

INSTYTUT CYBERNETYKI TECHNICZNEJ
POLITECHNIKI WROCŁAWSKIEJ
Raport serii: PREPRINTY Nr 13/81

PODZIAŁ TEMATYCZNY BAZ DANYCH
W DOKUMENTACYJNYCH SYSTEMACH
INFORMACYJNYCH

/rozprawa doktorska/

Włodzimierz Wnuk

Promotor:

prof. dr hab. inż. Adam SIELICKI

Słowa kluczowe: system infor-
macyjny, dzielona baza
danych, algorytm grupo-
wania

SPIS TREŚCI

Wykaz oznaczeń i skrótów używanych w pracy	3
1. Wprowadzenie	
1.1. Charakterystyka dotychczasowego rozwoju dziedziny .	7
1.2. Charakterystyka automatyzowanych systemów informa- cyjnych	8
1.2.1. Systemy informacyjne faktograficzne	8
1.2.2. Systemy informacyjne dokumentacyjne	9
1.3. Umiejscowienie i cel rozprawy	11
2. Problem podziału bazy danych	13
2.1. Wstęp	13
2.2. Proces grupowania - pojęcia podstawowe, opis formalny	15
2.3. Sposoby dokonywania podziału	17
3. Analiza porównawcza wybranych grafowych algorytmów podziału	24
3.1. Definicje teorii grafów	25
3.2. Algorytmy podziału wykorzystujące MDS i NDH	30
3.3. Algorytmy podziału wykorzystujące MPP i KS	36
3.4. Metody wykorzystujące pojęcie k-spójności	39
3.5. Podsumowanie	42
4. Opis eksperymentów i analiza wyników algorytmów grupowania	46
4.1. Plan eksperymentów	46
4.2. Pojęcia wstępne	48
4.2.1. Podobieństwo między obiektami	48
4.2.2. Podobieństwo między czasopismami	53
4.2.3. Podobieństwo między grupowaniami	53
4.3. Analiza związków tematycznych między czasopismami .	55
4.4. Analiza proponowanych funkcji podobieństwa obiektów	61
4.5. Analiza wyników zastosowania algorytmów grupowania	64
5. Zakończenie	71
6. Bibliografia	75
Dodatek 1: Opis struktury cytowań	80
Dodatek 2: Własności bibliometryczne empirycznej bazy danych	82

Wykaz skrótów i oznaczeń używanych w pracy

Skróty:

b.d.	baza danych
m.w.i.	moduł wyszukiwania informacji
s.i.	system informacyjny
s.i.d.	dokumentacyjny system informacyjny
s.i.f.	faktograficzny system informacyjny
DF	definicja
F	fakt /twierdzenie bez dowodu/
WN	wniosek
MDS	minimalne drzewo spinające
MPP	maksymalny podgraf pełny
NDH	najkrótsza droga hamiltonowska
k-BL	k - blok
k-GW	k - gwiazda
k-K	k - koło
k-KL	k - klika
k-KS	k - komponenta spójna
k-SP	k - spojenie
k-FN	reguła k-ty najdalszy sąsiad
k-NN	reguła k-ty najbliższy sąsiad

Oznaczenia:

- 1/ wykaz oznaczeń nie zawiera standardowych oznaczeń matematycznych,
- 2/ oznaczenia nie wymienione w wykazie mają układ liniowy, tzn. każde z nich ma znaczenie lokalne.

Wykaz oznaczeń:

$Y = \{y_1, y_2, \dots, y_n\} = \{y \mid \varphi(y)\}$	skończony zbiór Y wszystkich takich elementów y, które spełniają funkcję zdaniową $\varphi(y)$,
$\text{card } Y = n$	moc zbioru Y
$\arg \varphi(x)$	argument spełniający funkcję zdaniową $\varphi(x)$
2^Y	rodzina wszystkich podzbiorów zbioru Y

