $\mathcal{V}$,
 w jest wagą deskryptora $r(i)$ w kwerendzie $\mathcal{V}$,
 oraz

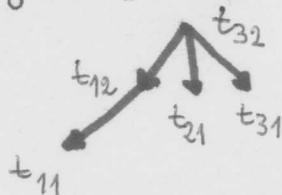
$$H(\mathcal{V})_j = \begin{cases} k(r,i) & i \in I(\mathcal{V}) \wedge i = j \\ k(r,j) & \exists i \in I(\mathcal{V}) \quad i \neq j \wedge j \in J(r(i)) \\ m_j & \text{w. p.p.} \end{cases} \quad (1.3.11)$$

gdzie $H(\mathcal{V})_j$ oznacza j -tą współrzędną opisu kostki.

Pierwszą część par $A(t)_i, B(\mathcal{V})_i$ nazywać będziemy poprzednikami a drugą następnikami rzutów. Poprzedniki rzutów określają położenie kostek względem początku układu współrzędnych a następniki położenie punktów wewnątrz kostek.

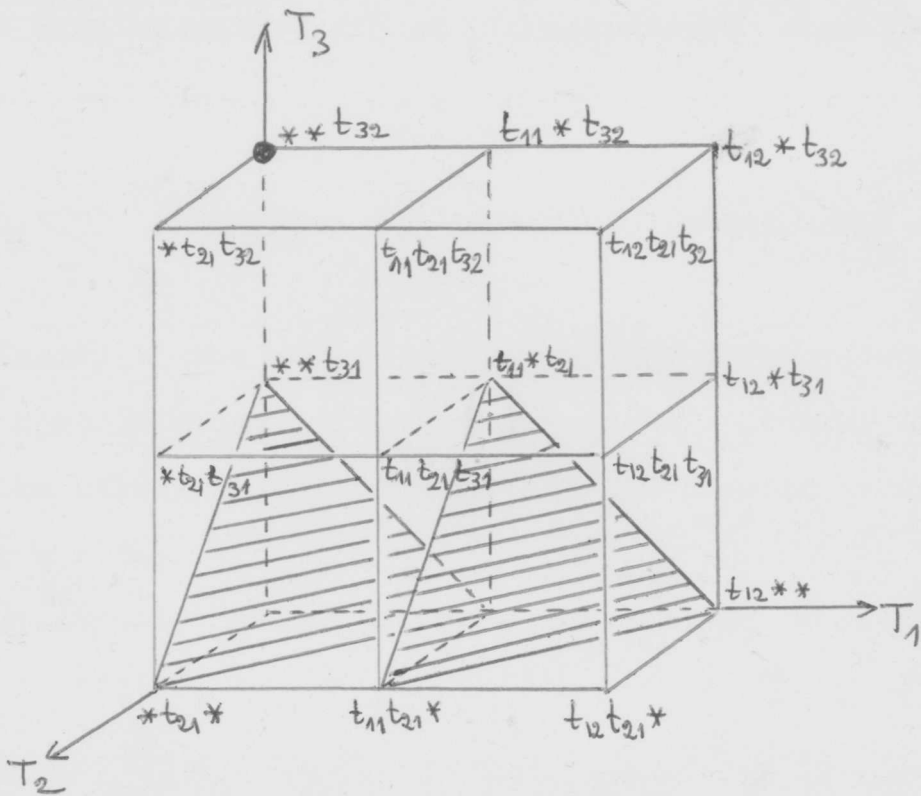
Przykład 1.3.2.

Jeśli tezaurus T_0 ma postać

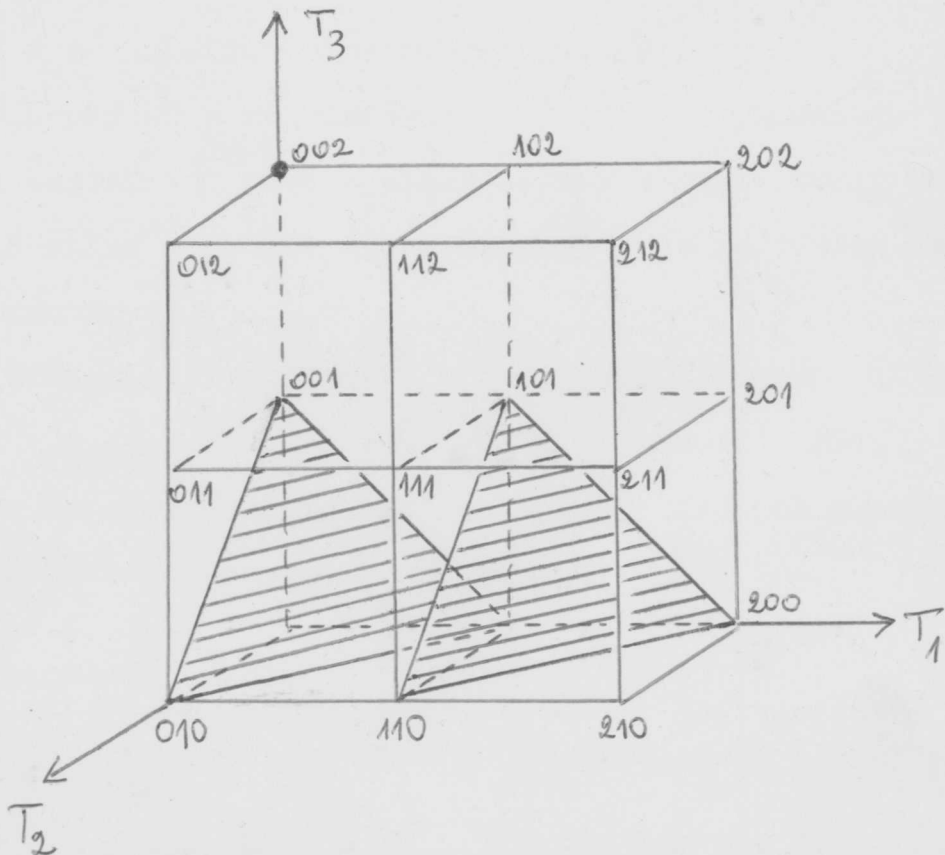


oraz zbiory T_i : $T_1 = \{t_{11}, t_{12}\}$, $T_2 = \{t_{21}\}$, $T_3 = \{t_{31}, t_{32}\}$

to opis kostek za pomocą współrzędnych deskryptorowych ma postać :



Opis kostek za pomocą współrzędnych liczbowych jest postaci :



W przestrzeni PD dla tezaursu T_0 punkty opisów dokumentów będą znajdować się na zakreskowanych obszarach rysunku.

1.3.3. Własności przestrzeni dokumentów PD.

Każdej kostce można przyporządkować s-elementowy ciąg, który opisuje wierzchołek kostki najdalej położony względem początku układu współrzędnych. Ciąg ten nazywać będziemy opisem kostki.

Otoczeniem dowolnego punktu d nazywać będziemy zbiór $X(d, c)$ postaci:

$$X(d, c) = \{d' : d' \in PD \wedge \mu(\alpha(\beta(d)), \alpha(\beta(d'))) \geq c\}$$

gdzie $\alpha(\beta(d)), \alpha(\beta(d'))$ są współrzędnymi deskryptorowymi punktów d i d' , μ jest miarą podobieństwa dla punktów d i d' zdefiniowaną wzorem (1.1.7).

Dla punktu o współrzędnych $b(\mathcal{V})_i$ wyznaczamy wierzchołek kostki $h(\mathcal{V})$ wyznaczający podprzestrzeń kostek, których opisy składają się z deskryptorów ściślejszych od opisu kostki $h(\mathcal{V})$.

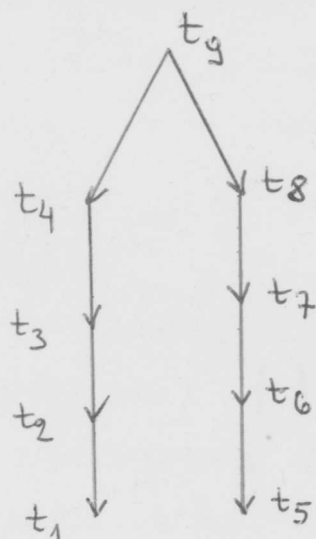
Odpowiedzią na kwerendę $b(\mathcal{V})$ jest otoczenie $X(b(\mathcal{V}), c)$ punktu $b(\mathcal{V})$. Wyznaczenie podprzestrzeni $PD(h(\mathcal{V}))$ pozwala na ograniczenie liczby kostek, których punkty będą należeć do otoczenia punktu $b(\mathcal{V})$.

Przykład 1.3.3.

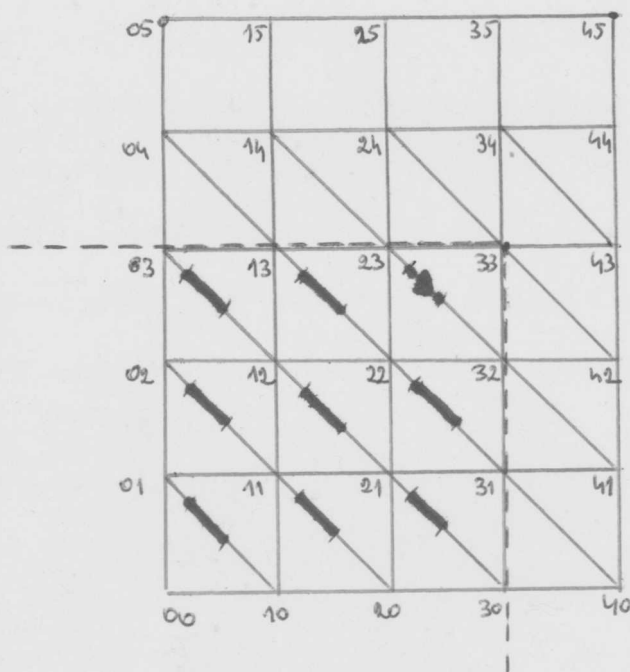
Niech zbiory deskryptorów T_i zawierają elementy:

$$T_1 = \{t_1, t_2, t_3, t_4\}, \quad T_2 = \{t_5, t_6, t_7, t_8, t_9\} \quad \text{oraz}$$

graf relacji hierarchii R pomiędzy deskryptorami ma postać :



Przestrzeń dokumentów PD można przedstawić następująco:



Jeśli $\mathcal{V} = \{(t_3, 0.4), (t_7, 0.6)\}$ jest kwerendą to

$$\mathbb{B}(\mathcal{V}) = \{(3, 0.4), (3, 0.6)\} \text{ i } \mathbb{H}(\mathcal{V}) = (3, 3).$$

W przestrzeni PD dokumenty leżą na przekątnych kwadratów i punkcie $(0,5)$.

Na rysunku trójkątem  zaznaczono położenie punktu kwerendy w przestrzeni, wierzchołek kostki $(3,3)$ wyznacza pod-

przestrzeń kwadratów, których opisy składają się z deskryptorów ściślejszych od opisu kwadratu (3,3). Otoczenie punktu - kwerendy zaznaczono odcinkami na przekątnych. Zbiór ten jest odpowiedzią na kwerendę $B(\mathcal{V})$. Za c przyjęto dokładność równą 0.5.

Przy pomocy zbiorów będących otoczeniami punktów leżących na osiach współrzędnych można łatwo znaleźć otoczenie dowolnego punktu w przestrzeni PD. Otoczeniem dowolnego punktu $B(\mathcal{V})$ w przestrzeni będzie przecięcie zbiorów będących otoczeniami punktów, które są rzutami punktu $B(\mathcal{V})$ na poszczególne osie układu współrzędnych. Poprzedniki rzutu punktu $B(\mathcal{V})$ określają zbiory a następniki rzutu i liczba c wyznaczają podzbiory tych zbiorów. Przecięcie obliczamy dla podzbiorów. Otoczenie punktu leżącego na osi układu współrzędnych nazywać będziemy listą inwersyjną, ponieważ otoczenie punktu ma postać zbliżoną do powszechnie znanej listy inwersyjnej. W następnym rozdziale zostaną podane szczegółowe opisy list inwersyjnych oraz zbadane ich własności.

1.4. Własności list inwersyjnych.

Lista inwersyjna $L(t)$ dla deskryptora $t \in T_0$ jest zbiorem par (dokument, waga). Do zbioru $L(t)$ należą dokumenty, które posiadają deskryptor t w swoim opisie. Waga dokumentu jest równa wadze z jaką deskryptor t wystąpił w opisie tego dokumentu.

Definicja 1.4.1.

Niech $\mathcal{L}(d) = \{(t_1, w_1), (t_2, w_2), \dots, (t_{k(d)}, w_{k(d)})\}$

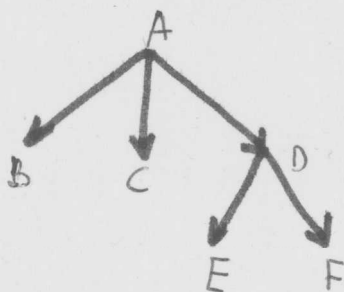
będzie opisem dokumentu $d \in D$.

$$\forall t_i \in T_0 \quad L(t_i) = \{ (d, v) : (t_i, w_i) \in \mathcal{G}(d) \wedge v = w_i \}$$

Liczba list inwersyjnych równa jest liczbie deskryptorów ze zbioru T_0 . Dokument, którego opis zawiera 1 deskryptorów, będzie należał do 1 list inwersyjnych. W dalszej części rozdziału ze względów praktycznych dokument utożsamiać będziemy z jego numerem (adresem).

Przykład 1.4.1.

Niech zbiór deskryptorów T_0 ma postać $T_0 = \{ A, B, C, D, E \}$, graf G przedstawia relacje hierarchii R pomiędzy deskryptorami



a opisy dokumentów są następujące :

$$\mathcal{B}_1 = \{ (B, 0.4), (C, 0.3), (E, 0.3) \}$$

$$\mathcal{B}_2 = \{ (C, 0.7), (D, 0.3) \}$$

$$\mathcal{B}_3 = \{ (C, 0.4), (D, 0.6) \}$$

$$\mathcal{B}_4 = \{ (B, 0.1), (C, 0.4), (D, 0.5) \}$$

$$\mathcal{B}_5 = \{ (E, 0.4), (F, 0.6) \}$$

Listy inwersyjne dla poszczególnych deskryptorów są postaci :

$$L(A) = \emptyset$$

$$L(B) = \{ (1, 0.4), (4, 0.1) \}$$

$$L(C) = \{ (1, 0.3), (2, 0.7), (3, 0.4), (4, 0.4) \}$$

$$L(D) = \{ (2, 0.3), (3, 0.6), (4, 0.5) \}$$

$$L(E) = \{(1, 0.3), (5, 0.4)\}$$

$$L(F) = \{(5, 0.6)\}$$

Na listach inwersyjnych $L(t)$ określone są działania sumy $\cup$ i iloczynu $\cap$. Suma i iloczyn list inwersyjnych jest listą inwersyjną.

Definicja 1.4.2.

$$a) (d_1, v_1) = (d_2, v_2) \iff d_1 = d_2 \wedge v_1 = v_2$$

$$b) (d, v) \in L(t_1) \cup L(t_2) \iff \\ \exists_{d_1 \neq d_2} [(d, v) = (d_1, v_1) \vee (d, v) = (d_2, v_2)] \vee \\ \exists_{d_1 = d_2} [(d, v) = (d_1, v_1 + v_2)]$$

$$c) (d, v) \in L(t_1) \cap L(t_2) \iff \bigvee_{d_1 = d_2} (d, v) = (d_1, v_1 + v_2)$$

przy czym $(d_1, v_1) \in L(t_1)$ i $(d_2, v_2) \in L(t_2)$

Jeżeli dany dokument należy do kilku różnych list inwersyjnych to waga dokumentu należącego do sumy lub przekroju tych list jest sumą wag jakie dokument posiadał w poszczególnych listach.

Przykład 1.4.2.

Opisy dokumentów i postać list inwersyjnych jest taka sama jak w przykładzie 1.4.1. Suma list $L(B)$ i $L(D)$ ma postać:

$$L(B) \cup L(D) = \{(1, 0.4), (2, 0.3), (3, 0.6), (4, 0.6)\}$$

Iloczyn list $L(B)$ i $L(C)$ ma postać:

$$L(B) \cap L(C) = \{(1, 0.7), (4, 0.5)\}.$$

Twierdzenie 1.4.1.

$$\bigvee_{t_1 R t_2} L(t_1) \cap L(t_2) = \emptyset$$

Dowód

Założmy, że $\exists_{(d,v)} (d,v) \in L(t_1) \cap L(t_2)$

Z definicji 1.4.2.c. wynika, że $\exists_{(t_1,v_1)} (t_1,v_1) \in \mathcal{G}(d) \wedge$

$\wedge \exists_{(t_2,v_2)} (t_2,v_2) \in \mathcal{G}(d)$ co daje sprzeczność z warunkiem (1.1.2).

Definicje 1.4.1 i 1.4.2 nie uwzględniają relacji hierarchii deskryptorów.

Definicja 1.4.3.

Oznaczmy przez X_i zbiór wszystkich deskryptorów ściślejszych od deskryptora t_i , a więc dla każdego $t_i \in T_0$

$$X_i = \{t : t \in T_0 \wedge t_i R t\}$$

Lista inwersyjna $L_R(t_i)$ dla deskryptora t_i uwzględniająca relację hierarchii R ma postać :

$$L_R(t_i) = \bigcup_{t \in X_i} L(t)$$

Przykład 1.4.3.

Zbiór deskryptorów i opisy dokumentów są takie same jak w przykładzie 1.4.1.

Lista inwersyjna dla deskryptora D z uwzględnieniem relacji R ma postać :

$$L_R(D) = L(D) \cup L(E) \cup L(F) = \\ = \{(1, 0.3), (2, 0.3), (3, 0.6), (4, 0.5), (5, 1.0)\}$$

Obliczanie sumy $\bigcup_{t \in X_i} L(t)$ zgodnie z definicją

1.4.2.b pozwala na usunięcie powtórzeń dokumentów. Obliczone wagi dokumentów z listy $L_R(t_i)$ zawierają sumy wag dokumentów, w opisach których znajdują się deskryptory ściślejsze od deskryptora t_i . W czasie pracy systemu wyszukiwania informacji tak obliczona suma list inwersyjnych pozwala na bardzo proste obliczanie stopnia podobieństwa kwerendy Q do opisu dokumentu $\theta(d)$ za pomocą miary (1.1.8) bez żadnych dodatkowych obliczeń. Wagi dokumentów znajdujące się na listach inwersyjnych będą składnikami sumy (1.1.8).

Definicja 1.4.4.

Niech

$$L_R(t) = \{(d_1, v_1), (d_2, v_2), \dots, (d_k, v_k)\}$$

będzie listą inwersyjną dla deskryptora t .

Na zbiorze $L_R(t)$ określona jest relacja porządku liniowego $\gg$, szeregująca elementy zbioru $L_R(t)$ ze względu na wielkości wag, tj.

$$\bigwedge_{\substack{1 \leq i, j \leq k \\ i \neq j}} (d_i, v_i) \gg (d_j, v_j) \iff v_i \geq v_j$$

W dalszych rozważaniach zakładać będziemy, że na listach inwersyjnych $L_R(t)$ elementy są umieszczone w kolejności malejących wag dokumentów

Twierdzenie 1.4.2.

Pomiędzy zbiorami $L_R(t_i)$ i $\bigcup_{t \in X_i} L(t)$ zachodzi

następujący związek :

$$\overline{L_R(t_i)} \subseteq \sum_{t \in X_i} \overline{L(t)}$$

Dowód wynika z definicji 1.4.2.b. W czasie obliczania sumy $\bigcup$ list inwersyjnych usuwane są powtórzenia dokumentów.

Definicja 1.4.5.

a) Mówimy, że lista $L(t_2)$ zawiera się w liście $L(t_1)$ i piszemy $L(t_2) \subset L(t_1)$ wtedy i tylko wtedy gdy

$$\forall (d_2, v_2) \in L(t_2) \quad \exists (d_1, v_1) \in L(t_1) \quad d_1 = d_2 \wedge v_1 \leq v_2$$

b) Mówimy, że lista $L(t_2)$ zawiera się w liście $L(t_1)$ z dokładnością b i piszemy $L(t_2) \subset_b L(t_1)$ wtedy i tylko wtedy gdy

$$\forall (d_2, v_2) \in L(t_2) \quad \exists (d_1, v_1) \in L(t_1) \quad \exists d_1 = d_2 \wedge v_1 \cdot b \leq v_2 \quad 0 \leq b \leq 1$$

Uwaga: Przy $b = 1$ definicja "zawierania się z dokładnością b " pokrywa się z definicją "zawierania się".

Twierdzenie 1.4.3.

$$\forall t \in T_0 \quad L(t) \subset L_R(t)$$

Dowód wynika z definicji 1.4.3.

Twierdzenie 1.4.4.

$$\bigvee_{t_1 R t_2} \exists \quad L_R(t_2) \subset_{b'} L_R(t_1) \quad 0 \leq b' \leq a$$

gdzie

$$a = \min_{\substack{(d, v_i) \in L_R(t_1) \\ (d, v_j) \in L_R(t_2)}} \left\{ \frac{v_j}{v_i} \right\}$$

Dowód

Niech

$$X_1 = \{t : t \in T_0 \wedge t_1 R t\}$$

$$X_2 = \{t : t \in T_0 \wedge t_2 R t\}$$

będą zbiorami deskryptorów ściślejszych odpowiednio od t_1, t_2 .

$$\bigvee_{t_1, t_2} t_1 R t_2 \Rightarrow X_2 \subset X_1$$

Listy inwersyjne dla deskryptorów t_1, t_2 mają postać:

$$L_R(t_1) = \bigcup_{t \in X_1} L(t)$$

$$L_R(t_2) = \bigcup_{t \in X_2} L(t)$$

Każdy dokument z listy $L_R(t_2)$ należy do listy $L_R(t_1)$. Sprawdzić musimy czy jest spełniony warunek w definicji 1.4.5.b. dotyczący wag dokumentów. Ponieważ b' zostało określone jako minimum z ilorazów wag dokumentu d z list $L_R(t_2), L_R(t_1)$ dla wszystkich dokumentów ze zbioru $L_R(t_2)$, więc dla każdej wagi dokumentu d z listy $L_R(t_2)$ jest spełniony warunek $v_1 b' \leq v_2$.