$\text{diam } X$	średnica zbioru X
$A = \{A_1, \dots, A_m\}$	zbiór atrybutów, $m = \text{card } A$
$B = \{b_1, \dots, b_n\}$	zbiór obiektów bazy danych, $n = \text{card } B$
$P^q = \{P_1, P_2, \dots, P_q\}$	podział q -elementowy zbioru B , $\forall (i=1, 2, \dots, q)$ ($\text{card } P_i = n_i$)
q	liczba grup
$R_p \subset B \times B$	relacja podobieństwa
$\Omega_k, k=1, 2, \dots$	relacja k -cytowania
$p(b_i, b_j)$	funkcja podobieństwa obiektów b_i, b_j / $p : B \times B \longrightarrow R$ /
$d(b_i, b_j)$	funkcja odległości obiektów b_i, b_j /dla metryki $d(b_i, b_j) := \varphi(b_i, b_j) /$
$p_g(k)$	podobieństwo grafowe obiektów
$p_r(k)$	podobieństwo k -retrospektywne obiektów
$p^*(P', P'')$	funkcja podobieństwa podziałów P', P''
$p^\circ(P', P'')$	funkcja podobieństwa podziałów słabych P', P''
$\Upsilon(P', P \text{ wzorcowe})$	średnia precyzja podziału P'
$\Theta(P', P'')$	łączna miara zgodności podziałów P', P''
$\triangleq$	symbol równości definicyjnej
$:=$	symbol podstawiania
$K(P^q)$	kryterium oceny podziału P^q
$O(n^k)$	złożoność algorytmu /maksymalna liczba operacji dla n obiektów/ jest ograniczona od góry wielomianem k -tego stopnia
$V(G), E(G)$	zbiór wierzchołków, krawędzi grafu G
$\mu : E(G) \longrightarrow R$	funkcja obciążeń krawędzi
$D(v_i, v_j)$	droga między wierzchołkami v_i, v_j
$\tilde{D}(v_i, v_j)$	długość drogi $D(v_i, v_j)$
$D^\Sigma(v_i, v_j)$	obciążenie drogi $D(v_i, v_j)$
$r(G)$	promień grafu
v^*	wierzchołek centralny

$\text{diam } G$	średnica grafu G /spójnego/
$\text{deg } v_i$	stopień wierzchołka v_i /liczba krawędzi przyległych do v_i /
$C(G_1, G_2)$	przekrój grafu G
$\nabla(G_1, G_2)$	szerokość przekroju $C(G_1, G_2)$
$\lambda(G_1, G_2)$	wieź przekroju $C(G_1, G_2)$
T_G	drzewo grafu G
e^*	krawędź niezgodna, $e^* \in E(G)$
$G(t)$	t -ekstrakcja grafu G

W pracy przeanalizowano możliwości dokonywania podziału tematycznego dokumentacyjnej bazy danych przy pomocy grafowych algorytmów grupowania. Zaproponowano cztery współczynniki podobieństwa między obiektami bazy danych poparte analizą struktury odesłań między pracami. Takie podejście umożliwia grupowanie dokumentów bez potrzeby ich manualnego indeksowania słowami kluczowymi. Na empirycznej bazie danych przetestowano pracę dziesięciu grafowych algorytmów grupowania. Wynikowe podziały tematyczne porównano z grupowaniami ekspertów przy pomocy zmodyfikowanej miary Randa. Eksperymenty wykazały, że uzyskane podziały są precyzyjniejsze niż podziały według słów kluczowych. Są one również stabilne. Przetestowano wpływ selekcji opisów, uwzględnianej głębokości cytowań i stopnia spójności grup na podziały wynikowe.

1. WPROWADZENIE

Każdy proces decyzyjny będący składową takich działań jak projektowanie, zarządzanie, itp. wymaga częstego oraz sprawnego wyszukiwania informacji, ponieważ wartość osiągniętego wyniku /projektu/ silnie zależy między innymi od zasobu posiadanych informacji oraz umiejętności i szybkości ich wykorzystania.

Jednakże wyraźnie zauważalna jest sprzeczność między chęcią zwiększenia prawdopodobieństwa znalezienia użytecznej informacji a dążeniem do zmniejszenia pracochłonności poszukiwań. Ponadto indywidualne poszukiwania literaturowe są niezwykle czasochłonne. Wobec tego powstały, bardzo dziś wyspecjalizowane i rozbudowane, ośrodki informacyjne /uczelniane, branżowe, patentowe/, świadczące usługi związane z gromadzeniem, wyszukiwaniem i udostępnianiem dokumentów. Poprzez wykorzystanie stwierdzonych prawidłowości bibliograficznych oraz rozsadną automatyzację typowych, powtarzalnych procedur dokumentacyjno-wyszukiwawczych można osiągnąć znaczny wzrost jakości tych usług.

Fakty przytoczone w [30, 52, 61, 82, 85, 116, 131] upoważniają do stwierdzenia, że:

- konieczna stała się automatyzacja wybranych procesów gromadzenia i wyszukiwania informacji, ponieważ manualne poszukiwania według pewnego profilu choć tanie, wygodne, niezastąpione i dające dobre wyniki w przypadku małej ilości nietypowych pytań informacyjnych są zbyt żmudne i czasochłonne,

- zagadnienia dotyczące automatyzowanych systemów informacyjnych są ważne, aktualne i trudne,

- istnieje duże zapotrzebowanie na usługi informacyjne mimo tego, że ich jakość nie zawsze jest zadowalająca.

Powszechnie akcentowanymi zaletami automatyzowanych systemów informacyjnych są: szybkie przeszukiwanie dużych baz danych /oszczędność czasu/, łatwość obsługi złożonych pytań informacyjnych, możliwość "oprofilowania" zainteresowań /selektywność/, możliwość zmagazynowania rezultatów wyszukiwań dla przyszłych klientów systemu; natomiast wadami pozostają jeszcze: niekiedy wyższy koszt obsługi [33, 36], niepełność baz danych /od 1970 r./, brak reprezentacji pewnych dyscyplin w bazach danych, subiektywne trudności współpracy człowiek-maszyna.

1.1. Charakterystyka dotychczasowego rozwoju dziedziny

Aktualnie ukazuje się kilkadziesiąt periodyków omawiających zasadnicze zagadnienia systemów gromadzenia i wyszukiwania informacji. Już z tego tylko powodu nie można wyczerpująco przedstawić historycznego rozwoju tej dziedziny.

Przyjmijmy następujący podział automatyzowanych systemów informacyjnych:

- systemy informacyjne faktograficzne /s.i.f./;
odpowiadają na pytania typu: podać identyfikatory obiektów w bazie danych /m.in. dokumentów/ o wymienionych cechach: $\alpha, \beta, \dots, \omega$ / $\alpha, \beta, \dots, \omega$ - dane a'priori/,
- systemy informacyjne dokumentacyjne /s.i.d./;
odpowiadają na pytania typu: podać identyfikatory obiektów najbliższych /podobnych/, w określonym sensie, obiektowi wzorcowemu [~].

Wśród prekursorów tworzących w latach sześćdziesiątych i siedemdziesiątych XX w. podstawy funkcjonowania s.i.d. należy wymienić: G. Saltona, J.A. Swetsa, J. Farradane'a, H. Borko, R.A. Fairthorne'a, C.W. Cleverdon, M.E. Marona, S.E. Robertsona, C.J. Rijsbergena, K. Spärck-Jones, L.B. Doyle'a, B.C. Vickery'ego, D.J. Hillmana.

W Polsce, rozwój nauki o s.i.f. zawdzięczamy m.in. pracom /z lat siedemdziesiątych/ Z. Pawlaka, W. Marka, M. Dąbrowskiego, W. Lipskiego, W.M. Turskiego; niezależnie od bardzo podobnych prac prowadzonych na Zachodzie przez np. E. Wonga, S.P. Ghosha. W analizie funkcjonowania i budowie modeli opisujących s.i.d.

wykorzystywano: probabilistykę [50, 99, 119, 124] a w niej: teorię podejmowania decyzji [63, 64, 65, 70, 76] i teorię obsługi masowej [91], teorię mnogości i logikę [26, 32, 77, 80, 92], teorię grafów [4, 45, 74], lingwistykę [111, 122], algebrę [53], teorię informacji [17, 62], teorię klasyfikacji [39, 109, 116, 147], teorię zbiorów rozmytych [86, 93, 94, 97, 105, 125].

Jest rzeczą zastanawiającą, że mimo dużego zaangażowania zasobów kadrowych i materialnych w stosunkowo długim okresie czasu /ok. 15 lat/ nie dopracowano się efektywnych s.i.d. Częściowo można to wytłumaczyć zbyt optymistycznymi prognozami dotyczącymi automatycznego przekładu tekstu. Ponadto procesy wyszukiwania /poświęcono im najwięcej uwagi/, które są w pewnym sensie lokalnie najefektywniejsze, potęgują trudności związane z procesami porządkującymi system /np. wymuszają konieczność gradacji znaczeń atrybutu obiektu, określenia stopnia korelacji między atrybutami/. Eksploatowanymi systemami dokumentacyjnymi są obecnie systemy dystrybutorskie, które zazwyczaj wykorzystują technikę przeszukiwania pełnego.

1.2. Charakterystyka automatyzowanych systemów informacyjnych

1.2.1. Systemy informacyjne faktograficzne

Syntetyczne, choć niewyczerpujące, przedstawienie metod organizacji m.w.i. zawierają monografie [27, 59, 10]. Poprzestaniemy na omówieniu głównych cech s.i.f.:

1/ zalety:

- uniwersalność /jej skutkiem jest jednak niższa efektywność systemu/,
- przejrzystość i prostota, które są wymuszone rygorystycznymi wymaganiami dotyczącymi postaci pytań i opisów,

2/ wady:

- zazwyczaj stosunkowo duża redundancja zapisów,
- pomijanie szeregu zależności między deskryptorami,
- uzyskiwanie poprawy parametrów systemu drogą nadmiernej rozbudowy modułu porządkującego s.i., poprzez precyzyjniejszy tezaurus, dokładniej określone relacje, większą adekwatność opisów, itp.,

- brak mechanizmów samoregulacji,
- zazwyczaj znacznie utrudniona aktualizacja i modyfikacja bazy danych m.w.i.

W modelach matematycznych s.i.f. zakładamy milcząco, że umiemy tworzyć opisy obiektów, opisy te są adekwatne i uwzględniają specyfikę środowiska, potrafimy określać relacje hierarchiczności, posiadamy dobry tezaursus /równomiernie "gęsty", z ustalonymi deskryptorami głównymi, pomocniczymi, zakazanymi, z usuniętą homo- i synonimią/.