Z ostatniego twierdzenia wynika następujące twierdzenie:

Twierdzenie 1.4.5.

$$\bigvee_{t_2 R t_1} L_R(t_1) \cap L_R(t_2) \neq \emptyset$$

Definicja 1.4.6.

Listę $L_R(t) \Big|_a$ nazywać będziemy listą obciętą przez

liczbę a . Do listy $L_R(t) \Big|_a$ będą należeć pary (dokument, waga) z listy $L_R(t)$, których wagi są nie mniejsze od liczby a , tj.

$$L_R(t) \Big|_a = \{ (d, v) = (d, v) \in L_R(t) \wedge v \geq a \}$$

Łatwo zauważyć, że dla $a = 0$, $L_R(t) \Big|_a = L_R(t)$ oraz

dla $0 < a \leq 1$, $L_R(t) \Big|_a \subset L_R(t)$.

Podstawowe własności obciętych list można przedstawić w postaci twierdzenia.

Twierdzenie 1.4.6.

- a) $L_R(t_1) \subset L_R(t_2) \Rightarrow L_R(t_1) \Big|_a \subset L_R(t_2) \Big|_a$
- b) jeśli $a_1 \geq a_2$ to $L_R(t) \Big|_{a_1} \subset L_R(t) \Big|_{a_2}$
- c) $\exists_{t_1, t_2} t_1 R t_2 \wedge L_R(t_2) \Big|_{a_1} \subset_b L_R(t_1) \Big|_{a_2} \iff \exists_{0 \leq b \leq 1} a_1 \cdot b \leq a_2$

Dowód związków a i b wynika z definicji 1.4.5a i 1.4.6.

Dowód związku c.

Jeśli $\exists_{t_1, t_2} t_1 R t_2$ to na podstawie twierdzenia

1.4.4. istnieje takie b, że $L_R(t_2) \subset L_R(t_1)$.

Z definicji 1.4.5.b otrzymujemy warunek dla wag dokumentu należącego do obu list, tj. $v_1 \cdot b \leq v_2$. Stąd wynika, że liczby a_1, a_2 muszą spełniać podobną zależność, tj. $a_1 \cdot b \leq a_2$.

W dalszych rozważaniach będziemy zakładać, że liczba a ma postać $a = c \cdot w$, gdzie w jest wagą deskryptora w kwerendzie odwołującej się do danej listy inwersyjnej, c jest dokładnością odpowiedzi wymaganą przez użytkownika (znaczenie c jest takie same jak w rozdziale 1.2.1). W następnej definicji zostanie podany sposób obliczania iloczynu $\cap$ list obciętych przez liczbę $a = c \cdot w$.

Definicja 1.4.7.

$$\left. (d, v) \in L_R(t_1) \right|_{c \cdot w_1} \cap \left. L_R(t_2) \right|_{c \cdot w_2} \Leftrightarrow \bigvee_{d_1 = d_2} (d, v) = (d_1, \min(w_1, v_1) + \min(w_2, v_2))$$

przy czym $\left. (d_1, v_1) \in L_R(t_1) \right|_{c \cdot w_1}$ i

$$\left. (d_2, v_2) \in L_R(t_2) \right|_{c \cdot w_2}$$

Definicja 1.4.7. jest uogólnieniem definicji 1.4.2.c.

Jeśli w definicji 1.4.7 przyjmiemy $c = 0$ i $w_1 = w_2 = 1$

to otrzymamy definicję równoważną definicji 1.4.2.c.

Definicja 1.4.8. (odpowiedź na kwerendę)

Niech $\mathcal{V} = \{(r_1, w_1), (r_2, w_2), \dots, (r_k, w_k)\}$ będzie kwerendą, c żadaną dokładnością odpowiedzi. Odpowiedzią na kwerendę $\mathcal{V}$ z dokładnością c nazywać będziemy

$$\mathcal{S}(\mathcal{V}, c) = \left\{ (d, v) : (d, v) \in \bigcap_{i=1}^k (L_R(r_i) \mid_{c \cdot w_i}) \right\}$$

Dla każdej pary (d, v) należącej do odpowiedzi v jest stopniem podobieństwa kwerendy $\mathcal{V}$ do opisu dokumentu $\mathcal{U}(d)$.

Przykład 1.4.4.

Zbiór deskryptorów i opisy dokumentów są takie same jak w przykładzie 1.4.1.

Uporządkowane listy inwersyjne z uwzględnieniem relacji hierarchii mają postać:

$$L_R(A) = \{(1, 1.0), (2, 0.0), (3, 1.0), (4, 1.0), (5, 1.0)\}$$

$$L_R(B) = \{(1, 0.4), (4, 0.1)\}$$

$$L_R(C) = \{(2, 0.7), (3, 0.4), (4, 0.4), (1, 0.3)\}$$

$$L_R(D) = \{(5, 1.0), (3, 0.6), (4, 0.5), (1, 0.3), (2, 0.3)\}$$

$$L_R(E) = \{(5, 0.4), (1, 0.3)\}$$

$$L_R(F) = \{(5, 0.6)\}$$

Niech $\mathcal{V} = \{(C, 0.5), (D, 0.5)\}$ i wymagana dokładność odpowiedzi $c = 0.8$.

Odpowiedź na kwerendę $\mathcal{V}$ z dokładnością c ma postać:

$$\mathcal{S}(\mathcal{V}, 0.8) = L_R(C) \mid_{0.8 \cdot 0.5} \cap L_R(D) \mid_{0.8 \cdot 0.5} =$$

$$= \{(2, 0.7), (3, 0.4), (4, 0.4), (1, 0.3)\} \Big|_{0.8 \times 0.5} \quad \Delta$$

$$\Delta \{(5, 1.0), (3, 0.6), (4, 0.5), (1, 0.3), (2, 0.3)\} \Big|_{0.8 \times 0.5} =$$

$$= \{(3, 0.9), (4, 0.9)\}$$

Do odpowiedzi na kwerendę $\mathcal{F}$ należą dokumenty 3 i 4 z dokładnością 0.9. Do odpowiedzi nie należą dokumenty 1 i 2 ponieważ wagi tych dokumentów na listach inwersyjnych są mniejsze od 0.4. Z drugiej strony sprawdzając opisy dokumentów 1 i 2 widzimy, że nie jest spełniony warunek (1.1.9) w definicji 1.1.2 dla opisów dokumentów.

Można teraz postawić pytanie, czy przy odpowiedziach systemu wyszukiwania informacji zdefiniowanych przy pomocy list inwersyjnych zostanie zachowana własność inkluzyjności modelu podana w rozdziale 1.1.2. Jeśli dla kwerend składających się z jednego deskryptora oznaczmy przez a_1 wymagana dokładność odpowiedzi dla jednej z kwerend, przez a_2 dokładność drugiej odpowiedzi i przez b dokładność relacji podobieństwa dla kwerend to odpowiedź na to pytanie wynika z twierdzenia 1.4.6.c. W przypadkach kiedy kwerenda składa się z większej liczby deskryptorów to odpowiedzią na kwerendę jest przekrój odpowiednich list inwersyjnych. W następnym twierdzeniu udowodnimy zawieranie się przekroju list inwersyjnych dla deskryptorów szczegółowych w przekroju list inwersyjnych dla deskryptorów ogólniejszych oraz, że własność inkluzyjności modelu jest zachowana. Dowód przeprowadzimy dla przekroju dwóch list. Dla dowolnej większej liczby list dowód można przeprowadzić za pomocą indukcji matematycznej.

Twierdzenie 1.4.7.

Definicja 1.4.8 zachowuje własność inkluzyjności modelu.

Dowód

Niech

$$\mathcal{V}_1 = \{(r_1, w_1), (r'_1, w'_1)\}$$

$$\mathcal{V}_2 = \{(r_2, w_2), (r'_2, w'_2)\}$$

będą kwerendami, w których $r_1 R r_2$ i $r'_1 R r'_2$.

Najpierw pokażemy, że jeśli dokument d z wagą v_2 należy do $L_R(r_2) \Big|_{a_2=w_2 c_2}$ i z wagą r'_2 należy do

$L_R(r'_2) \Big|_{a'_2=w'_2 c_2}$ z dokładnością c_2 to waga v

dokumentu należącego do przekroju tych list, tj

$$(d, v) \in L_R(r_2) \Big|_{a_2=w_2 c_2} \cap L_R(r'_2) \Big|_{a'_2=w'_2 c_2}$$

jest nie mniejsza od c_2 .

Korzystając z definicji 1.4.7 otrzymujemy, że

$$(d, v) = (d, \min(v_2, w_2) + \min(v'_2, w'_2)).$$

Aby dokument d z wagą v należał do przekroju list inwersyjnych z dokładnością nie mniejszą niż c_2 , musi zachodzić warunek $v \geq c_2$

Z własności $(d, v_2) \in L_R(r_2) \Big|_{a_2=w_2 c_2}$,

$$(d, v'_2) \in L_R(r'_2) \Big|_{a'_2=w'_2 c_2}$$

i definicji 1.4.6 wynika,

że $v_2 \geq w_2 c_2$ i $v'_2 \geq w'_2 c_2$

Korzystając z powyższych nierówności i $c_2 w_2 \leq w_2$,

$c_2 w'_2 \leq w'_2$ oraz warunku (1.1.3) otrzymujemy

$$c_2 = c_2(w_2 + w'_2) = c_2 w_2 + c_2 w'_2 \leq \min(v_2, w_2) + \min(v'_2, w'_2) = v$$

Pokazaliśmy, że $v \geq c_2$.

Analogicznie sprawdzając nierówności $v' \geq c_1$ dla

$$(d, v) \in L_R(r_1) \mid a_1 = w_1 c_1 \quad \Delta \quad L_R(r'_1) \mid a'_1 = w'_1 c_1$$

dowodzimy, że w tym przypadku dokładność c_1 jest także zachowana.

Korzystając z twierdzenia 1.4.6.c otrzymujemy, że przekrój list inwersyjnych dla deskryptorów szczegółowych zawiera się w przekroju list inwersyjnych dla deskryptorów ogólniejszych z dokładnością c oraz $c_1 c \leq c_2$

Przy obliczaniu przekrojów list inwersyjnych nie korzystamy z pełnych list inwersyjnych, bierzemy tylko te dokumenty, które spełniają warunek $\mu(\mathcal{P}, \mathcal{U}(d)) \geq c$. Przy wymaganych dużych dokładnościach odpowiedzi czas wyszukiwania może być więc krótszy niż w systemach bez wag deskryptorów.

Dokładniejsze uzasadnienie jest podane w rozdziale 4.3

(por. twierdzenie 4.3.1).

1.5. Wnioski

Modele systemów wyszukiwania informacji oparte na teaurusie podane przez W.M. Turskiego w pracy [7] i M. Dąbrowskiego w pracy [1] są szczególnymi przypadkami modelu systemu opi-

sanego w tym rozdziale.

Przypisując deskryptorom jednakowe wagi postaci $1/f_j$ (f_j - liczba różnych deskryptorów występujących w j - tym opisie dokumentu), odpowiedź z dokładnością 0 zawiera te same dokumenty co odpowiedź dla tej samej kwerendy w systemach [1] i [7].

Z praktyki wiadomo, że nie zawsze opis dokumentu jest adekwatny do treści dokumentu. Wynika to z faktu, że trudno jest wyspecyfikować dokładnie treść dokumentu za pomocą niewielkiej liczby deskryptorów występujących w jego opisie. Wprowadzenie wag deskryptorów pozwala w pewnym stopniu na precyzyjniejsze przedstawienie treści dokumentu za pomocą deskryptorów, bez potrzeby rozbudowy tezaursusa.

Weźmy na przykład opisy dwóch dokumentów $\{(algebra, 0.3), (geometria, 0.7)\}$ i $\{(algebra, 0.8), (geometria, 0.2)\}$. Opisy tych dokumentów nie będą rozróżnialne w systemach [1], [7], tymczasem pierwszy dokument bardziej zapewne zainteresuje użytkowników interesujących się geometrią, a drugi - użytkowników zajmujących się algebrą. Deskryptory z wagami w kwerendach pozwalają lepiej odzwierciedlić rzeczywiste potrzeby użytkowników. Wprowadzenie wag do opisów dokumentów pozwala na wyeliminowanie z odpowiedzi systemu dokumentów z małymi wagami deskryptorów szczegółowych względem deskryptorów kwerendy.

Rozdział 2. Modyfikacja przestrzeni dokumentów

2.1. Cele modyfikacji

Z uzyskanych w ostatnich latach ocen automatycznych systemów wyszukiwania informacji wynika, że efektywność wyszukiwania osiągnięta, obecnie i w możliwej do przewidzenia przyszłości, nie spełnia oczekiwań większości potencjalnych użytkowników. Co więcej, postęp teoretyczny w analizie językowej i organizacji danych zapowiada względnie nieznaczną poprawę w przyszłości.

Najbardziej widoczną poprawę efektywności wyszukiwania można otrzymać za pomocą adaptacyjnych metod współdziałania użytkownika z systemem, pozwalających na reorganizację danych na podstawie informacji otrzymanej od użytkownika.

Jedną z technik polepszenia efektywności automatycznego wyszukiwania informacji jest modyfikacja kwerend oparta na subiektywnych ocenach relewancji dokumentów wcześniej odnalezionych i dostarczonych użytkownikom. Dla każdej otrzymanej kwerendy system dokonuje najpierw wyszukiwania wstępnego, zaś małą część odnalezionych dokumentów przedstawia użytkownikowi. Użytkownik bada otrzymaną odpowiedź, klasyfikując każdy dokument jako relewantny lub nierelewantny. Oceny relewancji wracają później do systemu, który na ich podstawie dokonuje automatycznego przekształcenia pierwotnej kwerendy polegającego na „wzmocnieniu” (przez zwiększenie wag) deskryptorów występujących w dokumentach relewantnych, oraz „osłabieniu” deskryptorów znajdujących się w dokumentach nierelewantnych. Wynikiem tego procesu jest zmodyfikowana kwerenda, która jest bardziej przystosowana do podzbioru dokumentów relewantnych i bardziej „odległa”

od zbioru nierelevantnego. Nowotworzona kwerenda zostaje przekazana systemowi, który na jej podstawie przystępuje do drugiego cyklu wyszukiwania. Cykl ten może doprowadzić do odnalezienia dalszych dokumentów relevantnych, lub przynajmniej pozwoli uzyskać wyższe korelacje pozycji relevantnych z kwerendą zmodyfikowaną niż z kwerendą źródłową. Nowoodnalezione pozycje relevantne mogą być znów ocenione przez użytkownika, a odpowiednie oceny użyte do utworzenia następnej wersji kwerendy. Proces ten można powtarzać aż do momentu pełnego usatysfakcjonowania użytkownika.

Zastosowania i wyniki różnych badań opisanej wyżej modyfikacji kwerend można znaleźć w kilku artykułach zamieszczonych w pracy zbiorowej [6], autorów: J.J.Rocchio jr (pp. 328 - 338), G.Saltona (pp. 339 - 352), E.Ide (pp. 353 - 373), Y.K.Changa, C.Cirillo, J.Razona (pp. 374 - 389), A.Borodina, L.Keera, F. Lewisa (pp. 391 - 413).

Metody zwiększenia efektywności wyszukiwania przez modyfikację kwerend mają pewne wady. Podstawową wadą tych metod jest to, że modyfikacje kwerend nie mają stałego charakteru, nawet jeśli mogą one poprawić obsługę konkretnego zapytania, to jednak nie mogą ulepszyć samego systemu. Wprowadzenie innych metod, których zadaniem byłoby realizowanie pożytecznych stałych zmian w systemie może przynieść stałe polepszenie efektywności wyszukiwania. Jediną stałą strukturą w systemie jest przestrzeń wektorów opisów dokumentów, która powinna być przedmiotem modyfikacji.

Zmiany w przestrzeni wektorów opisów dokumentów mogą mieć wielorakie przeznaczenie.

1. Kwerendom jednolitej pod względem specyfikacji zawodowej grupy użytkowników odpowiadającą podobne zbiory dokumentów. Wykorzystanie tego faktu może uprościć wyszukiwanie dokumentów re-

lewantnych dla podobnych kwerend stawianych w przyszłości.

2. Mimo standaryzacji, słownictwo wielu dziedzin naukowych nie pozostaje stałe, zmienia się wraz z rozwojem nowych dziedzin, napływem personelu i innymi czynnikami. W rezultacie po pewnym czasie (np. po 5 latach) może się okazać, że opisy dokumentów straciły na swej aktualności. Metoda modyfikacji polegająca na wprowadzaniu nowych pojęć do wektorów opisów na podstawie zgłaszanych kwerend może zapewnić stałą aktualizację słownictwa występującego w tych wektorach.

3. Wagi deskryptorów w nowych wektorach są ustalane zgodnie z częstością występowania pojęć w streszczeniach danych dokumentów. Ustalenie wag sprowadza się więc do prostego procesu przeliczenia. Jednakże założenie, że najważniejsze są pojęcia najczęściej powtarzające się w streszczeniu nie zawsze jest właściwe. Aktualizacja wag wielu pojęć dokumentu na podstawie subiektywnej opinii użytkownika pozwala utworzyć wektory opisów dokumentów lepiej odzwierciedlające rzeczywistą treść.

2.2. Sposoby modyfikacji przestrzeni dokumentów

Istnieje wiele metod usprawniania wyszukiwania za pomocą modyfikacji wektorów opisów dokumentów. Poniższe metody zostały zaproponowane przez E. Ide w pracy ([6], pp. 410 - 413) oraz T.L. Brauena w pracy ([6], pp. 482 - 483) przy badaniach zbiorów Cranfield.

Po otrzymaniu odpowiedzi na swoją kwerendę użytkownik ocenia, które z uzyskanych dokumentów są relewantne, a które nie. Oceny użytkownika są przesyłane do systemu na zasadzie sprzężenia zwrotnego. Zgodnie z tymi wskazówkami użytkownika przeprowa-

dzona jest modyfikacja dokumentów. Modyfikację wektora opisu dokumentu można przeprowadzić według następujących metod:

A. Jeśli użytkownik oceni dokument należący do odpowiedzi jako relewantny względem jego kwerendy $\mathcal{V}$ to wektor opisu dokumentu $\mathcal{t}$ zostanie przekształcony przy pomocy wzoru:

$$\mathcal{t} := (1 - \lambda) \mathcal{t} + \lambda \mathcal{V} \quad (2.2.1.a)$$

W przypadkach gdy dokument oceniony jest jako nierelewantny, to wektor opisu dokumentu jest przekształcony według wzoru:

$$\mathcal{t} := (1 + \lambda) \mathcal{t} - \lambda \mathcal{V} \quad (2.2.1.b)$$

B. W tej metodzie do modyfikacji wektorów opisów relewantnych i nierelewantnych dokumentów stosujemy odpowiednio wzory:

$$\mathcal{t} := (1 - \lambda) \mathcal{t} + \lambda \mathcal{V} \quad (2.2.2.a)$$

oraz

$$\mathcal{t} := (1 - \lambda) \mathcal{t} - \lambda \mathcal{V} \quad (2.2.2.b)$$

C. W tej metodzie modyfikujemy tylko wektory opisów relewantnych dokumentów stosując wzór:

$$\mathcal{t} := (1 - \lambda) \mathcal{t} - \lambda \mathcal{V} \quad (2.2.3)$$

Wektory opisów dokumentów nierelewantnych pozostawimy bez zmiany.

Parametr λ występujący we wszystkich wzorach jest parametrem charakteryzującym szybkości zmian wag w opisach dokumentów w czasie modyfikacji. Przeprowadzone badania przez Brauera wykazały, że najlepsze rezultaty można otrzymać dla $\lambda \leq 0.2$. Wyższe wartości λ prowadzą do „niezamierzonych” wyników.

Powyższa postać wzorów modyfikacji wektorów opisów dokumentów nie uwzględnia relacji hierarchii. Wzory te zostały podane w celu wyjaśnienia postaci wzorów (2.2.4) – (2.4.6) dokładnie

pokazujących zmiany wag w czasie modyfikacji wektorów opisów z uwzględnieniem relacji hierarchii.

Niech $\mathcal{T} = \{(t_1, w_1), (t_2, w_2), \dots, (t_n, w_n)\}$ będzie opisem dokumentu a $\mathcal{V} = \{(r_1, v_1), (r_2, v_2), \dots, (r_m, v_m)\}$ kwadrantą. Zbiór $L_1^*(r_i) = \{t_j : r_i R t_j \wedge 1 \leq j \leq n\}$ dla $1 \leq i \leq m$.

Wzór (2.2.1.a) możemy zapisać w postaci

$$\forall_{1 \leq j \leq n} w_j := \begin{cases} (1 - \lambda) w_j + \frac{w_j}{\sum_{t_k \in L_1^*(r_i)} w_k} \lambda v_i & \text{jeśli } r_i R t_j \\ (1 - \lambda) w_j & \text{w p.p.} \end{cases} \quad (2.2.4)$$

Wzór (2.2.1.b) przyjmuje postać:

$$\forall_{1 \leq j \leq n} w_j := \begin{cases} \left| (1 + \lambda) w_j - \frac{w_j}{\sum_{t_k \in L_1^*(r_i)} w_k} \lambda v_i \right| & \text{jeśli } r_i R t_j \\ (1 + \lambda) w_j & \text{w p.p.} \end{cases} \quad (2.2.5)$$

oraz wzór (2.2.2.b)

$$\forall_{1 \leq j \leq n} w_j := \begin{cases} \left| (1 - \lambda) w_j - \frac{w_j}{\sum_{t_k \in L_1^*(r_i)} w_k} \lambda v_i \right| & \text{jeśli } r_i R t_j \\ (1 - \lambda) w_j & \text{w p.p.} \end{cases} \quad (2.2.6)$$

Aby uniknąć wartości ujemnych wag, które mogłyby się niekiedy pojawić, we wzorach (2.2.5) i (2.2.6) obliczamy wartości bezwzględne różnic. Po obliczeniu wag według wzoru (2.2.6), lub jeśli we wzorze (2.2.5) przynajmniej jedna różnica przyjęła

wartość ujemną, dokonujemy normalizacji wag, tj.

$$\forall 1 \leq j \leq n \quad w_j := \frac{w_j}{\sum_{k=1}^n w_k}$$

Modyfikacja opisów dokumentów przeprowadzona metodami A, B i C pozwala na zmianę wag i usuwanie deskryptorów, których wagi zmalały do zera. W powyższych metodach nie jest możliwe dołączenie nowych deskryptorów do opisów dokumentów ze względu na konieczność spełnienia warunku $\sigma \leq \frac{1}{2}$. Wadą powyższych metod usuwa się poprzez dodatkową modyfikację wektorów dokumentów. Przy przesyłaniu do systemu oceny o relewantności dokumentu użytkownik może podać nowy deskryptor z wagą, który według niego powinien wystąpić w opisie tego dokumentu. Dołączony deskryptor nie może być ani ogólniejszy ani szczegółowszy od wszystkich deskryptorów występujących w opisie dokumentu.