Rodzajem studium porównującego parametry funkcjonalne modeli s.i.f. są prace [27, 109, 116].

Nieefektywność stosowania w m.w.i. algebry boole'owskiej pokazano w [129], a nieefektywność /objętość pamięci/ organizacji zwartej Ghosha w [132] i pośrednio w [32, 73]. Listy odwrócone są nieefektywne dla przypadków gdy: a/ średnia liczba cech w opisie obiektu jest duża, b/ średnia liczba cech w pytaniu jest duża, c/ średnia liczba obiektów o tej samej cesze jest duża. Multilista jest nieefektywna czasowo z wyjątkiem specyficznych przypadków gdy: a/większość cech jest skojarzona z bardzo małą liczbą obiektów, b/ liczba cech w opisie obiektu jest bardzo mała, c/ liczba cech w pytaniu jest bardzo mała.

Reasumując: wymienione modele dają przesłanki do realizacji s.i.f. funkcjonujących efektywnie dla zwartej tematycznie, statycznej /mało aktualizacji/ i niezbyt licznej bazy danych.

1.2.2. Systemy informacyjne dokumentacyjne

Cechami charakterystycznymi s.i.d. są:

- duża liczba obiektów w bazie danych /rzęd 10^6 /,
- duża liczba cech obiektów /rzęd 10^4 /,
- duża dynamika bazy danych /duża częstość aktualizacji - dopisań, kasowań/,
- zmiennność z czasem znaczeń opisów obiektów,
- możliwość określenia nietrywialnej relacji podobieństwa semantycznego między obiektami oraz cechami opisów,
- mniejsze znaczenie ma zagadnienie zabezpieczenia bazy danych przed niepożądanym dostępem, natomiast duże znaczenie mają: automatyzacja procesów indeksowania [20, 99, 108, 147],

kompresja zapisów [25, 48, 75], konwersja baz danych [49], rozwój sieci komputerowych i tryb pracy on-line /możliwość "strojenia" pytań/, przetwarzanie rozproszone.

Warto podkreślić, że popularne w systemach dokumentacyjnych sposoby organizacji b.d. są zdegenerowanymi przypadkami podziału /np. listy odwrócone : $P^m = \{P_j | (b_i \in P_j) \Leftrightarrow (\exists (k \in \{1, 2, \dots, m\}) (b_{ik} = d_j \in D)) \}$ $j = 1, 2, \dots, m$ - cecha generuje grupę; listy proste $P^n = \{P_j | P_j = \{b_j\} \}$ $j = 1, 2, \dots, n$ - obiekt generuje grupę/.

Poniżej podano jakościową charakterystykę typowych metod organizacji b.d., z której wynika, że dzielenie b.d. jest metodą kompromisową łączącą zalety przeciwstawnych typów: struktury list odwróconych i prostych.

cechy typ struktury	typ pytań informacyjnych	wymagana pamięć	czas wyszuki- wania	trudności dopisa- nia/skasowania informacji	"sztywność struktury"	nieprecyzja odpowiedzi	pracochłonność eksploatacji	uzyskanie dodatkowych informacji
listy proste	$p(x,y) > \alpha$	1	5	1/4	1	1	1	-
listy odwrócone	$p(x,y) = \max$	5	1	2/2	4	2	2	-
listy liniowe	$p(x,y) = \max$	1	3	4/1	5	1	4	-
listy haszowe	$p(x,y) = \max$	2	2	3/2	4	2	4	-
dzielona baza danych	$p(x,y) > \alpha$	2	3	2/3	2	3	3	+

Skala:

- | | |
|-----------------|---------------------|
| 1 - bardzo małe | + - jest możliwość |
| 2 - małe | - - brak możliwości |
| 3 - średnie | |
| 4 - duże | |
| 5 - bardzo duże | |

Przytoczmy, na koniec, opinię Lancastera:

..."Jest możliwe, że wykorzystanie logiki boole'owskiej w automatyzowanych s.w.i. było błędem ..., którego przyczyną jest podobieństwo zasady działania pierwszych zautomatyzowanych systemów do funkcjonowania kart przeziernych lub obrzeźnie perforowanych. W systemach tego typu nie ma możliwości zadawania granicznego współczynnika zgodności pytania i dokumentu, jak również szeregowania dokumentów w odpowiedzi nierosnąco względem tej zgodności. Dla systemów dokumentacyjnych takie ograniczenia nie mają sensu..."

1.3. Umiejscowienie i cel rozprawy

Okazuje się, że współczesne systemy selektywnego rozpowszechniania informacji, wykorzystujące logikę boole'owską i strategię poszukiwania pełnego, nawet w przypadku skomplikowanego oprofilowania pytania informacyjnego, udzielają odpowiedzi o niezadowalającej pełności i precyzji. Systemy te korzystają z tezaurusów słów kluczowych. Niestety zazwyczaj czasopisma przy indeksowaniu prac posługują się różnymi tezaurusami i regułami indeksowania ponieważ prace normalizacyjne nie zakończyły się pełnym sukcesem. Jakkolwiek znane są metody pozwalające w tym przypadku polepszać efekt wyszukiwania poprzez: redukcję ogólności pytania, rozłożenie na pytania prostsze, wykorzystanie asocjacji słów, korekcji pytania, itp. [109, 116], to pojawia się pytanie:

- czy takie efekty wyszukiwań nie wynikają właśnie z przyjęcia jako opisu dokumentu ciągu słów kluczowych.