Modyfikacja wektora dokumentu przebiega według wzoru:

$$\exists 1 \leq i \leq n \quad w_i := \lambda w' \quad (2.2.7)$$

gdzie w' jest wagą nowego deskryptora.

Po dołączeniu deskryptora z wagą dokonuje się normalizacji opisu, tj.

$$\forall 1 \leq i \leq n \quad w_i := \frac{w_i}{1 + \lambda w'}$$

Użytkownicy wymagający małych dokładności odpowiedzi powinni mieć mniejszy wpływ na modyfikację opisów dokumentów niż użytkownicy wymagający dużych dokładności odpowiedzi. W odpowiedziach z małą dokładnością znajdują się dokumenty, których stopień podobieństwa do kwerend jest mały, wtedy wzrasta możliwość nieprawidłowej klasyfikacji wśród różnych orzekających o relewantności

dokumentu. Dlatego też we wzorach (2.2.4) - (2.2.6) za parametr można wziąć $\lambda = 0.1 \cdot (c + 1)$. Użytkownicy wymagający dużych dokładności c odpowiedzi mają wtedy większy wpływ na modyfikację opisów dokumentów.

Bardzo prosta postać wzorów (2.2.4) - (2.2.6) pozwala na łatwe stosowanie w praktyce podanych wyżej metod modyfikacji wektorów dokumentów.

Modyfikacja wektorów opisów relewantnych (nierелеwantnych) dokumentów ma na celu zwiększenie (zmniejszenie) stopnia podobieństwa μ opisu dokumentu $\mathcal{D}$ do kwerendy $\mathcal{Q}$.

W przykładzie 2.2.1 pokażemy jak szybko wzrasta stopień podobieństwa pomiędzy kwerendą i opisem dokumentu.

Przykład 2.2.1

Niech

$$\mathcal{Q}_1 = \{(r_1, .2), (r_2, .3), (r_3, .3), (r_4, .1), (r_5, .1)\}$$

$$\mathcal{D}_1 = \{(t_1, .1), (t_2, .1), (t_3, .1), (t_4, .2), (t_5, .2), (t_6, .1), (t_7, .1), (t_8, .1)\}$$

$$\mathcal{Q}_2 = \{(r'_1, .2), (r'_2, .6), (r'_3, .1), (r'_4, .1)\}$$

$$\mathcal{D}_2 = \{(t'_1, .1), (t'_2, .1), (t'_3, .1), (t'_4, .2), (t'_5, .2), (t'_6, .1), (t'_7, .1), (t'_8, .1)\}$$

$$\mathcal{Q}_3 = \{(r''_1, .2), (r''_2, .1), (r''_3, .2), (r''_4, .5)\}$$

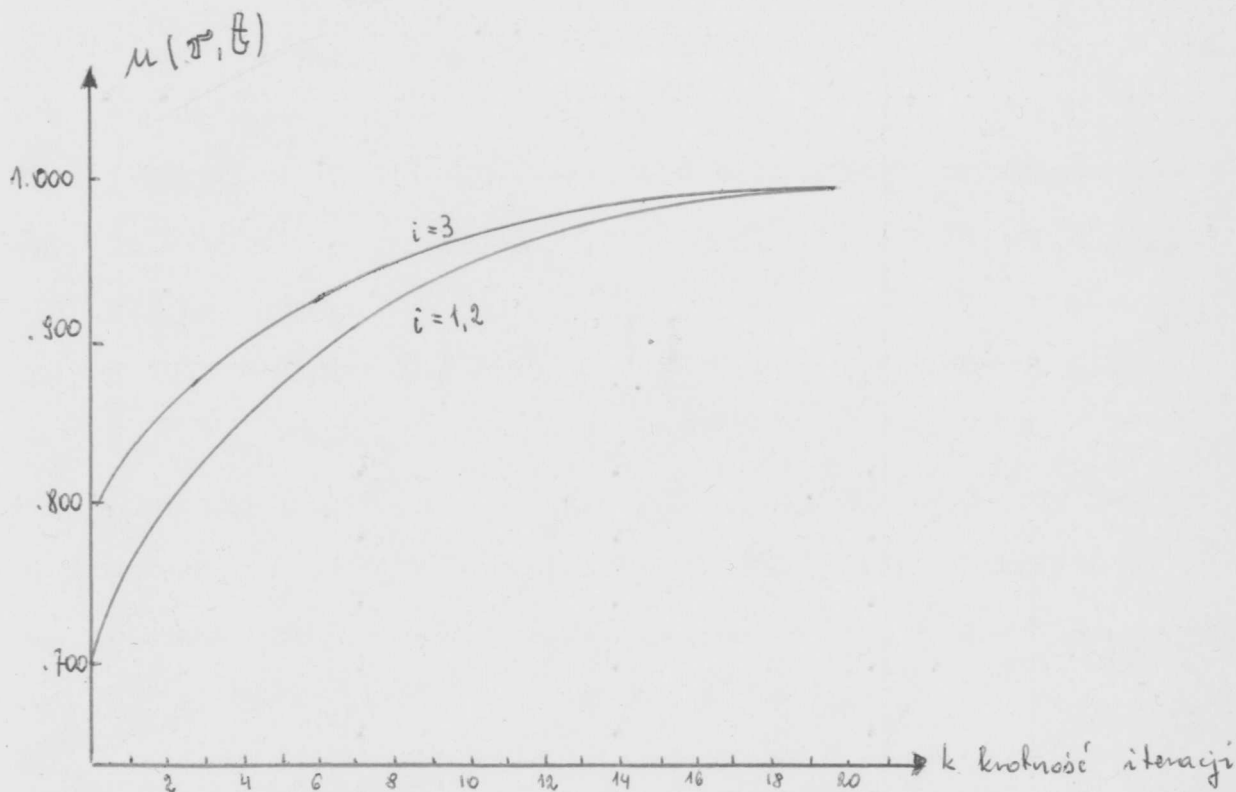
$$\mathcal{D}_3 = \{(t''_1, .1), (t''_2, .1), (t''_3, .1), (t''_4, .1), (t''_5, .2), (t''_6, .1), (t''_7, .1), (t''_8, .2)\}$$

będą odpowiednio kwerendami i opisami dokumentów. Strzałki pomiędzy deskryptorami kwerend i opisów wskazują relację hierarchii.

Dokumenty o opisach t_1, t_2, t_3 ocenione są jako relewantne względem kwerend v_1, v_2, v_3 . Modyfikację przeprowadzono według wzoru (2.2.4), za wartość parametru λ przyjęto 0.16. Poniższa tabelka przedstawia $\mu(v_i, t_i)$, w kolejnych iteracjach. Na wykresie przedstawiono zależność miary podobieństwa μ od liczby dokonanych modyfikacji opisu dokumentu.

$\mu(v_i, t_i)$

k	i = 1	i = 2	i = 3
0	.700	.700	.800
1	.748	.748	.821
2	.787	.787	.859
3	.823	.823	.881
4	.850	.850	.900
5	.874	.874	.916
6	.895	.895	.930
7	.910	.910	.941
8	.925	.925	.950
9	.937	.937	.958
10	.949	.949	.965
11	.955	.955	.969
12	.964	.964	.975
13	.970	.970	.979
14	.973	.973	.983
15	.979	.979	.985
16	.982	.982	.988
17	.985	.985	.990
18	.988	.988	.991
19	.988	.988	.993
20	.991	.991	.994



Wykres 2.2.1. Zależność miary μ od krotności iteracji

Po 20 iteracjach opisy t_1, t_2, t_3 przyjmują postać

$$t_1' = \{(t_1, .003), (t_2, .003), (t_3, .197), (t_4, .297), (t_5, .297), (t_6, .100), (t_7, .100), (t_8, .003)\}$$

$$t_2' = \{(t_1', .003), (t_2', .003), (t_3', .197), (t_4', .297), (t_5, .297), (t_6', .100), (t_7', .100), (t_8', .003)\}$$

$$t_3' = \{(t_1'', .100), (t_2'', .100), (t_3'', .100), (t_4'', .197), (t_5'', .248), (t_6'', .124), (t_7'', .124), (t_8'', .006)\}$$

Opisy t_1', t_2' są jednakowe, ponieważ deskryptory z wagami $(r_2, .3), (r_3, .3)$ w kwerendzie $\mathcal{V}_1$ zostały zastąpione przez $(r_2, .6)$ w kwerendzie $\mathcal{V}_2$.

Dokument relewantny w wyniku modyfikacji opisu zostanie przesunięty do kwerendy a nierelevantny odsunięty od kwerendy.

Przykład 2.2.2.

Niech $\mathcal{Q} = \{(t_1, 0.2), (t_2, 0.8)\}$ będzie kwerendą a

$\mathcal{B} = \{(t_1, 0.7), (t_2, 0.3)\}$ opisem dokumentu.

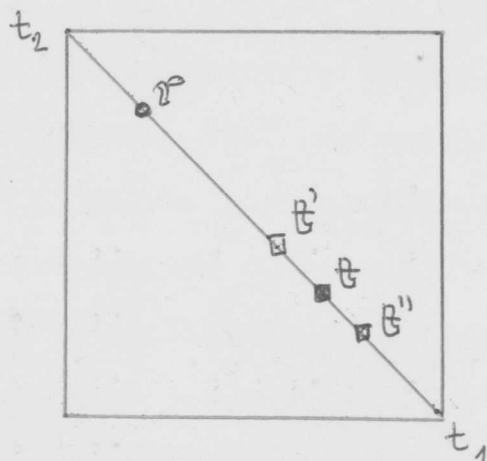
Miara podobieństwa $\mu(\mathcal{Q}, \mathcal{B}) = 0.5$.

Jeśli dokument zostanie oceniony jako relewantny względem kwerendy, to w wyniku modyfikacji (2.2.4) dla $\lambda = 0.2$ opis dokumentu przyjmie postać

$$\mathcal{B}' = \{(t_1, 0.7 \cdot 0.8 + 0.2 \cdot 0.2), (t_2, 0.3 \cdot 0.8 + 0.8 \cdot 0.2)\} = \\ = \{(t_1, 0.6), (t_2, 0.4)\} \text{ oraz } \mu(\mathcal{Q}, \mathcal{B}') = 0.6.$$

W przypadku gdy dokument o opisie $\mathcal{B}$ zostanie oceniony jako nierelevantny, to w wyniku modyfikacji (2.2.5) otrzymamy opis

$$\mathcal{B}'' = \{(t_1, 0.7 \cdot 1.2 - 0.2 \cdot 0.2), (t_2, 0.3 \cdot 1.2 - 0.8 \cdot 0.2)\} = \\ = \{(t_1, 0.8), (t_2, 0.2)\} \text{ i } \mu(\mathcal{Q}, \mathcal{B}'') = 0.4$$



Na przekątnej kwadratu zaznaczono położenie kwerendy $\mathcal{Q}$ i opisów dokumentów $\mathcal{B}$, $\mathcal{B}'$, $\mathcal{B}''$.

2.3. Wnioski i zalecenia

W czasie eksploatacji systemu oceny użytkowników o relewantności dokumentów należących do odpowiedzi są gromadzone w postaci modyfikacji wag dyskryptów w opisach dokumentów. Po pewnym okresie pracy systemu niektóre deskryptory tracą swoje poprzednią ważność w opisach dokumentów przez zmniejszenie wag. Deskryptory, których wagi zmalały do zera powinny być usunięte z opisów dokumentów. Możliwe jest również pojawienie się nowych deskryptów, które powinny być włączone do opisów dokumentów. Stała postać list inwersyjnych po pewnym czasie nie odzwierciedla nowych postaci opisów dokumentów. Dlatego też co pewien czas będą aktualizowane listy inwersyjne aby nie obniżyć efektywności systemu.

Ze zmodyfikowanej przestrzeni dokumentów korzyści czerpać będą nie tylko ci użytkownicy, którzy brali udział w modyfikowaniu opisów dokumentów. Nowi użytkownicy korzystać będą z „doświadczeń” zebranych podczas poprzednich wyszukiwań.

Uwagi użytkowników o opisach dokumentów, gromadzonych w systemie w czasie aktualizacji opisów, będą umieszczone na listach inwersyjnych podczas reorganizacji list. Ze zmodyfikowanych opisów dokumentów użytkownicy będą mogli skorzystać dopiero po reorganizacji list, ponieważ w czasie modyfikacji opisów dokumentów listy inwersyjne pozostają bez zmiany.

Jeżeli reorganizację list inwersyjnych będziemy przeprowadzać po każdej kwerendzie, to użytkownicy będą mogli korzystać z bieżącej modyfikacji przestrzeni opisów dokumentów. Powiemy wtedy, że system jest bardzo „wrażliwy” na modyfikacje opisów dokumentów należących do odpowiedzi na poszczególne kwerendy. Wysoka „wrażliwość” systemu na modyfikację opisów wiąże się z wy-

sokimi kosztami eksploatacji systemu.

Jeśli reorganizację list będziemy przeprowadzać rzadziej, to system będzie mniej „wrażliwy” na pojedyncze modyfikacje opisów dokumentów. Wtedy system uwzględnia „uśrednione” informacje od użytkowników i nie jest „wrażliwy” na pomyłki w ocenie dokumentów przez pojedynczych użytkowników oraz koszt eksploatacji takiego systemu może być kilkakrotnie niższy niż w systemach, w których reorganizacji list dokonuje się po każdej kwerendzie.

W systemie można dokonać wyboru częstotliwości przeprowadzanie reorganizacji list inwersyjnych. W jednym z następnych podrozdziałów zostaną podane uwagi o kosztach modyfikacji opisów i reorganizacji list inwersyjnych.

Wyszukiwanie informacji metodą pełnego przeglądu dokumentów jest bardzo „wrażliwe” na pojedyncze modyfikacje opisów, ale jest bardzo kosztowne. Zadaniem może być zbadanie „wrażliwości” na modyfikacje opisów dokumentów innych metod wyszukiwania.

W czasie oceniania dokumentów należących do odpowiedzi użytkownicy wydają orzeczenia o dokumencie w postaci binarnej (relevantny (R) lub nirelevantny (N)). W przypadkach granicznych użytkownicy mogą mieć wątpliwości w ocenach. Można wprowadzić stopnie relewantności dokumentów przyjmujących wartości liczbowe z przedziału $[-1, 1]$. Dla dokumentów relewantnych stopnie relewantności byłyby bliskie jedności, dla nirelevantnych bliskie -1 . Stopnie bliskie zera określałyby dokumenty, które trudno ocenić jako relewantne lub nirelevantne. Parametr λ występujący we wzorach modyfikujących opisy dokumentów może zmieniać się w zależności od stopni relewancji dokumentów. Następnym zadaniem może być zbadanie zachowania się

systemu przy zmieniających się parametrach) dla każdego
modyfikowanego opisu dokumentu.

Rozdział 3. Symulacja procesu wyszukiwania informacji.

W rozdziale następnym pokażemy, że modyfikacja przestrzeni opisów dokumentów przynosi korzyści użytkownikom w postaci zwiększenia efektywności systemu. Analityczne badanie zmian efektywności jest bardzo trudne do wykonania, dlatego też opierać będziemy się na badaniach empirycznych. W tym przypadku są dwie możliwości:

1) Zbadanie zmian w przestrzeni dokumentów istniejących systemów.

2) Zbadanie zmian w przestrzeni dokumentów przy pomocy symulacji pracy systemu.

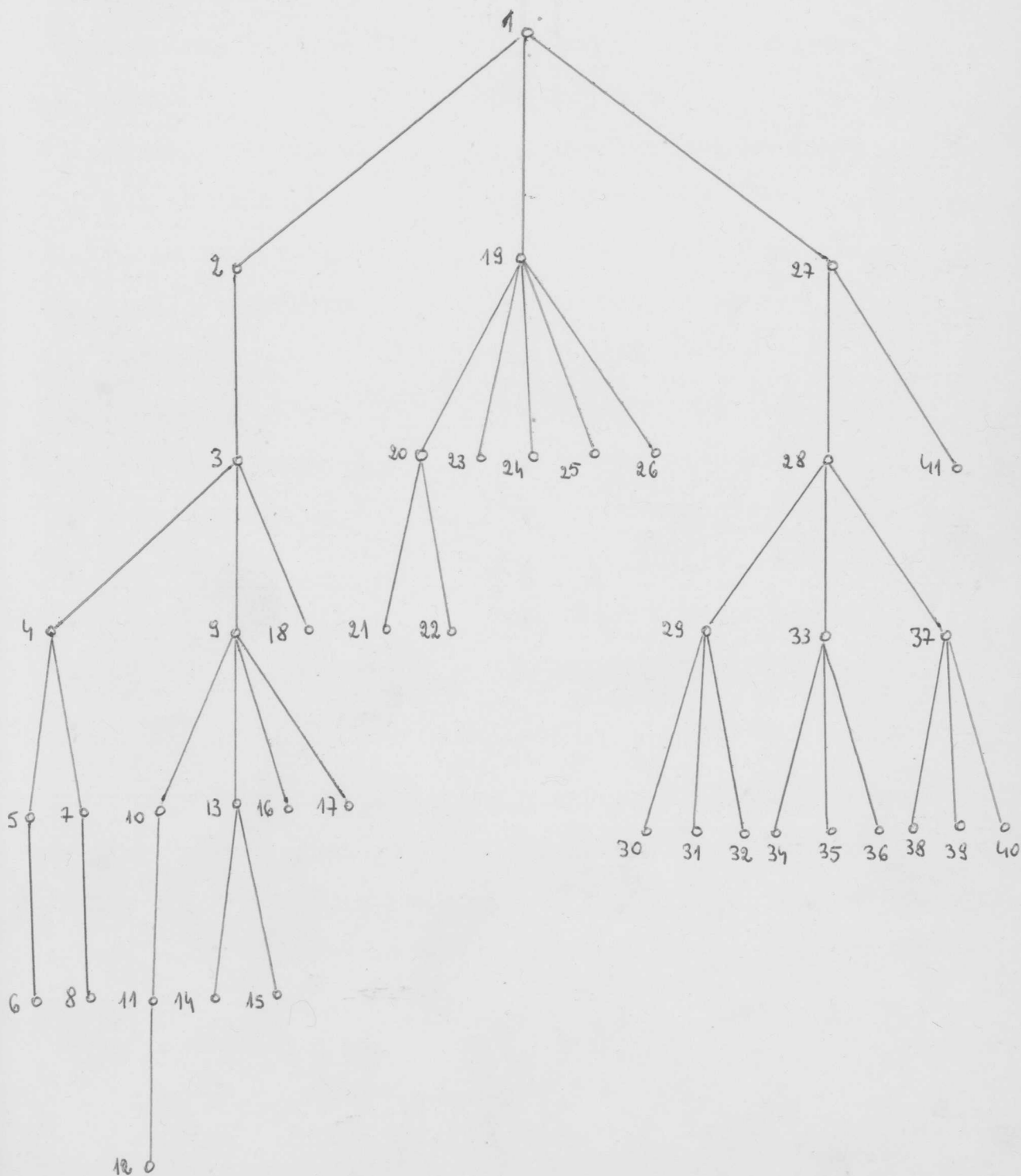
Badania przeprowadzać będziemy poprzez symulację pracy systemu, ponieważ w praktyce jeszcze nie używa się systemów wyszukiwania informacji spełniających założenia modelu podane w rozdziale 1.

W pierwszej części rozdziału podana jest postać tezauryusa wybranego do badań symulacyjnych, reprezentacja graficzna relacji hierarchii zachodząca pomiędzy deskryptorami, podział zbioru deskryptorów na podzbiory T_i wyznaczające osie układu współrzędnych przestrzeni opisów dokumentów oraz macierz INDEX, która dla dowolnego opisu wyznacza odpowiednie kostki przestrzeni opisów dokumentów.

W drugiej części podany jest opis symulacji podejmowania decyzji o relewantności dokumentów.

3.1. Realizacja maszynowa modelu systemu.

W celu sprawdzenia zachowania się systemu podczas modyfikacji dokumentów zbudowano przykładowy system wyszukiwania informacji. Zbiór deskryptorów składa się z 41 elementów ; graficznie tezaurus T można przedstawić jak następuje:



Lista deskryptorów

- | | |
|---------------------------|-----------------------------|
| 1. informatyka ETO | 22. system wielodostępny |
| 2. maszyna cyfrowa | 23. języki programowania |
| 3. organizacja m.c. | 24. programy usługowe |
| 4. urządzenia peryferyjne | 25. programy biblioteczne |
| 5. urządzenia wejściowe | 26. programy tłumaczące |
| 6. czytnik kart | 27. zastosowanie ETO |
| 7. urządzenia wyjścia | 28. metody zastosowań |
| 8. drukarka | 29. modele obliczeniowe |
| 9. pamięć | 30. algebra |
| 10. pamięć operacyjna | 31. analiza matematyczna |
| 11. pamięć rdzeniowa | 32. statystyka |
| 12. rdzeń ferrytowa | 33. modele morfologiczne |
| 13. pamięć zewnętrzna | 34. kombinatoryka |
| 14. pamięć taśmowa | 35. rachunkowość |
| 15. pamięć dyskowa | 36. wyszukiwanie informacji |
| 16. pojemność pamięci | 37. modele semantyczne |
| 17. czas dostępu | 38. rozpoznawanie postaci |
| 18. procesor | 39. heurystyka |
| 19. programowanie | 40. analiza semantyczna |
| 20. system operacyjny | 41. dziedziny zastosowań |
| 21. system tradycyjny | |

Część tezaursusa dotycząca maszyn cyfrowych i oprogramowania zaczerpnięto z pracy [4], natomiast część dotycząca zastosowań informatyki z pracy ([10], rozdz. 10). W dalszej części każdy deskryptor utożsamiać będziemy z jego numerem na liście.

Deskryptory ze zbioru T_0 podzielono na 24 podzbiory T_i spełniające warunki (1.3.1) - (1.3.4).