Dokument, jak wiemy, składa się z:

- a/ personaliów autora /nazwisko, imię, nazwa instytucji, itp./,
- b/ klasyfikacji tematycznej /słowa kluczowe, kody klasyfikacji, itp./,
- c/ tytułu,
- d/ abstraktu,
- e/ treści pracy,
- f/ bibliografii /cytowania/.

W dotychczasowych eksperymentach wykorzystywano z reguły do indeksowania słowa wybierane z b/, c/, d/. Wykorzystaniu cytowań do indeksowania poświęcono mało uwagi. Salton [106] wybierał słowa znaczące z tytułów prac cytowanych i dołączał do zestawu słów kluczowych..

W pracach [68, 84] pokazano, że może to prowadzić do pogorszenia jakości odpowiedzi. Small [114, 115] pokazał, że istnieje współzależność między słowami kluczowymi a bibliografią pracy. Schiminovich [110] i Eaton [12] wykorzystywali cytowania bezpośrednie /1 - cytowania/ do wyodrębniania: monotematycznych grup dokumentów. Ich wyniki były zachęcające, ale sprawdzone na pracach z fizyki - dziedziny o długiej historii. W pracy [2] próbowano grupować czasopisma z dziedziny nauk społecznych.

Zauważmy, że tylko punkty a/, c/, e/, f/ i niekiedy d/ zawierają informacje pochodzące bezpośrednio od autora pracy.

Warte sprawdzenia jest, czy najlepszym indeksującym nie jest może sam autor pracy.

W niniejszej pracy, która powstawała w latach 1974-80, przedstawiono koncepcję podziału bazy danych s.i.d. na moduły tematyczne przez wykorzystanie pewnych logicznych i semantycznych zależności między jej obiektami. Do ustalenia tych zależności zastosowano metody grupowania /ang. cluster analysis/. W pracy korzysta się /z pewnymi wyjątkami/ z aparatu kombinatoryki i teorii grafów. Spośród opracowań dotyczących grupowania wykorzystano częściowo monografię Anderberga [1] i Durana [31].

Najbliższymi tematycznie publikacjami w literaturze obcej są prace Croucha [24], Schiminovicha [110], Kwoka [68], Crofta [22], Bichtelera [12, 13], Smalla [114, 115], a w polskiej literaturze fachowej monografia M. Dąbrowskiego i K. Laus [27] /str. 23-26, 49-94, 121-129/.

Celem pracy jest obrona następującej tezy:

Na podstawie pewnych parametrów opisu obiektów dokumentacyjnej bazy danych można określić adekwatne współczynniki podobieństwa między tymi obiektami. Pozwala to dokonać podziału tematycznego bazy danych grafowymi metodami grupowania. Taki podział może mieć korzystny wpływ na parametry użytkowe dokumentacyjnego systemu informacyjnego.

Prawdziwość tej tezy miałaby duże znaczenie np. dla procesów automatycznego indeksowania i wyszukiwania dokumentów w systemach wyszukiwania informacji /polepszenia precyzji odpowiedzi, szerszy zestaw obsługiwanych pytań informacyjnych, samoaktualizacja podziału b.d. w czasie, itp./.

Zadania cząstkowe wykonane dla obrony tezy to:

- zweryfikowanie hipotez o:

- a/ adekwatności zawartości informacyjnej obiektu b.d. i jego opisu /rozszerzonej w określony sposób bibliografii/,
- b/ tematycznym grupowaniu się obiektów w niezdegenerowanej b.d.,

- zbadanie własności zdefiniowanych miar podobieństwa i zestawu grafowych algorytmów podziału.

Po sformułowaniu zadania podziału tematycznego i przedstawieniu sposobów dokonywania podziału /rozdział 2/ w rozdziale następnym poddano analizie zestaw wybranych grafowych algorytmów grupowania /tzw. algorytmy niehierarchiczne/. Część z nich będzie wykorzystywana w przebiegu eksperymentów na rzeczywistej, wyselekcjonowanej bazie danych. Pewne parametry tej b.d. zbadano w Dodatku 2. Na podstawie przyjętej postaci opisu obiektów z bazy danych zdefiniowano /rozdz. 4/ trzy funkcje podobieństwa określone na parach obiektów. W rozdziale 4 opisano przeprowadzone testy i przeanalizowano ich wyniki. Testy te miały dać podstawę do oceny: własności zdefiniowanych miar podobieństwa, własności grafowych algorytmów grupowania, cech uzyskanego podziału tematycznego.