Do zbioru T_i ($1 \leq i \leq 24$) należą deskryptory o numerach:

$T_1 = \{6,5\}$	$T_{13} = \{25\}$
$T_2 = \{8,7,4\}$	$T_{14} = \{26,19,1\}$
$T_3 = \{12,11,10\}$	$T_{15} = \{30,29,28\}$
$T_4 = \{14,13\}$	$T_{16} = \{31\}$
$T_5 = \{15\}$	$T_{17} = \{32\}$
$T_6 = \{16\}$	$T_{18} = \{34\}$
$T_7 = \{17,9\}$	$T_{19} = \{35\}$
$T_8 = \{18,3,2\}$	$T_{20} = \{36,33\}$
$T_9 = \{21\}$	$T_{21} = \{38\}$
$T_{10} = \{22,20\}$	$T_{22} = \{39\}$
$T_{11} = \{23\}$	$T_{23} = \{40,37\}$
$T_{12} = \{24\}$	$T_{24} = \{41,27\}$

W wektorze TEZAURUS została zapamiętana relacja hierarchii pomiędzy deskryptorami. Dla i - tego deskryptora element TEZAURUS(i) wskazuje numer ostatniego deskryptora, pomiędzy którymi zachodzi relacja hierarchii. Chcąc sprawdzić czy zachodzi relacja $t_i R t_j$ pomiędzy dowolnymi deskryptorami t_i, t_j wystarczy sprawdzić warunek :

$$i \leq j \leq \text{TEZAUR}(i)$$

Elementy wektora TEZAURUS przyjmują wartości :

i	TEZAURUS(i)
1	41
2	18
3	18
4	8
5	6
6	6
7	8
8	8
9	17
10	12
11	12
12	12
13	15
14	14
15	15
16	16
17	17
18	18
19	26
20	22
21	21

i	TEZAURUS(i)
22	22
23	23
24	24
25	25
26	26
27	41
28	40
29	32
30	30
31	31
32	32
33	36
34	34
35	35
36	36
37	40
38	38
39	39
40	40
41	41

Przy pomocy tablicy INDEX można łatwo wyznaczyć zbiór T_i i miejsce w tym zbiorze dla dowolnego deskryptora ze zbioru T_0 .

Dla j - tego deskryptora $INDEX(j,1) = i$, wskazuje numer zbioru T_i , $INDEX(j,2)$ miejsce w zbiorze T_i .

Elementy tablicy INDEX przyjmują wartości:

j	INDEX(j,1)	INDEX(j,2)	j	INDEX(j,1)	INDEX(j,2)
1	14	3	22	10	1
2	8	3	23	11	1
3	8	2	24	12	1
4	2	3	25	13	1
5	11	2	26	14	1
6	11	1	27	24	2
7	2	2	28	15	3
8	2	1	29	15	2
9	7	2	30	15	1
10	3	3	31	16	1
11	3	2	32	17	1
12	3	1	33	20	2
13	4	2	34	18	1
14	4	1	35	19	1
15	5	1	36	20	1
16	6	1	37	23	2
17	7	1	38	21	1
18	8	1	39	22	1
19	14	2	40	23	1
20	10	2	41	24	1
21	9	1			

Teksty deskryptorów są zapamiętane znakowo w tablicy
DESKRYPTOR w kolejności alfabetycznej.

DESKRYPTOR(i,1), DESKRYPTOR(i,2) zawierają nazwy deskryptorów
(do 20 znaków). DESKRYPTOR(i,3) jest wskaźnikiem pozycji
i-tego deskryptora w wektorze TEZAURUS i tablicy INDEX

i	DESKRYPTOR(i,1), DESKRYPTOR(i,2)	DESKRYPTOR(i,3)
1	algebra	30
2	analiza matematyczna	31
3	analiza semantyczna	40
4	czas dostępu	17
5	czytnik kart	6
6	drukarka	8
7	dziedziny zastosowań	41
8	heurystyka	39
9	informatyka ETO	1
10	języki programowania	23
11	kombinatoryka	34
12	maszyna cyfrowa	2
13	metody zastosowań	28
14	modele morfologiczne	33
15	modele obliczeniowe	29
16	modele semantyczne	37
17	oprogramowanie	19
18	organizacja m.c.	3
19	pamięć	9
20	pamięć dyskowa	15
21	pamięć operacyjna	10

i	DESKRYPTOR(i,1), DESKRYPTOR(i,2)	DESKRYPTOR(i,3)
22	pamięć rdzeniowa	11
23	pamięć taśmowa	14
24	pamięć zewnętrzna	13
25	pojemność pamięci	16
26	procesor	18
27	programy biblioteczne	25
28	programy tłumaczące	26
29	programy usługowe	24
30	rachunkowość	35
31	rdzeń ferrytowy	12
32	rozpoznawanie postaci	38
33	statystyka	32
34	system operacyjny	20
35	system tradycyjny	21
36	system wielodostępny	22
37	urządzenia peryferyjne	4
38	urządzenia wejściowe	5
39	urządzenia wyjściowe	7
40	wyszukiwanie informacji	36
41	zastosowanie ETO	27

Dla dowolnego deskryptora z kwerendy tablicę DESKRYPTOR wykorzystujemy do znalezienia jego pozycji w wektorze TEZAUROS i tablicy INDEX.

3.2. Sylumacja ocen o relewancji dokumentów

Zbiór opisów dokumentów z przestrzeni PD oznaczać będziemy przez PD_A . Każdy dokument, którego opis należy do PD_A , ma swój obraz będący opisem idealnym dokumentu. Celem modyfikacji opisów dokumentów jest przybliżenie opisów dokumentów ze zbioru PD_A do opisów idealnych. Zbiór opisów idealnych oznaczać będziemy przez PD_I .

Dany dokument o opisie $t \in PD_A$ należący do odpowiedzi na kwerendę $\mathcal{V}$ jest porównywany ze swoim obrazem ze zbioru PD_I w czasie podejmowania decyzji o relewancji dokumentu względem kwerendy $\mathcal{V}$. Jeżeli kwerenda jest bardziej podobna (większy stopień podobieństwa) do obrazu dokumentu niż do samego opisu dokumentu to podejmuje się decyzję o relewantności dokumentu w celu przesunięcia dokumentu bliżej kwerendy (tym samym bliżej obrazu - dokumentu o opisie idealnym).

Niech t_A oznacza opis badanego dokumentu ze zbioru PD_A a t_I odpowiadający mu idealny opis dokumentu ze zbioru PD_I . Dokument o opisie t_A będziemy uważać za relewantny względem kwerendy $\mathcal{V}$ jeśli:

$$1) \mu(\mathcal{V}, t_I) \geq \mu(\mathcal{V}, t_A)$$

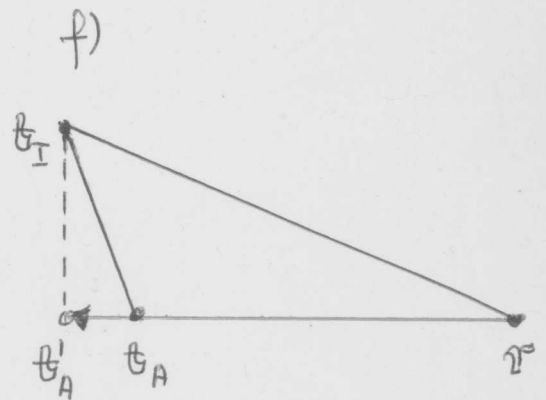
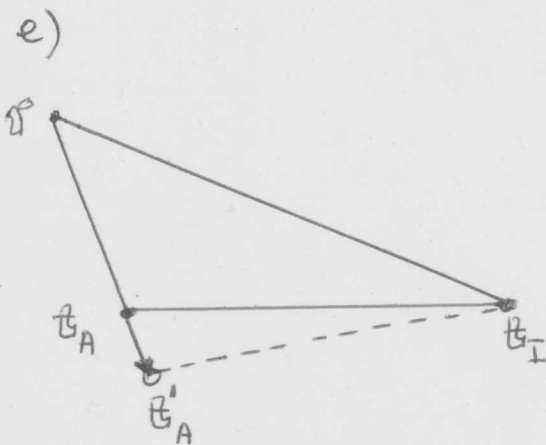
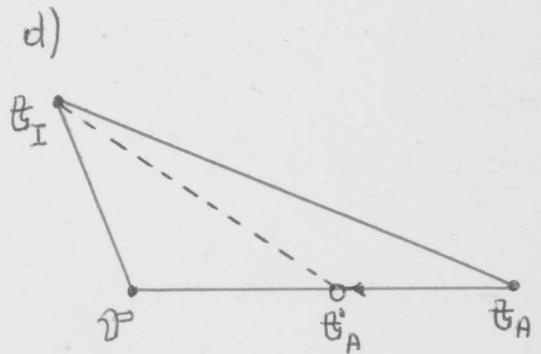
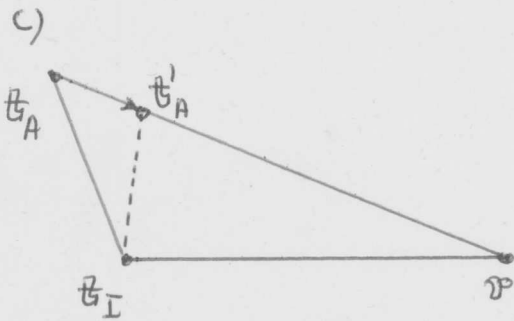
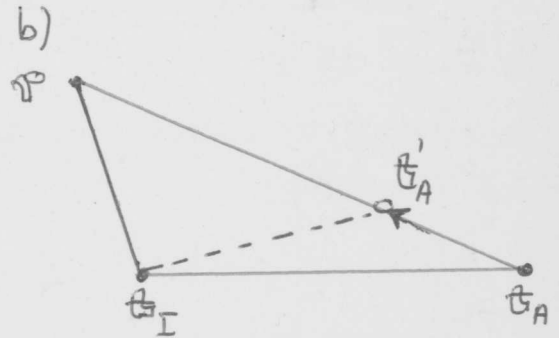
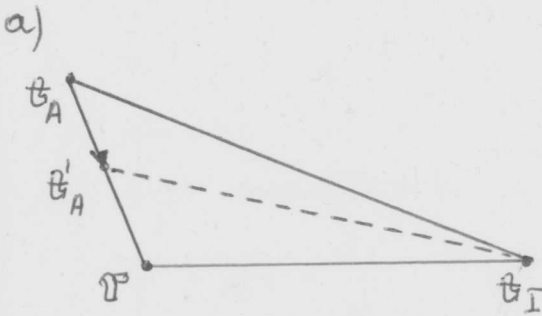
lub

$$2) \mu(\mathcal{V}, t_I) < \mu(\mathcal{V}, t_A) \wedge \mu(\mathcal{V}, t_I) \geq \mu(t_I, t_A)$$

W pozostałych przypadkach dokument o opisie t_A będziemy uważać za nirelewantny.

Powyższe warunki można zilustrować graficznie. Na rysunkach większy stopień podobieństwa pomiędzy kwerendą i opisem zazna -

czony jest krótszym odcinkiem.



Na rysunkach a, b, c i d dokumenty o opisach t_A zostaną ocenione jako relewantne względem kwerendy q , na rysunkach e i f dokumenty o opisach t_A zostaną ocenione jako nierelewantne. Dokumenty po modyfikacji zaznaczono na rysunkach przez t'_A . Stopień podobieństwa pomiędzy t_I i t'_A zaznaczono kreską przerywaną.

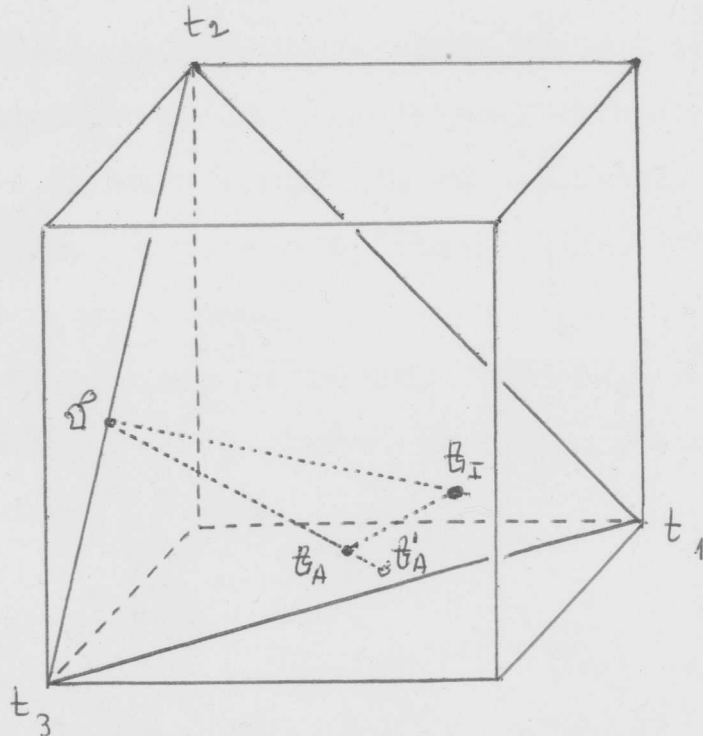
Przykład 2.4.1.

Niech $\mathcal{P} = \{(t_2, 0.4), (t_3, 0.6)\}$ będzie kwerendą,
 $\mathcal{E}_A = \{(t_1, 0.4), (t_2, 0.1), (t_3, 0.5)\}$, $\mathcal{E}_I = \{(t_1, 0.6), (t_2, 0.2), (t_3, 0.2)\}$
 będą odpowiednio opisami : dokumentu i odpowia-
 dającego mu dokumentu idealnego.

$$\mu(\mathcal{P}, \mathcal{E}_A) = 0.6 \quad \mu(\mathcal{P}, \mathcal{E}_I) = 0.4 \quad \mu(\mathcal{E}_I, \mathcal{E}_A) = 0.7$$

Na podstawie obliczonych stopni podobieństwa otrzymujemy,
 że dokument o opisie $\mathcal{E}_A$ jest nerelewantny względem
 kwerendy $\mathcal{P}$.

Modyfikując opis dokumentu według wzoru (2.25) z $\lambda = 0.2$
 otrzymujemy opis $\mathcal{E}'_A = \{(t_1, 0.48), (t_2, 0.04), (t_3, 0.48)\}$ i
 $\mu(\mathcal{E}'_A, \mathcal{E}_A) = 0.72$



Na przekątnej kostki jednostkowej przestrzeni dokumentów zazna-
 czono położenie kwerendy i dokumentów o opisach $\mathcal{E}_I$, $\mathcal{E}_A$, $\mathcal{E}'_A$.

3.3. Uwagi o przyjętych parametrach modelu.

Przestrzeń dokumentów PD zawiera 150 dokumentów. Do modyfikacji tej przestrzeni użyto 50 kwerend. Odpowiedzi na te kwerendy nie biorą później udziału przy badaniach efektywności modyfikacji. W celu zbadania efektywności modyfikacji systemu porównuje się odpowiedzi na 15 kwerend przed modyfikacją i po modyfikacji przestrzeni.

Kwerendy użyte do porównywania efektywności nie brały udziału w modyfikacji przestrzeni PD, dlatego też różnice w wydajności reprezentują miarę wartości tej modyfikacji.

Dla 50 kwerend biorących udział w modyfikacji dokładność odpowiedzi wynosi $c = 0.25$, tzn. modyfikuje się te opisy dokumentów dla których $\mu(\mathcal{V}, \mathcal{t}) \geq 0.25$. Dla 15 odpowiedzi, które porównywać będziemy przed i po modyfikacji, dokładność wynosi $c = 0.001$. Wartość parametru λ we wzorach (2.2.4) - (2.2.6) wynosi 0.14. W czasie tworzenia list inwersyjnych z opisów będą usuwane deskryptory, których wagi są mniejsze od 0.04.

Obliczenia wykonano na maszynie cyfrowej CDC 6400.

W dodatku podano treści programów realizujących poszczególne zadania systemu.

Rozdział 4. Analiza wyników symulacji pracy systemu wyszukiwania informacji

W rozdziale tym zostaną podane definicje współczynników korelacji porównujących opisy dokumentów przed i po modyfikacji z opisami idealnymi dokumentów należących do odpowiedzi na tę samą kwerendę. Zadaniem tych współczynników jest pokazanie w jakim stopniu „korelują” między sobą odpowiedzi. Zbadamy czy różnice pomiędzy odpowiednimi współczynnikami są przypadkowe czy też istotne. W końcu rozdziału zajmiemy się zbadaniem zależności pomiędzy współczynnikami korelacji a czasem modyfikowania opisów dokumentów.

4.1. Sposoby obliczania współczynników korelacji

Niech F będzie funkcją modyfikującą elementy ze zbioru PD_A należące do odpowiedzi na kwerendę $\mathcal{V}$. Dla każdego elementu ciągu $\mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_n$ przez $F_{\mathcal{V}_i}$ oznaczać będziemy przekształcenie takie, że

$$F_{\mathcal{V}_i} : PD_{i-1} \rightarrow PD_i \quad \text{dla } i = 1, 2, \dots, n$$

gdzie $PD_0 = PD_A$

Ponieważ przekształcenie zbiorów PD_i jest procesem iteracyjnym możemy zapisać

$$PD_i = F_{\mathcal{V}_i}(PD_{i-1}) = F_{\mathcal{V}_i} \circ F_{\mathcal{V}_{i-1}} \circ \dots \circ F_{\mathcal{V}_1}(PD_A)$$

Dla $i = n$ zbiór PD_n oznaczmy przez PD_B .

Zbiór PD_B jest zbiorem opisów dokumentów otrzymanych ze zbioru PD_A w wyniku modyfikacji opisów. Zbiór PD_B (po modyfikacji) jest zbiorem „pośrednim” pomiędzy zbiorami opisów

PD_A (przed modyfikacją) i zbiorem opisów idealnych PD_I , tzn. opisy dokumentów ze zbioru PD_B są bardziej podobne (występowanie deskryptorów, wielkości wag) do opisów idealnych niż opisy dokumentów ze zbioru PD_A .

Dla każdego zbioru: PD_A - opisów dokumentów przed adaptacją, PD_B - opisów po adaptacji i PD_I - idealnych opisów tworzymy listy inwersyjne. Otrzymujemy w ten sposób trzy rodzaje list inwersyjnych dla trzech rodzajów opisów dokumentów. Przy pomocy tych list wyszukujemy odpowiedni na 15 kwerend.

Miara podobieństwa odpowiedzi ρ (definicja zostanie podana w rozdziale 4.1.1) będzie porównywać odpowiedź na tę samą kwerendę zawierających dokumenty o opisach ze zbiorów PD_A , PD_B , PD_I . Miara ρ określa różnice (odchylenia) $\mu(\mathcal{V}, \mathcal{E}_A)$ od $\mu(\mathcal{V}, \mathcal{E}_I)$ oraz $\mu(\mathcal{V}, \mathcal{E}_B)$ od $\mu(\mathcal{V}, \mathcal{E}_I)$ dla opisów należących do odpowiednich zbiorów.

Ponadto pokażemy, że modyfikacja wpływa również na wzrost innych współczynników. Zajmiemy się współczynnikiem korelacji Pearsona i cosinusowej.

Jeśli stopnie podobieństwa μ kwerendy do opisów dokumentów należących do odpowiedzi będziemy traktować jako współrzędne odpowiednich wektorów, to możemy zbadać kąty pomiędzy wektorem odpowiedzi przed adaptacją lub po adaptacji a wektorem odpowiedzi idealnej. Wartość kąta zmienia się od 0° do 90° . Jeśli cosinus kąta bliski jest jedności, to współrzędne wektora niewiele różnią się od siebie. Można powiedzieć, że współczynnik korelacji cosinusowej jest miarą „różnic” współrzędnych.

Dla tych samych wektorów można zbadać współzależność liniową współrzędnych wektorów, tzn. czy wzrostowi wartości współrzędnych wektora odpowiedzi idealnych odpowiadają wyższe warto-

ści współrzędnych wektora odpowiedzi przed lub po adaptacji. W rozdziale 4.1.3 podamy szczegółowe definicje tych współczynników oraz obliczone współczynniki. Dla obliczonych współczynników dokonamy analizy statystycznej, sprawdzając czy występujące różnice pomiędzy odpowiednimi współczynnikami są istotne czy przypadkowe.

4.1.1. Miara podobieństwa odpowiedzi

Niech zbiór X zawiera numery (adresy) dokumentów należących do odpowiedzi w zbiorze PD_I , zbiór Y numery dokumentów w zbiorze PD_A , zbiór Z numery dokumentów w PD_B , które należą do odpowiedzi na kwerendę .

Zbiory X , Y i Z możemy przedstawić w postaci:

$$X = \{i : i = d(t) \wedge (t, u) \in \delta(\mathcal{P}) \text{ w zbiorze } PD_I\}$$

$$Y = \{i : i = d(t') \wedge (t', u) \in \delta(\mathcal{P}) \text{ w zbiorze } PD_A\}$$

$$Z = \{i : i = d(t'') \wedge (t'', u) \in \delta(\mathcal{P}) \text{ w zbiorze } PD_B\}$$

Gdzie $d(t)$ oznacza numer dokumentu o opisie $\mathcal{P}$,
 $u = \mu(\mathcal{P}, t)$.

Dane są funkcje

$$f_1, g_1 : X \cup Y \rightarrow [0, 1]$$

$$f_2, g_2 : X \cup Z \rightarrow [0, 1]$$

takie, że:

$$f_1(j) = \begin{cases} u & \text{dla } j \in X \\ 0 & \text{dla } j \notin X \end{cases}$$

$$g_1(j) = \begin{cases} u' & \text{dla } j \in Y \\ 0 & \text{dla } j \notin Y \end{cases}$$

$$f_2(j) = \begin{cases} u & \text{dla } j \in X \\ 0 & \text{dla } j \notin X \end{cases}$$

$$g_2(j) = \begin{cases} u'' & \text{dla } j \in Z \\ 0 & \text{dla } j \notin Z \end{cases}$$

Zbiory wartości funkcji f_1 i f_2 są jednakowe, ale dziedziny tych funkcji są różne.

Zachowanie się różnic miar podobieństwa $\mu(\mathcal{P}, \mathcal{Q})$ dla różnych kwerend w zbiorach PD_A i PD_I oraz PD_B i PD_I będziemy badać przez porównywanie współczynników podobieństw zbiorów X , Y i Z .

$$\rho(X, Y) = 1 - \frac{\sum_{j \in X \cup Y} |f_1(j) - g_1(j)|}{\sum_{i \in X} f_1(i) + \sum_{i \in Y} g_1(i)} \quad (4.1.1)$$

$$\rho(X, Z) = 1 - \frac{\sum_{j \in X \cup Z} |f_2(j) - g_2(j)|}{\sum_{i \in X} f_2(i) + \sum_{i \in Z} g_2(i)} \quad (4.1.2)$$

Jeżeli odpowiedzi w zbiorach PD_I i PD_A nie zawierają ani jednego wspólnego numeru dokumentu, to $X \cap Y = \emptyset$ oraz

$\rho(X, Y) = 0$. Jeżeli odpowiedzi w zbiorach PD_I i PD_A zawierają te same numery dokumentów i miary podobieństw $\mu(\mathcal{P}, \mathcal{Q})$ kwerend do opisów dokumentów są równe, to $\rho(X, Y) = 1$.

Współczynnik $\rho(X, Y)$ nazywać będziemy miarą podobieństwa odpowiedzi w zbiorach PD_I i PD_A (ze względu na stopnie podobieństwa u opisu dokumentu do kwerendy $\mathcal{P}$). Takie same własności ma współczynnik $\rho(X, Z)$ dla odpowiedzi w zbiorach PD_I i PD_B .

W poniższej tabeli podano współczynniki podobieństw dla

odpowiedzi na 15 kwerend, które nie brały udziału w modyfikowaniu zbioru PD_A . $\varphi(X,Y)$ jest współczynnikiem podobieństwa zbioru PD_A (przed adaptacją do zbioru idealnego PD_I .

$\varphi(X,Z_1)$ jest współczynnikiem podobieństwa zbioru PD_B do zbioru idealnego PD_I .

Dla $i = 1$, modyfikację przeprowadzono według wzorów (2.2.4), (2.2.5); dla $i = 2$ według wzorów (2.2.4), (2.2.6) oraz dla $i = 3$ według (2.2.4).

nr	$\varphi(X,Y)$	$\varphi(X,Z_1)$	$\varphi(X,Z_2)$	$\varphi(X,Z_3)$
1	.863	.961	.961	.955
2	.759	.944	.898	.874
3	.645	.900	.904	.879
4	.575	.786	.692	.663
5	.827	.927	.908	.907
6	.533	.769	.683	.669
7	.644	.780	.717	.706
8	.886	.960	.958	.957
9	.171	.823	.800	.650
10	.605	.811	.721	.704
11	.396	.794	.790	.605
12	.819	.899	.923	.863
13	.852	.956	.881	.925
14	.648	.900	.867	.883
15	.179	.810	.789	.627

Do obliczenia średnich wartości współczynników nie stosujemy średnich arytmetycznych ale średnie ważone, ponieważ odpowiedzi składają się z różnej liczby dokumentów. Odpowiedzi

z większą liczbą dokumentów powinny mieć większy wpływ na średnie wartości współczynników.

Średnie ważone współczynników wynoszą

$$\bar{\rho}(X, Y) = 0.702$$

$$\bar{\rho}(X, Z_1) = 0.899$$

$$\bar{\rho}(X, Z_2) = 0.873$$

$$\bar{\rho}(X, Z_3) = 0.841$$

4.1.2. Weryfikacja różnicy współczynników ρ przed i po adaptacji

Przy pomocy wnioskowania statystycznego ustala się czy wielkości współczynników przed adaptacją i po adaptacji różnią się między sobą „w sposób istotny”, czy występujące między nimi różnice są przypadkowe. We wnioskowaniu statystycznym badanie rozpoczyna się od sformułowania hipotezy statystycznej. Różnicę współczynników można uważać za istotną, jeśli obie próby pochodzą ze zbiorowości generalnych o różnych parametrach σ_1^* , σ_2^* (gdzie σ_1^* , σ_2^* oznaczają średnie wartości współczynników odpowiednio przed adaptacją i po adaptacji). Czy istotnie tak jest, to należy sprawdzić.

W tym celu formułuje się przypuszczenie, że obie próby losowe pochodzą z tej samej zbiorowości generalnej o różnych parametrach σ_1^* , σ_2^* . Przypuszczenie to jest właśnie hipotezą H_0 , które należy sprawdzić. Hipoteza zerowa wyraża przypuszczenie, że dwie próby losowe różnią się między sobą nie w sposób „istotny” lecz tylko przypadkowy. Wyrażamy to następująco

$$H_0 : \sigma_1^* - \sigma_2^* = 0$$

Przy weryfikacji hipotezy na gruncie prób losowych może występować błąd pierwszego rodzaju, który powstaje wówczas, gdy odrzucona zostaje hipoteza, zwany także błędem alfa. Podjęcie błędnych rozstrzygnięć zależy od poziomu ufności leżącego u podstaw tych rozstrzygnięć. Hipoteza, która przy przyjęciu prawdopodobieństwa błędu $\alpha = 0.05 \approx 5\%$ zostaje odrzucona stanowi uzasadnienie dla wniosków, o których mówimy, że „leżą one na poziomie ufności 5%”.

Ponieważ porównywane współczynniki podobieństwa są obliczone dla odpowiedzi na te same kwerendy nie bierzemy pod uwagę wartości współczynników lecz ich różnice. Stawiamy pytanie, czy średnia arytmetyczna różnic $d = f(X,Z) - f(X,Y)$ odbiega od zera tylko w sposób przypadkowy czy w sposób istotny.

Hipotezę H_0 sprawdzać będziemy przy pomocy testu „t”. Niech $d_i = f_i(X,Z) - f_i(X,Y)$ będzie różnicą współczynników dla i - tej kwerendy.

Obliczamy wartość t

$$t = \frac{\bar{d}}{\sqrt{\frac{\sum_{i=1}^n d_i^2 - n\bar{d}^2}{n(n-1)}}} \quad (4.1.3)$$

gdzie $\bar{d}$ jest średnią ważoną różnic d_i .

Obliczoną wartość t porównujemy z wartością ustaloną przy pomocy tabeli przy uwzględnieniu przyjętego prawdopodobieństwa błędu α oraz liczby stopni swobody $n - 1$, stąd otrzymujemy następujące kryterium:

Jeśli $t < t_{\alpha, n-1}$, to H_0 przyjąć,
jeśli natomiast $t \geq t_{\alpha, n-1}$, to H_0 należy odrzucić

Ze względu na tradycyjne przyjęte oznaczenia w statystyce literą t oznacza się zmienną losową t - Studenta. W tym miejscu nie powinno powstawać żadne niejasności związane z oznaczaniem w rozdziale 1 deskryptorów przez t .

Dla

$$d_{1i} = f_i(X, Z_1) - f_i(X, Y)$$

$$d_{2i} = f_i(X, Z_2) - f_i(X, Y)$$

$$d_{3i} = f_i(X, Z_3) - f_i(X, Y) \quad (i = 1, 2, \dots, 15)$$

obliczamy

$$\bar{d}_1 = 0.197, \quad \bar{d}_2 = 0.171, \quad \bar{d}_3 = 0.139$$

oraz

$$t_1 = 3.276, \quad t_2 = 2.985, \quad t_3 = 3.495.$$

Odczytana z tabeli ([11], p. 112) wartość $t_{\alpha, n-1}$ dla testu jednostronnego i $\alpha = 0.005$, $n - 1 = 14$ wynosi 2.97.

Ponieważ $t_1, t_2, t_3 \geq t_{0.05, 14}$ należy hipotezę zerową odrzucić.

Otrzymaliśmy, że współczynniki podobieństwa przed i po adaptacji są istotnie różne z prawdopodobieństwem błędu $\alpha = 0.005$.

4.1.3. Współczynnik korelacji Pearsona i cosinusowej

Niech miary podobieństw μ kwerendy $\mathcal{P}$ do idealnych opisów $\mathcal{G}_I$ dokumentów należących do opowiedzi będą współrzędnymi wektora X oraz miary podobieństw μ kwerendy $\mathcal{P}$ do opisów

$\mathcal{E}_A$ nie modyfikowanych (przed adaptacją) należących do odpowiedzi będą współrzędnymi wektora Y . Jeżeli dokument o opisie $\mathcal{E}_I$ (ze zbioru PD_I) należący do odpowiedzi z i - tą rangą należy również o opisie $\mathcal{E}_A$ (ze zbioru PD_A) do odpowiedzi na kwerendę $\mathcal{V}$ to $Y_i = \mu(\mathcal{V}, \mathcal{E}_A)$ w przeciwnym przypadku $Y_i = 0$.

Podobnie jak wektor Y tworzymy wektory:

Z_1 dla modyfikacji według wzorów (2.2.4), (2.2.5);
 Z_2 dla modyfikacji według wzorów (2.2.4), (2.2.6) oraz Z_3 dla (2.2.4).

Dla tak skonstruowanych wektorów X, Y, Z_j ($j = 1, 2, 3$) możemy zbadać współzależność liniową wektorów (X, Y) i (X, Z_j) , tzn. zbadać czy wysokim stopniom podobieństw opisów $\mathcal{E}_I$ odpowiadają wysokie stopnie podobieństw opisów $\mathcal{E}_A$ dokumentów należących do odpowiedzi na tą samą kwerendę. Jeśli wzrost stopni podobieństw jest w przybliżeniu jednostajny, tj. jeśli punkty na wykresie układają się dookoła pewnej prostej, to odpowiednią współzależność liniową nazywamy korelacją Pearsona.

Współczynnik korelacji miarowej Pearsona może przybierać wartości znajdujące się w przedziale liczbowym od -1 do 1 . Gdy współczynnik jest bliski -1 , znaczy to, że między obu wektorami istnieje zupełne przeciwieństwo lub innymi słowy, wysokim wartościom współrzędnych wektora X odpowiadają zawsze niskie wartości współrzędnych wektora Y . Gdy natomiast współczynnik bliski jest jedności, mamy do czynienia z pozytywnym, ściśłym liniowym związkiem obu wektorów, w którym wysokim wartościom współrzędnych wektora X odpowiadają wysokie wartości współrzędnych wektora Y .

Współczynnik korelacji Pearsona oblicza się ze wzoru:

$$P_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (4.1.4)$$

gdzie $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.

Dla tych samych wektorów odpowiedzi X, Y, Z_1, Z_2, Z_3 zastępując różnice $x_i - \bar{x}$, $y_i - \bar{y}$ we wzorze (4.1.4) przez x_i, y_i otrzymamy współczynnik korelacji cosinusowej.

$$Q_{XY} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2 \cdot \sum_{i=1}^n y_i^2}}, \quad (4.1.5)$$

Współczynnik Q_{XY} jest cosinusem kąta pomiędzy wektorami odpowiedzi: Y (przed adaptacją) i X (idealnej). Podobnie interpretuje się współczynniki Q_{XZ_i} ($i = 1, 2, 3$).

Korelacja cosinusa jest nieujemna ponieważ składowe są nieujemne. Współczynnik Q przyjmuje wartości z przedziału $[0, 1]$.

W poniższej tabeli podano współczynniki korelacji Pearsona dla 14 kwerend (odpowiedź na 7 - mą kwerendę zawiera tylko 2 elementy). W pierwszej kolumnie podano numery kwerend, w drugiej liczbę dokumentów w odpowiedzi, w trzeciej współczynniki korelacji P_{XY} dla wektorów X, Y , w kolejnych kolumnach współczynniki P_{XZ_i} dla wektorów X, Z_i ($i = 1, 2, 3$).

NR	N	P_{XY}	P_{XZ_1}	P_{XZ_2}	P_{XZ_3}
1	13	.385	.973	.977	.943
2	15	.183	.959	.647	.672
3	6	-.243	.902	.920	.840
4	4	-.513	.951	.856	-.137
5	39	.243	.616	.565	.468
6	4	-.224	.918	.895	.753
8	9	.814	.959	.964	.948
9	5	-.147	.977	.850	.814
10	4	-.543	.992	.972	.798
11	5	-.210	-.213	-.156	-.371
12	7	.859	.948	.952	.923
13	6	.447	.955	.839	.858
14	16	.143	.783	.690	.777
15	5	-.232	.911	.817	.879

Średnie ważne współczynniki wynoszą:

$$\overline{P}_{XY} = 0.229, \quad \overline{P}_{XZ_1} = 0.887, \quad \overline{P}_{XZ_2} = 0.815, \quad \overline{P}_{XZ_3} = 0.744.$$

Analizując pierwszą kolumnę tabeli widzimy, że korelacja liniowa pomiędzy odpowiedziami zawierającymi opisy idealne i nie modyfikowane jest niska a w niektórych przypadkach nawet ujemna. Dla pewnej części odpowiedzi wysokim stopniom podobieństwa opisów idealnych odpowiadają niższe stopnie podobieństwa opisów nie modyfikowanych do tej samej kwerendy i na odwrót.

W wyniku modyfikacji opisów różnice pomiędzy stopniami podobieństwa odpowiednich opisów zmalały, tzn. w przypadkach

gdy małemu stopniowi podobieństwa opisu idealnego t_I odpowiadał duży stopień opisu t_A nie modyfikowanego, to po modyfikacji opisów stopień podobieństwa opisu t_B do tej samej kwerendy φ jest stopniem zmniejszonym. Podobna sytuacja zachodzi w przypadkach odwrotnych.

Najlepsze efekty otrzymano dla modyfikacji według wzorów (2.2.4), (2.2.5). Porównując średnie ważne współczynniki korelacji Pearsona trzeba pamiętać, że najmniejszy wpływ na średnie moje odpowiedzi o numerach 1, 2, 5, 14 ponieważ zawierają najwięcej dokumentów.

Współczynniki korelacji cosinusowej dla poszczególnych kwerend przedstawiono w poniższej tabeli:

NR	N	Q_{XY}	Q_{XZ_1}	Q_{XZ_2}	Q_{XZ_3}
1	13	.884	.997	.997	.994
2	15	.841	.995	.945	.953
3	6	.666	.987	.989	.979
4	4	.956	.998	.924	.983
5	39	.910	.979	.960	.963
6	4	.937	.997	.934	.991
8	9	.958	.995	.996	.994
9	5	.410	.996	.994	.993
10	4	.942	.999	.941	.991
11	5	.601	.858	.858	.725
12	7	.933	.979	.984	.967
13	6	.915	.995	.951	.986
14	16	.728	.969	.940	.968
15	5	.383	.993	.989	.993

Średnie ważone współczynników korelacji wynoszą:

$$\bar{Q}_{XY} = 0.868, \quad \bar{Q}_{XZ_1} = 0.990, \quad \bar{Q}_{XZ_2} = 0.974, \quad \bar{Q}_{XZ_3} = 0.977$$

Do obliczania średnich współczynników wykorzystuje się transformację Fishera (4.1.6).

Obliczając funkcję arc cos dla współczynników korelacji cosinusowej otrzymujemy wartości kątów pomiędzy wektorami.

Kąt pomiędzy wektorami X i Y wynosi $29^{\circ} 40'$; X i Z_1 - $8^{\circ} 10'$; X i Z_2 - 13° oraz X i Z_3 - $12^{\circ} 20'$.

Modyfikacja według wzorów (2.2.4) i (2.2.5) dla najlepsze wyniki.

Obliczenia współczynników korelacji Pearsona i cosinusowej wykonano również dla list inwersyjnych ze względu na większą liczbę elementów w listach niż w badanych odpowiedziach.

Współczynniki korelacji Pearsona i cosinusowej można również obliczyć dla list inwersyjnych idealnych, przed adaptacją i po adaptacji dla trzech metod modyfikacji. Wektory X, Y, Z_1 , Z_2 , Z_3 tworzymy tak samo, jak dla odpowiedzi na kwerendy. Obliczenie współczynników korelacji dla list inwersyjnych ma sens, ponieważ listy inwersyjne składają się z numerów (adresów) dokumentów z wagami oraz w listach określony jest porządek liniowy dokumentów ze względu na wagi. W następnych trzech tabelach podane są współczynniki korelacji dla list inwersyjnych. W pierwszej kolumnie tabeli podaje się numery list inwersyjnych, w drugiej liczbę elementów w liście, w trzeciej i czwartej współczynniki korelacji Pearsona P_{XY} i P_{XZ_1} (w kolumnie zaznaczonej przez R0, R1), w piątej i szóstej współczynniki korelacji cosinusowej Q_{XY} i Q_{XZ_1} (w kolumnie Q0, Q1). W drugiej

NR	N	R0	R1	Q0	Q1
2	75	.7498	.8965	.9416	.9782
3	69	.7735	.9271	.9467	.9849
4	36	.6317	.8876	.8933	.9701
5	19	.6142	.8935	.9221	.9799
7	21	.6070	.9367	.8934	.9808
9	45	.7947	.9484	.9442	.9868
10	21	.5706	.8176	.9012	.9677
11	6	.5785	.4674	.8205	.8498
13	19	.7807	.9688	.9246	.9892
14	7	.7756	.9649	.9266	.9894
15	8	.8376	.9891	.9386	.9959
16	10	-.1905	.3095	.7869	.9345
17	10	.4444	.8416	.9042	.9620
18	5	.2100	.3236	.8664	.8952
19	92	.4863	.8238	.8552	.9611
20	33	.4645	.8301	.8085	.9364
21	7	.9727	.9867	.9806	.9900
22	7	.4949	.7284	.8447	.9297
23	26	.5702	.8739	.7918	.9542
24	20	.4050	.7868	.7973	.9331
25	32	-.2332	.7902	.5645	.9456
26	19	.7227	.8440	.9035	.9562
27	107	.6740	.9082	.9351	.9810
28	87	.6175	.9221	.9257	.9840
29	53	.4749	.7776	.8669	.9471
30	22	.2386	.4006	.7326	.7957
31	17	.1827	.5295	.7976	.8782
32	18	.8481	.9021	.9481	.9651
33	44	.4959	.7671	.8621	.9355
34	14	.0849	.1768	.7009	.7574
35	3	.8660	.9852	.9487	.9867
36	17	.3231	.6846	.8623	.9306
37	21	.7326	.9495	.8889	.9816
38	8	.3698	.8771	.7769	.9585
39	7	.0601	.6644	.7327	.9608
40	6	.5246	.7904	.9133	.9730
41	36	.0952	.5290	.7365	.8831

Wartości średnie:

$$\overline{R_0} = 0.584$$

$$\overline{R_1} = 0.865$$

$$\overline{Q_0} = 0.892$$

$$\overline{Q_1} = 0.968$$

NR	N	R0	R2	Q0	Q2
2	75	.7498	.9028	.9416	.9795
3	69	.7735	.9319	.9467	.9858
4	36	.6317	.8832	.8933	.9662
5	19	.6142	.8686	.9221	.9738
7	21	.6070	.9293	.8934	.9777
9	45	.7947	.9459	.9442	.9869
10	21	.5706	.7723	.9012	.9594
11	6	.5785	.3595	.8205	.8270
13	19	.7807	.9699	.9246	.9885
14	7	.7756	.9621	.9266	.9888
15	8	.8376	.9876	.9386	.9952
16	10	-.1905	.2425	.7869	.9265
17	10	.4444	.9089	.9042	.9741
18	5	.2100	.3846	.8664	.9028
19	92	.4863	.8485	.8552	.9655
20	33	.4645	.8765	.8085	.9531
21	7	.9727	.9726	.9806	.9823
22	7	.4949	.7708	.8447	.9400
23	26	.5702	.8840	.7918	.9595
24	20	.4050	.8165	.7973	.9415
25	32	-.2332	.7616	.5645	.9367
26	19	.7227	.8427	.9035	.9547
27	107	.6740	.9149	.9351	.9818
28	87	.6175	.9313	.9257	.9856
29	53	.4749	.7926	.8669	.9485
30	22	.2386	.4181	.7326	.8019
31	17	.1827	.5304	.7976	.8646
32	18	.8481	.9033	.9481	.9579
33	44	.4959	.7787	.8621	.9385
34	14	.0849	.1817	.7009	.7629
35	3	.8660	.9925	.9487	.9809
36	17	.3231	.6708	.8623	.9323
37	21	.7326	.9609	.8889	.9856
38	8	.3698	.8813	.7769	.9602
39	7	.0601	.6748	.7327	.9619
40	6	.5246	.8085	.9133	.9748
41	36	.0952	.5645	.7365	.8837

Wartości średnie:

$$\overline{R_0}=0.584$$

$$\overline{R_2}=0.872$$

$$\overline{Q_0}=0.892$$

$$\overline{Q_2}=0.969$$

NR	N	R0	R3	Q0	Q3
2	75	.7498	.8478	.9416	.9690
3	69	.7735	.8797	.9467	.9760
4	36	.6317	.8128	.8933	.9538
5	19	.6142	.7487	.9221	.9546
7	21	.6070	.8861	.8934	.9652
9	45	.7947	.9175	.9442	.9804
10	21	.5706	.7932	.9012	.9640
11	6	.5785	.5533	.8205	.8548
13	19	.7807	.9278	.9246	.9768
14	7	.7756	.9468	.9266	.9824
15	8	.8376	.9578	.9386	.9732
16	10	-.1905	.0689	.7869	.9192
17	10	.4444	.6494	.9042	.9500
18	5	.2100	.3174	.8664	.8626
19	92	.4863	.7228	.8552	.9397
20	33	.4645	.6650	.8085	.8814
21	7	.9727	.9778	.9806	.9817
22	7	.4949	.6023	.8447	.8881
23	26	.5702	.8496	.7918	.9469
24	20	.4050	.6822	.7973	.9044
25	32	-.2332	.7801	.5645	.9424
26	19	.7227	.8159	.9035	.9515
27	107	.6740	.8541	.9351	.9715
28	87	.6175	.8615	.9257	.9725
29	53	.4749	.7170	.8669	.9338
30	22	.2386	.3787	.7326	.7912
31	17	.1827	.4266	.7976	.8723
32	18	.8481	.9350	.9481	.9801
33	44	.4959	.6566	.8621	.9116
34	14	.0849	.0926	.7009	.7461
35	3	.8660	.9788	.9487	.9924
36	17	.3231	.5154	.8623	.9091
37	21	.7326	.9191	.8889	.9702
38	8	.3698	.8409	.7769	.9474
39	7	.0601	.6208	.7327	.9553
40	6	.5246	.7200	.9133	.9666
41	36	.0952	.4273	.7365	.8506

Wartości średnie:

$$\bar{R}_0 = 0.584$$

$$\bar{R}_3 = 0.798$$

$$\bar{Q}_0 = 0.892$$

$$\bar{Q}_3 = 0.953$$

i trzeciej tabeli umieszczone są odpowiednio współczynniki

$$P_{XZ_2}, Q_{XZ_2} \text{ i } P_{XZ_3}, Q_{XZ_3}.$$

W tabelach brak jest współczynników dla list o numerach: 1, 6, 8, 12. Współczynnik korelacji dla pierwszej listy jest równy 1 a dla list o numerach 6, 8, 12 równy 0.