2. PROBLEM PODZIAŁU BAZY DANYCH

2.1. Wstęp

Baza danych /b.d./ jest integralną częścią systemu gromadzenia i wyszukiwania informacji /s.i./. Jest ona zbiorem opisów obiektów, wśród których poszukujemy co pewien czas opisów zgodnych /relewantnych/ z pytaniami informacyjnymi. Od zawartości i sposobu organizacji b.d. w sposób zasadniczy zależy efektywność s.i. O ile dla b.d. typu faktograficznego opracowano wiele struktur fizycznych danych zorientowanych na szybkość wyszukiwania /listy inwersyjne/ bądź na oszczędność przechowywania danych /listy proste/, o tyle dla b.d. typu dokumentacyjnego /szczególnie o znacznych rozmiarach/ zagadnienie właściwej, efektywnej organizacji b.d. jest nadal otwarte.

Stwierdzono empirycznie / [13, 22, 68, 106, 108, 109, 114, 116]/ następujące fakty:

- 1/ opisy deskryptorowe w odpowiednio licznej b.d. mają tendencję do koncentrowania się w dość regularne, zwarte grupy,
- 2/ użytkownik potrafi określić wzorcowy obiekt zainteresowań A,
- 3/ pytania informacyjne, z powodu specjalizacji tematycznej użytkowników s.i.d., koncentrują się w prawie rozłącznych grupach,
- 4/ preferencje użytkowników nieznacznie zmieniają się w czasie,
- 5/ zazwyczaj bardzo trudno określić wagi deskryptorów,
- 6/ istnieje zgodność między częstością korzystania z publikacji a częstością jej cytowania,
- 7/ fakty wymienione w Dodatku 2 do niniejszej pracy.

Uwzględniając to, że s.i.d. obsługuje polecenia typu: znaleźć x opisów w określony sposób najbliższych opisowi A oraz fakty 1÷4 można sformułować hipotezę, że podział b.d. na moduły tematyczne i przydzielenie pytań informacyjnych do odpowiednich modułów pozwoli osiągnąć mały czas wyszukiwań /mniejsze zbiory do przeszukania/ i rozsądną wielkość pamięci potrzebnej do przechowania b.d. /redukcja wielokrotnych zapisów/ przy prawie niezmiennych pozostałych parametrach s.i.d. /np. zawartość odpowiedzi systemu/.

Natomiast na podstawie faktów 5÷7 można przypuszczać, że określenie podobieństwa, a co za tym idzie - dokonanie podziału, na podstawie związków bibliograficznych będzie wygodniejsze i nie pogorszy wyników s.i.d. w porównaniu z s.i.d. wykorzystującym tradycyjny tezaurus słów kluczowych.

Konieczność podzielenia b.d. na moduły wynika z dużych rozmiarów dokumentacyjnych b.d. oraz z możliwości technicznych współczesnych maszyn cyfrowych. Możliwe są co najmniej trzy typy podejścia do rozwiązania tego zagadnienia. Ich istotą jest preferowanie:

- a/ zawartości tematycznej b.d. - podział b.d. na grupy obiektów wzajemnie tematycznie podobnych /zakładamy, że pytanie informacyjne można złożyć z fragmentów jednorodnych tematycznie/,
- b/ parametrów technicznych sprzętu cyfrowego - podział b.d. na grupy o równej liczności /bloki, stronice, itp./ w celu minimalizacji /przy założeniu równomiernego rozkładu częstości pytań informacyjnych/ liczby zwrotów do pamięci zewnętrznej [71] co, jak wiadomo, ma decydujący wpływ na szybkość wyszukiwania,
- c/ zainteresowań użytkowników /np. metoda Yu [109] / - wyodrębnionym grupom użytkowników o wspólnych zainteresowaniach przypo-

rzadkowuje się części b.d., które zawierają obiekty relewantne do pytań informacyjnych tej grupy. Pewne obiekty b.d. mogą nie zostać jednak przydzielone. Ponadto zniekształca się zależnośći tematyczne w b.d. Zyskiem jest obniżenie prędkości klasyfikacji, ponieważ liczba użytkowników jest zazwyczaj dużo mniejsza od liczności b.d.

Następne rozdziały dotyczą głównie podziału tematycznego b.d.