W następnym rozdziale zbadamy czy różnice pomiędzy współczynnikami przed adaptacją i po adaptacji są istotne oraz obliczymy przedziały ufności dla tych współczynników.

4.1.4. Weryfikacja różnicy współczynników korelacji Pearsona i cosinusowej

Rozkład zmiennej losowej P_{XY} jest wysoce asymetryczny i mimo, że przy $n \rightarrow \infty$ jest on zbieżny do rozkładu normalnego, to jednak zbieżność ta jest bardzo powolna ([2], p. 378). Rozkład ten można uważać za normalny dla dużych n , nie mniejszych od 500.

Aby móc zastosować klasyczną metodę weryfikowania testem „t” trzeba dokonać normalizacji rozkładu u przez przekształcenie Fishera „Z”

$$u = \frac{1}{2} \log \left(\frac{1+P}{1-P} \right) = \operatorname{artgh} (P) \quad (4.1.6)$$

Wariancja rozkładu wartości u jest niezależna od korelacji ρ^* w zbiorowości generalnej. Dzięki temu można za pośrednictwem u weryfikować współczynnik korelacji nawet, gdy zmienne X, Y w zbiorowości generalnej są w wysokim stopniu skorelowane (tzn., gdy ρ^* leży blisko 1).

Transformację u wykorzystuje się również do obliczenia średniej ważonej P_{XY} współczynników P_{XY} .

Rozkład próby losowej u jest w przybliżeniu normalny, co pozwala na zastosowanie klasycznej metody weryfikowania testem „t”.

Podobnie jak w rozdziale 4.1.2 obliczamy wartość t według wzoru (4.1.3).

Obliczoną wartość t porównujemy z wartością ustaloną przy pomocy tabeli przy uwzględnieniu przyjętego prawdopodobieństwa błędu α oraz liczby stopni swobody $n - 1$, otrzymujemy kryterium

Jeśli $t < t_{\alpha, n-1}$, to należy przyjąć hipotezę H_0 :
Różnica współczynników korelacji jest przypadkowa

Jeśli $t \geq t_{\alpha, n-1}$, to należy odrzucić hipotezę H_0 :
Różnica współczynników korelacji jest istotna.

Obliczając dla odpowiedzi na kolejne kwerendy

$$d_j = \text{artgh}(P_{XZ_j}) - \text{artgh}(P_{XY}) \quad (j = 1, 2, 3)$$

$$d'_j = \text{artgh}(Q_{XZ_j}) - \text{artgh}(Q_{XZ_j}) \quad (j = 1, 2, 3)$$

otrzymujemy średnie różnice

$$\bar{d}_1 = 0.826, \quad \bar{d}_2 = 0.721, \quad \bar{d}_3 = 0.620$$

$$\bar{d}'_1 = 0.866, \quad \bar{d}'_2 = 0.684, \quad \bar{d}'_3 = 0.721$$

oraz wartość t ze wzoru (4.1.3)

$$t_1 = 3.336, \quad t_2 = 3.135, \quad t_3 = 3.497,$$

$$t'_1 = 4.164, \quad t'_2 = 3.015, \quad t'_3 = 3.777.$$

Odczytana z tabeli ([11], p. 112) wartość $t_{\alpha, n-1}$ dla testu

jednostronnego i $\alpha = 0.005$, $n - 1 = 13$ wynosi 3.012. Z prawdopodobieństwem błędu 0.005 odrzucamy hipotezę H_0 ; różnice współczynników korelacji Pearsona i cosinusowej są istotne.

Takie same obliczenia wykonujemy dla współczynników korelacji dla kolejnych list inwersyjnych.

Średnie różnice wynoszą

$$\bar{d}_1 = 0.569, \quad \bar{d}_2 = 0.588, \quad \bar{d}_3 = 0.402$$

$$\bar{d}'_1 = 0.551, \quad \bar{d}'_2 = 0.560, \quad \bar{d}'_3 = 0.410$$

oraz wartość t ze wzoru (4.1.3)

$$t_1 = 13.770, \quad t_2 = 13.391, \quad t_3 = 9.813$$

$$t'_1 = 11.578, \quad t'_2 = 13.043, \quad t'_3 = 8.199.$$

Odczytana wartość $t_{\alpha, n-1}$ z tabeli ([11], p. 112) dla testu jednostronnego i $\alpha = 0.0005$, $n - 1 = 36$ wynosi 3.582.

Z prawdopodobieństwa błędu 0.0005 odrzucamy hipotezę H_0 , różnice współczynników korelacji Pearsona i cosinusowej dla list inwersyjnych są istotne.

4.2. Zależność współczynników korelacji od liczby kwerend biorących udział w modyfikacji

W poniższych tabelach przedstawiono wartości współczynników korelacji Pearsona dla list inwersyjnych w zależności od liczby kwerend. W pierwszej kolumnie tabeli podano numery kwerend, po których obliczona współczynniki korelacji. W drugiej, trzeciej i czwartej kolumnie podano wartości współczynników dla modyfikacji wg wzorów (2.2.4, 2.2.5), (2.2.4, 2.2.6) i (2.2.4).

W czwartej kolumnie podano liczbę dokumentów, które modyfikowano. W drugiej tabeli przedstawiono współczynniki dla korelacji cosinusowej. Wielkości współczynników korelacji dla list inwersyjnych zależą od liczby modyfikowanych opisów. Im więcej było modyfikowanych opisów dokumentów, tym większy jest wzrost współczynników. Przyrosty współczynników są większe na początku modyfikacji niż przy końcu modyfikacji.

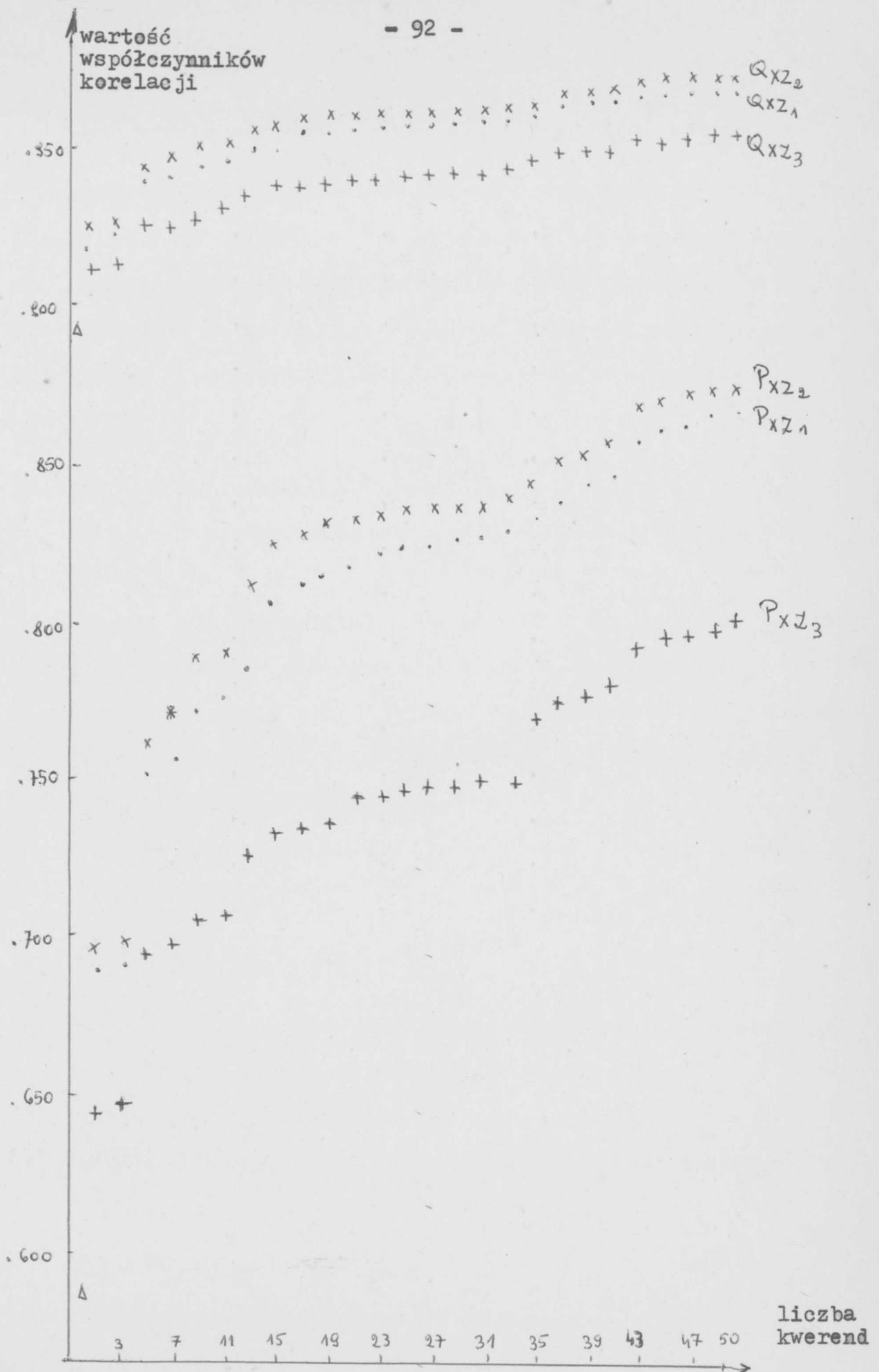
Zależność współczynników korelacji Pearsona i cosinusowej od liczby kwerend biorących udział w modyfikacji przedstawiono na wykresie nr 4.2.1.

i	P_{XZ_1}	P_{XZ_2}	P_{XZ_3}	n
0	.584	.584	.584	0
1	.690	.698	.647	82
3	.692	.700	.650	19
5	.754	.767	.699	116
7	.758	.773	.700	15
9	.774	.790	.708	23
11	.776	.791	.709	10
13	.779	.814	.728	31
15	.809	.825	.734	59
17	.812	.827	.736	9
19	.815	.830	.737	14
21	.819	.831	.743	29
23	.823	.834	.744	25
25	.825	.835	.746	11
27	.825	.836	.747	23
29	.826	.836	.747	8
31	.826	.835	.747	13

1	P_{XZ_1}	P_{XZ_2}	P_{XZ_3}	n
33	.828	.839	.749	23
35	.833	.844	.769	60
37	.839	.850	.775	33
39	.844	.852	.778	41
41	.845	.855	.780	19
43	.858	.868	.793	67
45	.860	.870	.795	41
47	.861	.871	.795	5
49	.863	.872	.798	18
50	.865	.873	.799	7

1	Q_{XZ_1}	P_{XZ_2}	Q_{XZ_3}	n
0	.892	.892	.892	0
1	.924	.926	.913	82
3	.925	.926	.914	19
5	.941	.944	.926	116
7	.942	.945	.926	15
9	.946	.949	.929	23
11	.946	.949	.930	10
13	.952	.956	.936	31
15	.944	.957	.937	59
17	.955	.958	.937	9
19	.955	.959	.938	14

i	Q_{XZ_1}	Q_{XZ_2}	Q_{XZ_3}	n
21	.956	.959	.939	29
23	.956	.960	.940	25
25	.957	.960	.940	11
27	.958	.960	.941	23
29	.958	.960	.941	8
31	.958	.960	.941	13
33	.958	.961	.942	23
35	.960	.962	.946	60
37	.961	.965	.947	33
39	.962	.965	.948	41
41	.961	.965	.949	19
43	.966	.969	.952	67
45	.967	.969	.953	41
47	.967	.969	.953	5
49	.967	.969	.954	18
50	.967	.969	.954	7



Wykres 4.2.1. Zależność współczynników korelacji od liczby kwerend biorących udział w modyfikacji.

4.3. Czas wyszukiwania dokumentów

Czas wyszukiwania odpowiedzi na kwerendę zależy od liczby i ogólności deskryptorów występujących w kwerendzie. Czas wyszukiwania będzie dłuższy dla ogólniejszych deskryptorów lub dla większej liczby deskryptorów występujących w kwerendzie. W dalszych rozważaniach będzie używane średnie czasy.

Oznaczenie:

m - liczba dokumentów w zbiorze D ,

k_0 - średni czas uporządkowania dokumentów w jednej liście inwersyjnej (czas zależy od długości listy),

l_0 - liczba list inwersyjnych (= liczba deskryptorów),

m_1 - średnia liczba deskryptorów w opisie dokumentów,

k_1 - średni czas dopisania jednego deskryptora z wagą do listy inwersyjnej dla tego deskryptora i wszystkich list dla deskryptorów ogólniejszych,

K_1 - średni czas reorganizacji bazy danych (= czas utworzenia list inwersyjnych)

$$K_1 = m \cdot m_1 \cdot k_1 + l_0 \cdot k_0$$

m_2 - średnia liczba dokumentów relewantnych względem kwerendy

k_2 - średni czas modyfikacji opisu dokumentu,

m_3 - średnia liczba dokumentów nierelewantnych,

K_2 - średni czas adaptacji (modyfikacji) opisów dokumentów należących od odpowiedzi na kwerendę

$$K_2 = (m_2 + m_3) \cdot k_2$$

l_1 - średnia liczba deskryptorów w kwerendzie,

k_3 - średni czas obliczenia przekroju dwóch list inwersyjnych,

k_4 - średni czas znalezienia pozycji na liście, dla której wagi spełniają żądany warunek dokładności,

K_3 - średni czas wyszukania dokumentów z uwzględnieniem wag w opisach dokumentów

$$K_3 = (l_1 - 1)k_3 + l_1 \cdot k_4$$

Przy dużych dokładnościach odpowiedzi do obliczania przekrojów bierzemy ^{listy} obcięte (bez dokumentów z małymi wagami). Wtedy średni czas obliczania przekrojów list k_3 jest mniejszy od średniego czasu k'_3 obliczania przekroju pełnych list inwersyjnych.

K_0 - średni czas wyszukiwania odpowiedzi bez uwzględniania wag deskryptorów

$$K_0 = k'_3 (l_1 - 1)$$

Przez c oznaczmy różnicę czasów k'_3 i k_3

$$c = k'_3 - k_3$$

Twierdzenie 4.3.1.

$$K_0 > K_3 \iff c > \frac{l_1}{l_1 - 1} \cdot k_4$$

Dowód

$$\begin{aligned} K_0 > K_3 &\iff k'_3 (l_1 - 1) > (l_1 - 1) k_3 + l_1 \cdot k_4 \iff \\ &(k_3 + c)(l_1 - 1) > (l_1 - 1) k_3 + l_1 \cdot k_4 \iff \\ &c(l_1 - 1) > l_1 \cdot k_4 \end{aligned}$$

Czas reorganizacji (tworzenia) list inwersyjnych dla 150 opisów dokumentów i 41 deskryptorów zmieniał się w przedziale od 3.441 do 3.844 sekund. Średni czas K_1 z 50 -ciu

reorganizacji wynosi 3.581 sekund.

Czas zmiany wag deskryptorów w jednym opisie dokumentu wahał się w przedziale od 0.0005 do 0.032 sekundy. Średni czas wynosi 0.0067 sekundy.

Czas dopisywania nowych deskryptorów dla jednego opisu zmieniał się od 0.0005 do 0.031 sekundy. Średni czas - 0.0035 sekundy. Czas k_2 jest średnim czasem modyfikacji opisu dokumentu, $k_2 = 0.0035 + 0.0067 = 0.01$.

Czas wyszukiwania odpowiedzi zmieniał się od 0.0005 do 0.091 sekund. Średni czas K_3 wyszukania odpowiedzi dla jednej kwerendy wynosi 0.0219 sekundy.

W poniższej tabeli kolumna m zawiera liczbę dokumentów wyszukanych, kolumna n liczby dokumentów relewantnych w k odpowiedziach.

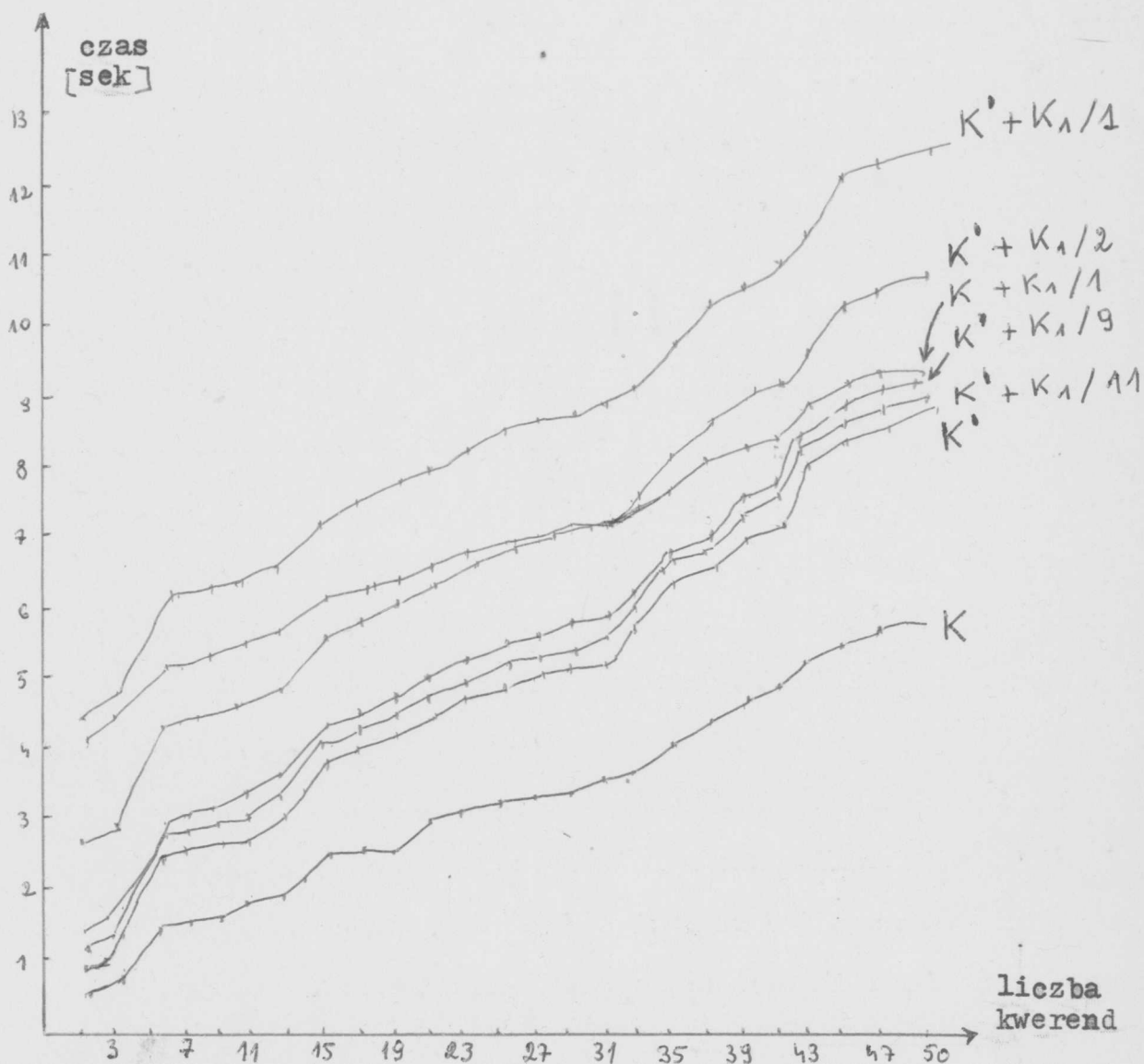
k	m	n
1	82	51
3	101	63
5	217	130
7	232	145
9	255	160
11	265	168
13	296	185
15	355	222
17	364	227
19	378	236
21	407	254
32	432	270
25	443	276

k	m	n
27	456	285
29	464	290
31	477	298
33	500	312
35	560	350
37	593	370
39	631	394
41	650	406
43	717	448
45	758	474
47	763	477
49	781	488
50	787	492

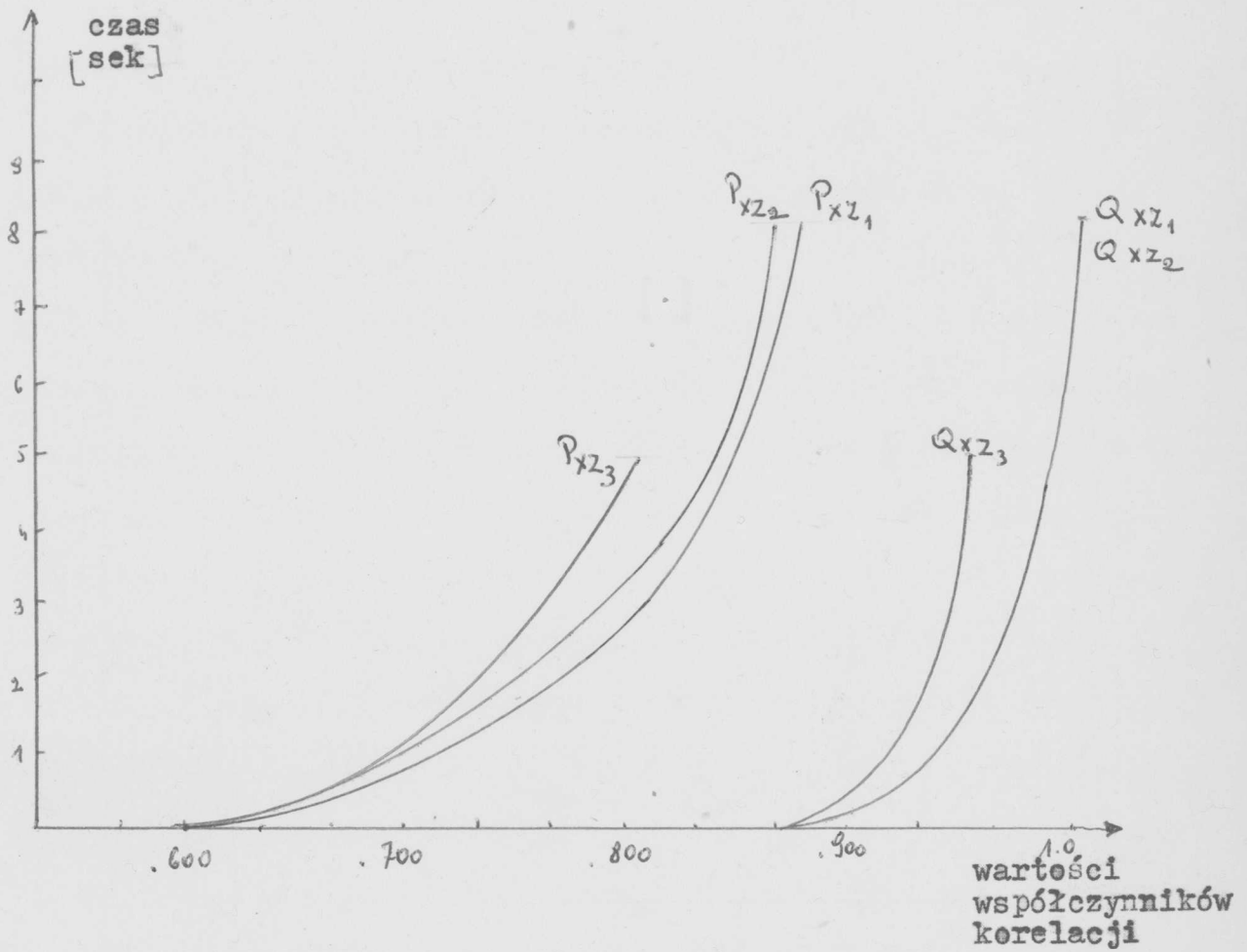
Mnożąc liczbę dokumentów przez średni czas modyfikacji otrzymujemy czas modyfikacji K_2 . Czas $k \cdot K_3$ jest czasem wyszukania k odpowiedzi. Sumą tych czasów jako funkcję liczby kwerend przedstawiono na wykresie nr 4.3.1.

Otrzymujemy dwie rodziny łamanych. Jedną w przypadku modyfikacji opisów dokumentów relewantnych, drugą w przypadku modyfikacji opisów dokumentów relewantnych i nirelewantnych. Łamana tych rodzin są przesunięte względem siebie w zależności od częstotliwości modyfikacji list inwersyjnych. Jeśli K jest czasem wyszukiwania z modyfikacją dokumentów relewantnych to rodzina łamanych ma postać $K + K_1/n$, gdzie n jest częstotliwością tworzenia list inwersyjnych, K_1 czasem utworzenia list inwersyjnych. Podobnie w przypadku gdy K' jest czasem wyszukiwania z modyfikacją relewantnych i nirelewantnych opisów dokumentów.

Korzystając z zależności współczynników korelacji od liczby k kwerend biorących udział w modyfikowaniu opisów i z zależności czasu modyfikacji opisów należących do k odpowiedzi możemy otrzymać zależność czasu modyfikacji od wartości współczynników korelacji (wykres 4.3.2).



Wykres 4.3.1. Czas wyszukania odpowiedzi dla k kwerend z modyfikacją opisów dokumentów.



Wykres 4.3.2. Zależność czasu modyfikacji opisów dokumentów od wartości współczynników korelacji.

4.4. Wnioski

1) Z obliczonych współczynników korelacji Pearsona i cosinusowej oraz podobieństwa odpowiedzi wynika, że modyfikacja opisów dokumentów relewantnych i nierelewantnych daje lepsze rezultaty niż w przypadkach modyfikacji tylko opisów dokumentów relewantnych.

2) Trudno jednoznacznie określić, która z dwu metod modyfikacji opisów dokumentów relewantnych i nierelewantnych daje lepsze wyniki. W pewnych przypadkach (miara podobieństw odpowiedzi, współczynniki korelacji dla odpowiedzi) lepsze wyniki otrzymano dla pierwszej z podanej w rozdziale (2.2) metody modyfikacji, a w innych przypadkach (współczynniki korelacji dla list inwersyjnych) lepsze wyniki otrzymano dla drugiej metody modyfikacji.

3) Przy pomocy badań przeprowadzonych na próbie losowej i znalezieniu przedziałów ufności dla współczynników korelacji możemy stwierdzić, że w zbiorowości generalnej współczynniki korelacji po modyfikacji opisów są większe niż przed modyfikacją.

4) Większej liczbie modyfikacji opisów dokumentów odpowiada większy przyrost współczynników korelacji. Przyrosty współczynników zależą również od opisów dokumentów. Jeśli pomiędzy opisami dokumentów i kwerendami występują duże różnice wag i deskryptorów (mały stopień podobieństwa kwerend do opisów) to przyrosty współczynników korelacji są większe niż gdy pomiędzy opisami i kwerendami występują małe różnice. Na początku modyfikacji opisów przyrosty współczynników korelacji są większe niż przy końcu ponieważ w czasie modyfikacji

opisy dokumentów „upodobniają się” do kwerend. (por. wykres 2.2.1 i wykres 4.3.1).

5) Mniejszym (większym) przyrostem czasu modyfikacji opisów dokumentów odpowiadają większe (mniejsze) przyrosty współczynników korelacji (por. wykres 4.3.2). Wynika to stąd, że przyrost stopnia podobieństwa opisu do kwerendy nie zależy od czasu modyfikacji ale od samego stopnia podobieństwa. Przyrost w czasie modyfikacji jest większy dla małych stopni niż dla dużych stopni, natomiast czas modyfikacji opisów jest jednakowy w obu przypadkach.

W dalszych badaniach można rozważać zachowanie się innych współczynników korelacji i miar podobieństwa odpowiedzi podczas modyfikacji opisów dokumentów oraz zachowanie się współczynników korelacji przy innych metodach symulacji oceniania relewantności dokumentów.

Na koniec postawimy sobie pytanie, czy współczynniki charakteryzujące efektywność systemu są zależne od różnych zbiorów orzeczeń o relewancji, wiadomo bowiem, że na podjęcie decyzji o relewancji danego dokumentu mogą wpływać różne czynniki, np. tematyka, stopień trudności dokumentów, dostępny czas, rozmiar zbioru dokumentów, doświadczenie i wykształcenie orzekającego, sposób orzekania o relewancji itp.

Jeżeli miarą efektywności systemu wyszukiwania informacji są wskaźniki kompletności (stosunek liczby dokumentów wyszukanych i relewantnych do liczby dokumentów relewantnych w zbiorze) i dokładności (stosunek liczby dokumentów wyszukanych i relewantnych do liczby dokumentów wyszukanych) to wskaźniki te nie zależą w istotny sposób od jednomyslności orzeczeń o relewancji,

nawet w przypadkach gdy jednomyslność decyzji o relewancji sięga zaledwie 30%. Przeprowadzone badania eksperymentalne przez M.E.Lesk i G.Saltona w pracy ([6], pp. 536 - 557) wykazały, że chociaż zbiory dokumentów są różnie zdefiniowane przy różnych orzekających o relewancji, to jednak w przypadku dokumentów najbardziej zbliżonych do kwerend orzeczenia w większości są jednomyslne. Ponieważ są to właśnie te dokumenty, które mają największy wpływ na wyznaczenie efektywności wyszukiwania, dlatego też wskaźniki efektywności nie zależą w istotny sposób od jednomyslności ocen.

W dalszej części cytowanej pracy autorzy przeprowadzili teoretyczne rozważania, w których przy pomocy statystycznych wyliczeń wykazali, że warunki nałożone na orzeczenia o relewancji są bardzo słabe. Wymaga się tylko, aby orzekający z większym prawdopodobieństwem zakwalifikował dokument relewantny jako „relewantny” niż „relewantnym” nazwał dokument nierelewantny.

Można przypuszczać, że różne zbiory ocen relewancji nie będą istotnie wpływały na inne miary efektywności systemu. Ale to wymaga przeprowadzenia eksperymentów lub rozważań teoretycznych.

Rozdział 5. Opis metody zmniejszenia redundancji dokumentów

Metoda list inwersyjnych w systemach wyszukiwania informacji z określoną relacją hierarchii wprowadza bardzo dużą redundancję dokumentów. Wprowadzenie wag deskryptorów zwiększa redundancję dwukrotnie, ponieważ w listach inwersyjnych adresy dokumentów występują z wagami. W czasie wyszukiwania dokumentów nie wszystkie adresy dokumentów z małymi wagami są wykorzystywane. Ta własność jest podstawą podanej w tym rozdziale metody zmniejszenia redundancji dokumentów.

5.1. Zmniejszenie redundancji a efektywność systemu.

Metoda list inwersyjnych w systemach wyszukiwania informacji z określoną relacją hierarchii R na deskryptorach wprowadza dużą redundancję względną dokumentów WRD .

$$WRD = 2 \left(\sum_{t \in T_0} \overline{L_R(t)} - m \right)$$

(m jest liczbą dokumentów).

We wzorze tym występuje podwojona suma ponieważ listy inwersyjne zawierają adresy dokumentów z wagami.

Do listy $L_R(t)$ należą wszystkie adresy dokumentów z wagami, które w swoich opisach zawierają deskryptor t oraz wszystkie adresy tych dokumentów, które w swoich opisach posiadają deskryptory ściślejsze względem deskryptora t .

W praktyce będziemy się posługiwać bezwzględną redundancją dokumentów, która wskazuje średnią krotność powtórzenia dokumentu w listach inwersyjnych. Bezwzględną redundancję doku-

mentów, która wskazuje średnią krotność powtórzenia dokumentu w listach inwersyjnych. Bezwzględną redundancję dokumentów oznaczać będziemy przez BRD i $BRD = WRD/m$.

Redundacja jest proporcjonalna zarówno do mocy zbioru dokumentów D , mocy zbioru deskryptorów T_0 jak i do stopnia hierarchiczności deskryptorów systemu. W przykładowym badanym systemie redundacja względna dla 150 dokumentów wynosi $2 \cdot 1035$, bezwzględna redundacja $2 \cdot 6.9$.

W dalszej części rozdziału zostanie podany opis jednego ze sposobów zmniejszenia redundacji dokumentów. Zmniejszenie redundacji odbywa się kosztem zmniejszenia pewnej liczby dokumentów w niektórych odpowiedziach.

Można założyć, że użytkownicy dla ogólniejszych deskryptorów zawartych w kwerendach będą używać większych wag i żądać większej dokładności niż dla deskryptorów szczegółowych. Liczby $b = w \cdot c$ (w - waga deskryptora występującego w kwerendzie, c - wymagana dokładność odpowiedzi) dla dowolnego ogólniejszego deskryptora w większości przypadków będą większe od liczb b dla deskryptorów szczegółowych. Dlatego też z list inwersyjnych dla ogólniejszych deskryptorów adresy z małymi wagami będą używane rzadziej niż adresy z większymi wagami.

Adresy z małymi wagami, bardzo rzadko używane, niepotrzebnie zwiększają redundację dokumentów. Najwięcej zbędnych adresów znajduje się na listach dla ogólniejszych deskryptorów ponieważ liczby $b = w \cdot c$ obcinające listy inwersyjne w czasie wyszukiwania będą w większości przypadków większe od liczb dla list deskryptorów szczegółowych oraz dla deskrypto-

rów ogólniejszych listy zawierają więcej deskryptorów z wagami niż listy dla deskryptorów szczegółowych. W dalszej części zostanie podane kryterium statystyczne wyboru wartości progowej b_k liczb b dla każdej listy inwersyjnej, z której można usunąć wszystkie adresy dokumentów o wagach mniejszych od wartości progowej b_k .

Jeżeli liczby $b = w \cdot c$ będą mniejsze od wartości progowej b_k to do odpowiedzi będą należeć tylko te adresy, których wagi są większe od b_k . W odpowiedzi będzie brakować adresów dokumentów, których wagi są z przedziału liczbowego (b, b_k) . W zależności od przyjętych parametrów systemu dla 1% lub 5% ogólnej liczby odwołań do listy inwersyjnej liczby $b = w \cdot c$ będą mniejsze od ustalonej wartości progowej. Można więc przypuścić, że usunięcie adresów o wagach mniejszych od wartości progowej b_k nie spowoduje zbyt dużego zmniejszenia efektywności systemu, zwłaszcza że wśród niepełnych odpowiedzi brakować może tylko dokumentów o małych wagach deskryptorów.

W przypadkach gdy liczba odwołań do danej listy w okresie między jedną a drugą reorganizacją jest zbyt mała aby wyznaczyć wartość progową to wtedy listę zostawiamy bez zmiany.

Między jedną a drugą reorganizacją list inwersyjnych będziemy zbierać informacje o liczbach b . W czasie reorganizacji list badać będziemy rozkład liczb b . W algorytmie wyboru wartości progowej liczb b korzystać będziemy z rozkładu beta, dlatego najpierw podamy podstawowe własności rozkładu beta.

5.2. Własności rozkładu beta.

Do empirycznych zdarzeń losowych można zastosować rozkład beta, dlatego gdyż:

- 1) funkcja gęstości rozkładu beta jest określona na odcinku $[0,1]$
- 2) funkcja gęstości może być zarówno symetryczna jak i asymetryczna w zależności od parametrów p i q .
- 3) bardzo prosto oblicza się parametry p i q , które są funkcjami pierwszego i drugiego momentu zwykłego m_1, m_2 .

Funkcja gęstości rozkładu beta jest postaci

$$f(x) = \begin{cases} \frac{x^p (1-x)^{q-1}}{B(p, q)} & 0 \leq x \leq 1 \\ 0 & x \leq 0 \wedge x > 1 \end{cases}$$

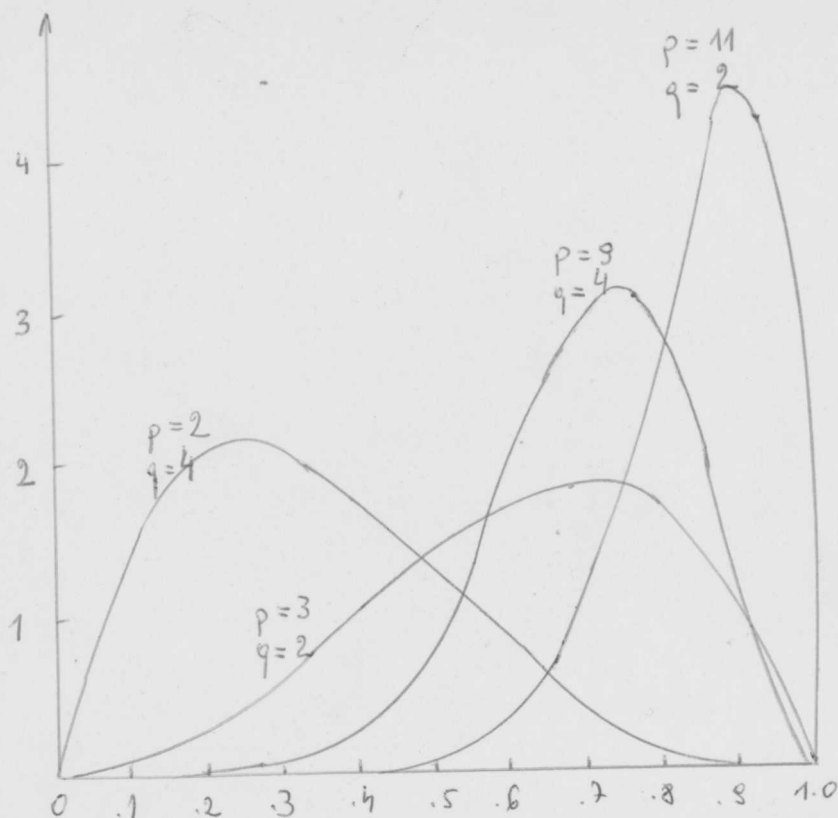
gdzie $B(p, q) = \frac{\Gamma(p) \Gamma(q)}{\Gamma(p+q)}$

Wykres funkcji gęstości w zależności od parametrów p, q przedstawiono na wykresie 5.2.1.

Funkcja gęstości wartości maksymalną przyjmuje w punkcie $(p-1)/(p+q-2)$.

Pierwszy i drugi moment zwykły rozkładu beta ma postać ([2], p. 167):

$$\begin{cases} m_1 = p/(p+q) & (5.2.1) \\ m_2 = \frac{p(p+1)}{(p+q)(p+q+1)} & (5.2.2) \end{cases}$$



Wykres 5.2.1

Rozwiązując układ dwóch równań (5.2.1) i (5.2.2) ze względu na niewiadome p, q otrzymujemy

$$p = m_1 \cdot \frac{m_1 - m_2}{m_2 - m_1^2} \quad (5.2.3)$$

$$q = (1 - m_1) \cdot \frac{m_1 - m_2}{m_2 - m_1^2} \quad (5.2.4)$$

5.3. Algorytm wyboru wartości progowych

Odcinek $[0,1]$ dzielimy na k podprzedziałów $(x_{i-1}, x_i]$ o długości $1/k$ ($x_i = x_{i-1} + 1/k$)

Zbieranie informacji w czasie wyszukiwania dokumentów o wielkościach liczb b przebiega następująco:

Z każdą listą $L_R(t)$ związany jest wektor P_t o k elementach. Jeśli $b \in (x_{i-1}, x_i]$ to $P_t(i) := P_t(i) + 1$. Po N kwerendach odwołujących się do listy inwersyjnej $L_R(t)$ elementy wektora $P_t(i)$ zawierają krot-ności (liczebność) liczb b z przedziału $(x_{i-1/k}, x_i)$, dla $i = 1, 2, \dots, k$.

Elementy wektora P_t są wartościami funkcji rozkładu częstości $g(X)$ zmiennej losowej X podającymi częstość występowania i - tej realizacji tej zmiennej

$$g \{ x_i - 1/k < X < x_i \} = n_i = P_t(i)$$

przy czym:

$$1) \quad n_i \geq 0$$

$$2) \quad \sum_{i=1}^k n_i = N$$

X jest zmienną losową typu skokowego o punktach skokowych x_i , skokach n_i/N .

Pierwszy moment zmiennej losowej X przyjmuje wartość

$$m_1 = N \sum_{i=1}^k x_i \frac{n_i}{N}$$

a drugi moment

$$m_2 = N \sum_{i=1}^2 x_i^2 n_i$$

Rozkład częstotliwości można przedstawić za pomocą histogramu. Zmienna rozkładu, której rozkład przedstawiono w postaci histogramu jest ciągła. Zwiększając nieograniczenie liczbę podprzedziałów na odcinku $[0,1]$ otrzymamy krzywą liczebności.

Wartość progową b_k znajdujemy rozwiązując nierówność

$$P(X \leq b_k) = \int_0^{b_k} \frac{1}{B(p,q)} x^{p-1} (1-x)^{q-1} dx \leq \alpha$$

ze względu na niewidomą b_k .

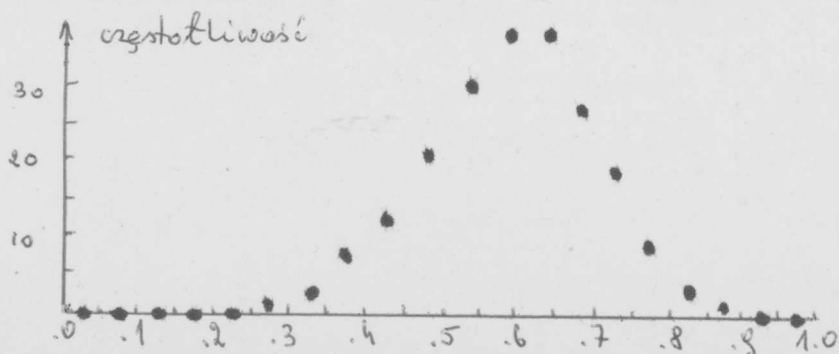
α jest prawdopodobieństwem popełnienia błędu.

Przykład 5.3.1.

Częstotliwość wystąpień liczb $b = w \cdot c$ w poszczególnych przedziałach $(x_{i-1/k}, x_i]$ dla $k = 20$ przedstawia się następująco:

i	6	7	8	9	10	11	12	13	14	15	16	17	18
n_i	1	3	7	13	21	29	34	34	28	19	10	3	1

Rozkład przedstawiony za pomocą histogramu ma postać



Obliczamy momenty rozkładu m_1, m_2 ze wzorów (5.2.1), (5.2.2)

$$m_1 = \frac{1}{N} \sum_{i=1}^{20} x_i n_i = 0.612$$

$$m_2 = \frac{1}{N} \sum_{i=1}^{20} x_i^2 n_i = 0.387$$

Korzystając ze wzorów (5.2.3), (5.2.4) obliczamy p i q

$$p = 11.055$$

$$q = 7.0086$$

Odczytując z tabeli ([11], pp. 133, 137) wartość progowa dla $p = 11$, $q = 7$ i $\alpha = 0.01$ otrzymujemy $b_k = 0.343$. Dla $\alpha = 0.05$, $b_k = 0.421$.

Z prawdopodobieństwem popełnienia błędu $\alpha = 0.05$ z listy możemy usunąć adresy tych dokumentów, których wagi są mniejsze od 0.4 (gdyż wartość kryterialna na poziomie istotności $P = 0.05$ wynosi 0.421). Do odpowiedzi nie będą należeć dokumenty, których wagi w listach są z przedziału (b, b_k) , przy tym 5% z ogólnej liczby odwołań do tej listy będzie spełniało warunek $b < b_k$

W czasie wyszukiwania dokumentów z pamięci zewnętrznych będą sprawdzane listy inwersyjne do pamięci operacyjnych. Zmniejszenie redundacji dokumentów może wpłynąć nie tylko na zmniejszenie obszaru pamięci operacyjnej, ale też może przynieść korzyści użytkownikom w postaci zmniejszenia kosztów korzystania z systemu. Pokażemy to na następującym przykładzie.

Przykład 5.3.2.