2.2. Proces grupowania - pojęcia podstawowe, opis formalny

Przyjmijmy następujące oznaczenia:

$A = \{A_1, \dots, A_m\}$ - zbiór atrybutów /niekiedy utożsamiany z R_+^m /,

$D_i = \{b_{i_1}, \dots, b_{i_{k_i}}\}$; $i = 1, 2, \dots, m$ - zbiór wartości /deskryptorów/ każdego z atrybutów,

$B = \{b_1, \dots, b_n\}$ - zbiór obiektów stanowiących bazę danych

$b_i \triangleq \langle b_{i_1}, \dots, b_{i_m} \rangle \in \prod_{j=1}^m D_j / \forall (j=1, 2, \dots, m) (b_{i_j} \in D_j) /$ -
- opis obiektu

$Z = \{z_1, \dots, z_q\}$ - zbiór nazw /symboli grup/,

$\xi: \prod_{j=1}^m A_j \rightarrow Z$ - przekształcenie, operator grupowania
najczęściej $\xi^{-1}(Z) \subset \prod_{j=1}^m A_j$ /,

$P_i = \{b_j \in B \mid \xi(b_j) = z_i\}$ - klasa /grupa/ - element podziału o nazwie z_i

Zdefiniujmy:

DF 2.2.1. $P_q = \{P_1, P_2, \dots, P_q\}$ jest q-podziałem /mocnym/ zbioru B
jeśli:

1/ $\forall (i, j = 1, 2, \dots, q ; i \neq j) (P_i \cap P_j = \emptyset, P_i \neq \emptyset, P_i \subset B)$,

2/ $\bigcup_{i=1}^q P_i = B$.

W podziale składowym dopuszcza się nierozłączność grup / $P_i \cap P_j \neq \emptyset$ /

Jednym z najważniejszych pojęć w grupowaniu jest pojęcie podobieństwa. Można definiować je na wiele sposobów [31, 118].

W niniejszej pracy przyjmujemy następujące określenie:

DF 2.2.2. a/ Relacja podobieństwa $R_p \subset B \times B$ jest to relacja zwrotna, symetryczna wyznaczana przez progową wartość funkcji podobieństwa.

b/ Funkcja podobieństwa $p : B \times B \rightarrow [0,1] \subset \mathbb{R}$ jest to funkcja symetryczna spełniająca warunek:

$$(b_i, b_j \in B) (p(b_i, b_i) \triangleq \max_{b_j \in B} \{p(b_i, b_j)\} \triangleq 1).$$

Sformułowanie zadania podziału /grupowania/:

- konstruktywnie /np. w postaci algorytmu/ zdefiniować przekształcenie $\mathcal{E}$ /na co najmniej B /, tak, aby spełniona była implikacja:

$$(*) \quad \forall (i, j, k, l) [(b_i, b_j \in P_l \wedge b_k \notin P_l) \Rightarrow (p(b_i, b_j) > \max \{p(b_i, b_k), p(b_j, b_k)\}) ,$$

gdzie:

$p(x, y)$ - współczynnik podobieństwa między obiektami x i y ; $P_l, P_k \in \mathcal{P}_q$

A więc podział b.d. to zorganizowanie pewnej struktury logicznej na jej elementach, tzn. pogrupowanie tych elementów. Grupowanie /clustering/ jest techniką klasyfikacji, dla której:

1. semantyka zagadnienia określona jest przez podanie podobieństw między obiektami,
2. zwykle parametry wszystkich obiektów podlegających grupowaniu znane są przed procesem grupowania,
3. nie ma a'priori zadanego ciągu uczącego,
4. podczas grupowania należy określić reprezentantów /nazwy/ klas.

Podobieństwo nie jest relacją typu równoważności /nieprzechodność!/ i dlatego B może nie podzielić się na rozłączne grupy /niespełnienie warunku 1/ z DF 2.2.1/. Wobec tego, przy szukaniu podziałów mocnych, utożsamia się zazwyczaj zbiór A z $\mathbb{R}_+^m$, w której zdefiniować można metrykę $\varphi : \mathbb{R}_+^m \times \mathbb{R}_+^m \rightarrow \mathbb{R}_+ \cup \{0\}$. Indukuje ona porządek na zbiorze par. O wyborze rodzaju metryki decydują własności uniwersum $\mathcal{B} \supset B$ oraz prostota i łatwość obliczeń.

W niektórych przypadkach można podać wzory [27, 118] pozwalające obliczać wartość podobieństwa na podstawie metryki oraz funkcji

odległości /w sensie Soergela/ na podstawie podobieństw /odwrotna proporcjonalność/.

Należy podkreślić, że chociaż w s.i. cechy w opisach obiektów mogą mieć różny charakter /jakościowe, ilościowe, ciągłe, dyskretne, itp./, to cechy w s.i.d. zwykle są jednorodne /najczęściej binarne/, co znacznie upraszcza analizę i obliczenia.

Pożądanymi własnościami miar /współczynników/ podobieństwa są m.in.:

- przejrzysta interpretacja fizyczna, geometryczna lub intuicyjna,
- niezależność od przekształceń skali cech /liniowe, multiplikacyjne/,
- duża "zdolność rozdzielająca" współczynnika,
- prostota i łatwość obliczeń, w tym łatwość dokonywania aktualizacji przy zmianie zestawu cech lub deskryptorów.

Wyczerpujący przegląd współczynników podobieństwa: obiekt -

- obiekt, grupa - grupa, cecha - cecha wraz z analizą ich własności oraz próbą systematyzacji przedstawiono w pracy [138].