Koszty korzystania z pamięci operacyjnej oblicza się następująco:

Najpierw obliczamy koszty wyrażone w jednostkach kilosłósekundach, tzn. wykorzystywany obszar pamięci wyrażony w kilosłowach mnożymy przez liczbę sekund korzystania z tej pamięci. Otrzymane wielkości dzielimy przez stały przelicznik zamieniający kilosłowo sekundy na sekundy. Niech $z = 16.5$ (taki przelicznik jest stosowany w przypadku korzystania z pamięci operacyjnej m.c. CDC 6400). Załóżmy, że w pamięci maszyny mamy dwie najdłuższe listy inwersyjne zawierające po 5000 adresów dokumentów i 5000 wag. Przy zbieraniu informacji o liczbach b korzystamy z dwóch wektorów 50-cio elementowych. Załóżmy ponadto, że dla pewnej liczby kwerend czas pracy systemu wynosi 500 sekund. Po znalezieniu wartości progowej dla tych list okazało się, że na każdej liście zbyteczne jest po 1250 adresów i wag. Koszt korzystania z pełnych list inwersyjnych wynosi

$$K = \frac{2(5 + 5 + .05) \cdot 500}{16.5} \text{ sek} = 609.9 \text{ sek}$$

Koszt korzystania z niepełnych list wynosi

$$K = \frac{2(3.75 + 3.75 + .05) \cdot 500}{16.5} \text{ sek} = 458,2 \text{ sek}$$

5.4. Wnioski

1) Redundancję można jeszcze bardziej zmniejszyć w przypadkach gdy z systemu korzystają użytkownicy tylko z kilku dziedzin tezaurya a nie z jego całości. Wtedy do pewnych list nie będzie odwołań lub będą bardzo rzadkie. Listy te można usunąć.

2) Gdyby w praktyce okazało się, że inne krzywe teoretyczne rozkładów prawdopodobieństwa lepiej przybliżają krzywą empiryczną na odcinku $[0,1]$ to sposób wyznaczania wartości progowych dla każdej z list pozostanie taki sam.

3) Wyższymi wartościami progowymi w poszczególnych listach odpowiada zmniejszenie redundancji dokumentów. Zmniejszenie redundancji wpływa z kolei na zmniejszenie efektywności systemu.

W zależności od potrzeb użytkowników można odpowiednio sterować wartościami progowymi w każdej z list.

Zakończenie.

Rozważania i otrzymane wyniki symulacji procesu wyszukiwania sugerują, że modyfikacja wektorów opisów dokumentów może stanowić pożyteczną pomoc w wyszukiwaniu informacji. Oczywiste jest, że metody takie wiążą się z korzyściami zarówno z punktu widzenia efektywności wyszukiwania jak i indeksowania oraz, że korzyści te wiążą się z dodatkowymi kosztami wyszukiwania.

Otrzymane wyniki mogą być podstawą do dalszych badań i zastosowań praktycznych.

Istniejące systemy wyszukiwania informacji oparte na tezaurusach można niezbyt wysokimi kosztami przystosować do warunków modelu podanych w rozdziale 1. Wystarczy wszystkim deskryptorom w opisach dokumentów nadać początkowe wagi równe $1/f_j$ (f_j - liczba różnych deskryptorów występujących w j - tym opisie dokumentów). Listy inwersyjne spełniające założenia z rozdziału 1.4 można otrzymać dokonując jednego pełnego przeglądu wszystkich opisów dokumentów. W ten sposób otrzymamy rzeczywisty system spełniający założenia badanego systemu, w którym możemy prowadzić dalsze badania i korzystać z systemu w celach praktycznych.

Literatura

1. Dąbrowski M. : Modele systemów wyszukiwania informacji. Prace Instytutu Organizacji i Kierowania - Cybernetyka Stosowana i Informatyka, seria B, zeszyt 16, Warszawa. 1975
2. Fisz M. : Rachunek prawdopodobieństwa i statystyka matematyczna, Warszawa 1967, PWN.
3. Hall M., Jr. : Combinatorial theory. Waltham Mass. 1967, Blaisdell.
4. Leska M., Leski K. : Tezaurus informacji naukowej. Prace - Studia - Przyczynki, 1(54), Warszawa 1972, IINTE.
5. Salton G. : Automatic Information Organization and Retrieval, New York 1968, McGraw - Hill.
6. Salton G. (ed.). : The Smart Retrieval System. New Jersey 1971, Prentice - Hall (Przekład polski : Salton G. (ed.). : Smart - automatyczny system wyszukiwania informacji, Warszawa 1975, WNT.).
7. Turski W. M. : On a Model of Information Retrieval System Based on Thesaurus. Information Storage and Retrieval, 7, 1971, pp. 88 - 94, 201.
8. Turski W. M. : A correspondence between thesaurus - based and multi - attribute retrieval systems. Information Storage and Retrieval, 8, 1972, pp. 309 - 313.
9. Turski W. M. : Niektóre zagadnienia struktur danych. Problemy przetwarzania informacji, tom 2, Warszawa 1974, WNT, pp. 153 - 197.

10. Turski W.M. : Propedeutyka informatyki. Warszawa 1975, PWN.
11. Zieliński R. : Tablice statystyczne, Warszawa 1972, PWN.

Dodatek (treści podprogramów)

PROGRAM START(INPUT,OUTPUT)

```
* * * * *
*
5  *   UWAGI  O  DEKLARACJACH  TABLIC
*
* * * * *

10  *   IMPLICIT INTEGER(A-U)
*   LOGICAL RELEW,MOD
*   DIMENSION R(25),WR(24),P(151),WP(151)
*   COMMON /OBSZAR/ TEZAU(41),INDEX(41,2),DESKRY(41,3)
*   +/BAZA/ DOKUM(150,25),WDOKUM(150,24)
*   +/IDBAZA/ ID(150,25),WID(150,24)
15  *   +/LISTA / PL(42),LIST(1250),WLIST(1250)
*
* * * * *
*
20  *   (R(I),I=1,24) - REPREZENTACJA DESKRYPTOROW KWERENDY ,
*   R(25) - NUMER KWERENDY .
*   (WR(I),I=1,24) - WAGI DESKRYPTOROW
*   TEZAU - TABLICA RELACJI HIERARHII .
*   INDEX(I,1) - NUMER ZBIORU T(I)
*   INDEX(I,2) - MIEJSCE DESKRYPTORA W ZBIORZE T(I) .
25  *   DESKRY(I,1),DESKRY(I,2) - NAZWY DESKRYPTOROW (ZNAKOWO)
*   DESKRY(I,3) - MIEJSCE I-TEGO DESKRYPTORA W TABLICY TEZAU
*   (DOKUM(I,J),J=1,24) - PRZEDSTAWIENIE DESKRYPTOROW I-TEGO OPISU
*   DOKUMENTU, DOKUM(I,25) - NUMER DOKUMENTU
*   WDOKUM(I,J),J=1,24) - WAGI DESKRYPTOROW I-TEGO OPSU DOKUMENTU
30  *   PL(I) - POCZATEK I-TEJ LISTY INWERSYJNEJ
*   LIST - ZBIOR NUMEROW DOKUMENTOW ZNAJDUJACYCH SIE NA
*   WSZYSTKICH LISTACH INWERSYJNYCH.
*   WLIST - ZBIOR WAG DOKUMENTOW
*
35  * * * * *

END
```

SUBROUTINE INWERS

```
* * * * *
*
*   TWORZENIE LIST INWERSYJNYCH.
*
* * * * *
*   IMPLICIT INTEGER (A-U)
*   DIMENSION D(25),WD(24),P(151),WP(151),LISTY(41,151),WLISTY(41,151)
*   COMMON /OBSZAR/ TEZAUR(41),INDEX(41,2),DESKRY(41,3)
10  +/LISTA / PL(42),LIST(1250),WLIST(1250)
*   DO 6 I=1,41
*   6 LISTY(I,1)=1
*   DO 5 I=1,150
*   READ(4,73) (D(J),J=1,25)
15  READ(4,74) (WD(J),J=1,24),NRW
* * * * *
*   CZYTANIE OPISU DOKUMENTU.
*
* * * * *
20  73 FORMAT(24I5,1X,I4)
*   74 FORMAT(24F5.3,1X,I4)
*   Z=0
*   DO 63 J=1,24
25  IF(WD(J).GE.0.04) GO TO 61
* * * * *
*   USUWANIE DESKRYPTOROW Z OPISOW DOKUMENTOW, KTORYCH WAGI SA
*   MNIEJSZE OD 0.04.
30  * * * * *
*   D(J)=0
*   WD(J)=0
*   61 Z=Z+WD(J)
35  63 CONTINUE
*   DO 62 J=1,24
*   62 WD(J)=WD(J)/Z
*   IF(D(25).EQ.NRW) GO TO 10
*   WRITE(2,55) D(25),NRW
40  55 FORMAT(* DESKRYPTORY NR=*,I3,5X,**WAGI NR=*,I3)
*   10 JK=1
*   DO 11 J=1,24
*   IF(D(J).EQ.0) GO TO 11
*   DO 12 J1= JK,41
45  IF(INDEX(J1,1).EQ.J.AND.INDEX(J1,2).EQ.D(J)) GO TO 14
*   12 CONTINUE
*   14 L=LISTY(J1,1)=LISTY(J1,1)+1
*   LISTY(J1,L)=D(25)
*   WLISTY(J1,L)=WD(J)
50  JK=J1
*   11 CONTINUE
*   5 CONTINUE
*   DO 7 IA=1,41
*   7 LISTY(IA,1)=LISTY(IA,1)-1
55  K2=1
*   DO 18 NR=1,41
*   DW=IB=LISTY(NR,1)+1
```

```
        DO 8 IA=1,IB
        WP(IA)=WLISTY(NR,IA)
60      8 P(IA)=LISTY(NR,IA)
        DOS=TEZAUR(NR)
        KNR=NR+1
        IF(KNR.GT.DOS) GO TO 111
        DO 17 SUM=KNR,DOS
65      ILE=LISTY(SUM,1)+1
        DO 21 J1=2,ILE
        KL=DW+1
        P(KL)=ELEM=LISTY(SUM,J1)
        WP(KL)=WLISTY(SUM,J1)
70      I=1
        20 I=I+1
        IF(P(I).NE.ELEM) GO TO 20
        IF(I.LE.DW) WP(I)=WP(I)+WP(KL)
        IF(I.GT.DW) DW=DW+1
75      21 CONTINUE
        17 CONTINUE
        111 P(1)=DW-1
        K=P(1)
        K1=K+1
80      DO 31 L=2,K1
        DO 32 L1=2,K
        IF(WP(L1).GE.WP(L1+1)) GO TO 32
        MN=P(L1)
        P(L1)=P(L1+1)
85      P(L1+1)=MN
        ZN=WP(L1)
        WP(L1)=WP(L1+1)
        WP(L1+1)=ZN
        32 CONTINUE
90      31 CONTINUE
        DO 92 DE=2,K1
        K3=K2+DE-2
        LIST(K3)=P(DE)
        WLIST(K3)=WP(DE)
95      92 CONTINUE
        PL(NR+1)=K2=K3+1
        18 CONTINUE
        PL(1)=1
        RETURN
100     END
```

SUBROUTINE TPRIM(NR)

```
* * * * *  
*  
* DOPISYWANIE NOWYCH DESKRYPTOROW DO OPISU DOKUMENTU  
* 0 ADRESIE NR.  
*  
* * * * *
```

IMPLICIT INTEGER(A-U)

DIMENSION TP(24),WTP(24)

COMMON /OBSZAR/ TEZAU(41),INDEX(41,2),DESKRY(41,3)

+ /BAZA/ DOKUM(150,25),WDOKUM(150,24)

+ /IDBAZA/ ID(150,25),WID(150,24)

XE=0.0

DO 63 IC=1,24

ROB=TP(IC)=ID(NR,IC)

WTP(IC)=WID(NR,IC)

IF(ROB.EQ.0) GO TO 63

OD=1

DO 60 I4=1,24

IF(DOKUM(NR,I4).EQ.0) GO TO 60

DO 61 J1=OD,41

IF(I4.EQ.INDEX(J1,1).AND.DOKUM(NR,I4).EQ.INDEX(J1,2)) GO TO 62

61 CONTINUE

62 OD=J1+1

K1=TEZAU(J1)

DO 64 JC=J1,K1

IF(IC.EQ.INDEX(JC,1).AND.ROB.EQ.INDEX(JC,2)) GO TO 65

64 CONTINUE

60 CONTINUE

GO TO 63

65 WTP(IC)=TP(IC)=0

XE=XE+WID(NR,IC)

63 CONTINUE

DO 2 I=1,24

DOKUM(NR,I)=DOKUM(NR,I)+TP(I)

2 WDOKUM(NR,I)=(WDOKUM(NR,I)+WTP(I))/(2-XE)

RETURN

END

SUBROUTINE MODWAG(R,WR,NR,XC,MOD,CYKL)

```
* * * * *
*
*   MODYFIKACJA OPISU DOKUMENTU O ADRESIE NR.
*
*   * * * * *
*
*   PARAMETR CYKL OKRESLA RODZAJ MODYFIKACJI,
*   MOD=.TRUE. - GDY OPIS DOKUMENTU JEST RELEVANTY WZGLEDEM
10  *   KWERENDY R, W PRZECIWNYM PRZYPADKU MOD=.FALSE.
*   XC OKRESLA SZYBKOSC ZMIAN WAG W OPISIE DOKUMENTU.
*
*   * * * * *
15  *   IMPLICIT INTEGER (A-U)
*   LOGICAL MOD
*   DIMENSION R(25),WR(24),RWD(25),WRWD(24)
*   COMMON /OBSZAR/ TEZAU(41),INDEX(41,2)
*   +/BAZA/ DOKUM(150,25),WDOKUM(150,24)
20  *   XALFA=.2
*   DO 1 I=1,24
*   RWD(I)=0
1  *   WRWD(I)=0
*   OD=1
*   DO 60 I4=1,24
25  *   IF(R(I4).EQ.0) GO TO 60
*   DO 61 J=OD,41
*   IF(I4.EQ.INDEX(J,1).AND.R(I4).EQ.INDEX(J,2)) GO TO 62
61  *   CONTINUE
62  *   WJ=0
30  *   OD=J+1
*   K1=TEZAU(J)
*   DO 63 IC=1,24
*   ROB=DOKUM(NR,IC)
*   IF(ROB.EQ.0) GO TO 63
35  *   DO 64 JC=J,K1
*   IF(IC.NE.INDEX(JC,1).OR.ROB.NE.INDEX(JC,2)) GO TO 64
*   WJ=WJ+WDOKUM(NR,IC)
*   RWD(IC)=I4
*   GO TO 63
40  *   64 CONTINUE
*   63 CONTINUE
*   IF(WJ.EQ.0) DOKUM(NR,I4)=R(I4)
*   WRWD(I4)=WJ
*   60 CONTINUE
45  *   DO 2 IC=1,24
*   I=RWD(IC)
*   IF(I.EQ.0) I=IC
*   XWR=XALFA*XC*WR(I)
*   IF(WRWD(I).EQ.0) GO TO 10
*   XWR=XWR*WDOKUM(NR,IC)/WRWD(I)
50  *   10 XA=1-XALFA*XC
*   IF(MOD) WDOKUM(NR,IC)=WDOKUM(NR,IC)*XA+XWR
*   GO TO (21,22,2),CYKL
21  *   IF(.NOT.MOD) WDOKUM(NR,IC)=ABS(WDOKUM(NR,IC)*(2-XA)-XWR)
55  *   GO TO 2
22  *   IF(.NOT.MOD) WDOKUM(NR,IC)=ABS(WDOKUM(NR,IC)*XA-XWR)
*   2 CONTINUE
```

60

```
      IF(MOD) GO TO 5
      Z=0
      DO 3 I=1,24
3     Z=Z+WDOKUM(NR,I)
      DO 4 I=1,24
4     WDOKUM(NR,I)=WDOKUM(NR,I)/Z
5     RETURN
65    END
```

SUBROUTINE ANSWER(R,WR,P,WP,XC)

```
* * * * *
*
* PO WYKONANIU PROCEDURY ANSWER W WEKTORZE P ZNAJDUJA SIE
5 * ADRESY DOKUMENTOW NALEZACYCH DO ODPOWIEDZI NA KWERENDE R,
* W WEKTORZE WP ZNAJDUJA SIE MIARY PODOBIENSTW,
* P(1) ZAWIERA LICZBE DOKUMENTOW W ODPOWIEDZI,
* WARTOSCIA PARAMETRU CX JEST WYMAGANA DOKLADNOSC ODPOWIEDZI .
*
10 * * * * *
    IMPLICIT INTEGER(A-U)
    DIMENSION R(25),WR(24),P(151),WP(151),Q(151),WQ(151)
    COMMON /OBSZAR/ TEZAU(41),INDEX(41,2),DESKRY(41,3)
    +/LISTA / PL(42),LIST(1250),WLIST(1250)
15    DO 1 I1=1,24
        IF(R(I1).EQ.0) GO TO 1
        NRL=1
        DO 10 J=NRL,41
            IF(I1.EQ.INDEX(J,1).AND.R(I1).EQ.INDEX(J,2)) GO TO 11
20    10 CONTINUE
        11 NRL=J
            K1=PL(NRL)
            K2=PL(NRL+1)-1
            J=1
25    DO 2 I2=K1,K2
            IF(WR(I1)*XC.GT.WLIST(I2)) GO TO 7
            J=J+1
            WP(J)=AMIN1(WR(I1),WLIST(I2))
            P(J)=LIST(I2)
30    2 CONTINUE
        7 P(1)=J-1
            IF(P(1).EQ.0) GO TO 6
            GO TO 4
        1 CONTINUE
35    4 I1=I1+1
            IF(I1.GT.24) GO TO 6
            DO 3 I=I1,24
                IF(R(I).EQ.0) GO TO 3
                DO 12 J=NRL,41
                    IF(I.EQ.INDEX(J,1).AND.R(I).EQ.INDEX(J,2)) GO TO 13
40    12 CONTINUE
                13 NRL=J
                    K1=PL(NRL)
                    K2=PL(NRL+1)-1
                    J=1
45    DO 5 I2=K1,K2
                    IF(WR(I)*XC.GT.WLIST(I2)) GO TO 8
                    J=J+1
                    WQ(J)=AMIN1(WR(I),WLIST(I2))
                    Q(J)=LIST(I2)
50    5 CONTINUE
        8 Q(1)=J-1
            IF(Q(1).EQ.0) GO TO 6
            CALL TUITU(P,WP,Q,WQ)
55    IF(P(1).EQ.0) GO TO 6
        3 CONTINUE
        6 RETURN
    END
```

SUBROUTINE TUITU(P,WP,Q,WQ)

```
* * * * *
*
* OBLICZANIE PRZEKROJU LIST INWERSYJNYCH P I Q.
*
* * * * *
*
* PO WYKONANIU PROCEDURY TUITU WEKTOR P ZAWIERA ADRESY
* DOKUMENTOW NALEZACYCH DO PRZEKROJU LIST INWERSYJNYCH.
*
* * * * *
*
* IMPLICIT INTEGER (A-U)
* DIMENSION P(151),Q(151),WP(151),WQ(151)
* N=P(1)+1
* M=Q(1)+1
* L=1
* DO 1 I=2,N
* DO 2 J=2,M
* IF(P(I).NE.Q(J)) GO TO 2
* L=L+1
* P(L)=P(I)
* WP(L)=WP(I)+WQ(J)
* GO TO 1
* 2 CONTINUE
* 1 CONTINUE
* P(1)=L-1
* RETURN
* END
```

```
      LOGICAL FUNCTION RELEW(R,WR,NR)
      * * * * *
      *
      *      BADANIE RELEWANTNOSCI DOKUMENTU O ADRESIE NR WZGLEDEM
5      *      KWERENDY R.
      *
      * * * * *
      *
10     *      JESLI DOKUMENT JEST RELEWANTNY TO FUNKCJA RELEW PRZYJMUJE
      *      WARTOSC .TRUE. W PRZECIWNYM PRZYPADKU WARTOSC .FALSE.
      *
      * * * * *
      *      IMPLICIT INTEGER(A-U)
      *      REAL MP
15     *      DIMENSION R(25),WR(24),T(25),WT(24),TP(25),WTP(24)
      *      COMMON /BAZA/ DOKUM(150,25),WDOKUM(150,24)
      *      +/IDBAZA/ ID(150,25),WID(150,24)
      *      XTT=0
20     *      DO 1 I=1,25
      *      T(I)=DOKUM(NR,I)
1      *      TP(I)=ID(NR,I)
      *      DO 2 I=1,24
      *      WT(I)=WDOKUM(NR,I)
25     *      2 WTP(I)=WID(NR,I)
      *      XRT=MP(R,WR,T,WT,.TRUE.)
      *      XTT=MP(T,WT,TP,WTP,.TRUE.)+MP(TP,WTP,T,WT,.FALSE.)
      *      XRTP=MP(R,WR,TP,WTP,.TRUE.)+MP(TP,WTP,R,WR,.FALSE.)
      *      IF(XRTP.GE.XRT) GO TO 3
      *      IF(XRTP.GE.XTT) GO TO 3
30     *      RELEW=.FALSE.
      *      GO TO 5
      *      3 RELEW=.TRUE.
      *      5 CONTINUE
      *      RETURN      $      END
```

```
REAL FUNCTION MP(R,WR,D,WD,PART)
```

```
* * * * *
*
* FUNKCJA MP OBLICZA MIARE PODOBIENSTWA KWERENDY R DO OPISU
* DOKUMENTU O ADRESIE NR.
*
* * * * *
```

```
IMPLICIT INTEGER(A-U)
```

```
LOGICAL PART
```

```
REAL MIARA,MI
```

```
DIMENSION R(25),WR(24),D(25),WD(24)
```

```
COMMON /OBSZAR/ TEZAUR(41),INDEX(41,2),DESKRY(41,3)
```

```
OD=1
```

```
MI=0.0
```

```
DO 60 I4=1,24
```

```
IF(R(I4).EQ.0) GO TO 60
```

```
DO 61 J1=OD,41
```

```
IF(I4.EQ.INDEX(J1,1).AND.R(I4).EQ.INDEX(J1,2)) GO TO 62
```

```
61 CONTINUE
```

```
62 OD=J1+1
```

```
K1=TEZAUR(J1)
```

```
MIARA=0.0
```

```
IF(PART) GO TO 69
```

```
J1=J1+1
```

```
IF(J1.GT.K1) GO TO 60
```

```
69 DO 63 IC=1,24
```

```
ROB=D(IC)
```

```
IF(ROB.EQ.0) GO TO 63
```

```
DO 64 JC=J1,K1
```

```
IF(IC.NE.INDEX(JC,1).OR.ROB.NE.INDEX(JC,2)) GO TO 64
```

```
MIARA=MIARA+WD(IC)
```

```
GO TO 63
```

```
64 CONTINUE
```

```
63 CONTINUE
```

```
MI=MI+AMIN1(MIARA,WR(I4))
```

```
60 CONTINUE
```

```
MP=MI
```

```
RETURN
```

```
END
```