Przedstawiono tam również sposoby tworzenia profili /reprezentacji/ grup, dla przyspieszania pracy algorytmów grupowania.

2.3. Sposoby dokonywania podziału

Efektem działania algorytmu podziału b.d. powinno być znalezienie podziału spełniającego pierwotne wymagania (*) /p. 2.2./.

Ponieważ podobieństwo nie jest ultrametryką, a zbiór rozwiązań dopuszczalnych jest dyskretny /być może pusty/ należy dokonać pełnego przeglądu wariantów. Metoda ta, przy spotykanych wielkościach b.d. i obecnych możliwościach technicznych jest praktycznie bezużyteczna.

Liczba $S(n)$ sposobów podziału zbioru n -elementowego na $q = 1, 2, \dots, n$, niepustych, rozłącznych podzbiorów jest rozwiązaniem rekurencji:

$$\begin{cases} S(n, q) = q S(n-1, q) + S(n-1, q-1), \\ \underline{S(1, 1) = 1 \quad ; \quad S(1, n) = 0 \quad \text{dla } n \neq 1} \end{cases}$$

i wynosi [60]:

$$S(n) = \sum_{q=1}^n S(n,q) = \sum_{q=1}^n \frac{1}{q!} \sum_{k=0}^n (-1)^{q-k} \binom{q}{k} k^n$$

$$\text{/np. } S(100) > 10^{67}/$$

Gdy grupy mają jednakową liczbę elementów $/kq = n/$, to:

$$S(n|k) = \frac{n!}{(k!)^q q!} ; \text{ i np. } S(2015) \approx 10^{10}$$

Nawet gdy $D \subset \mathbb{R}$ /zbiór dyskretny n-elementowy na prostej, $m = 1/$, a grupy są spójne, liczba podziałów na q-grup $T(n,q)$ wynosi:

$$T(n,q) = \binom{n-1}{q-1} \text{ i np. } T(n,2) = n-1 \ll 2^{n-1} - 1 = S(n,2)$$

Przypadek A: $\mathbb{A}^m$ jest przestrzenią metryczną i znany jest wektorowy opis każdego obiektu

Wobec powyższych trudności obliczeniowych, wzmacnia się warunek (* /p.2.2) przyjmując, że prawdziwa jest równoważność:

$$(**) (\varphi(b_i, b_j) \geq \varphi(b_k, b_l)) \Leftrightarrow (p(b_i, b_j) \leq p(b_k, b_l))$$

Ponadto wprowadza się pewne kryterium K jakości podziału. Ponieważ nie ma kryterium adekwatnego wymaganiom (*), jego wyboru dokonuje się arbitralnie /przegląd kryteriów w pracy [138] /.

Najczęściej spotykanymi zadaniami zastępczymi są:

- I/ szukanie wśród $P^q /q = 2, \dots, n-1/$ podziału P^* optymalizującego K ,
- II/ szukanie wśród P^q /q ustalone a'priori/ podziału P^* optymalizującego K .

Zadanie I/ można praktycznie rozwiązać dla przypadku, gdy B i A są małoliczne. Zadanie II/ rozwiązuje się najczęściej metodą iteracyjnej korekcji $P_{(i)}^q \rightarrow P_{(i+1)}^q$ monotonicznie zmniejszającej wartość $K(P^q)$, co nie gwarantuje /mimo otrzymania podziału stabilnego/ uzyskanie optymalnej wartości $K(P^q)$. Zauważmy ponadto, że liczba podziałów sąsiednich /różniących się nazwą tylko jednego obiektu/ podziałowi $P^q = \{P_1, P_2, \dots, P^q\}$ wynosi $n(q-1)$, gdy $\forall (i = 1, 2, \dots, q) (\text{card } P_i > 1)$. Wobec czego pełny przegląd rozwiązań jest nierealizowalny dla $n > 20$, natomiast programowanie

dynamiczne jest nieskuteczne dla $n \geq 100$.

Można wyróżnić następujące typy metod grupowania:

- 1/ aglomerujące [np. 1, 46, 54, 67, 69, 103, 133],
- 2/ dzielące /dekomponujące/ [1, 31, 67],
- 3/ inne [np. 14, 40, 55], np. programowanie całkowitoliczbowe [130]/.

Ponadto [98] metody te mogą być:

- a/ heurystyczne /strukturalne/,
- b/ ekstremalne deterministyczne,
- c/ probabilistyczne.

Inny podział metod grupowania podano w [31].

Każda z powyższych procedur może mieć realizację szeregową /wynik zależy od kolejności rozpatrywania danych/ lub równoległą.

Główne typy zadań podziału i rozwiązujących je procedur można przedstawić w następującej tablicy morfologicznej:

a		b		c	
1	2	1	2	1	2
R	S/R	S/R	S	H	H

gdzie:

- 1 - małoliczne b.d. /rząd kilkuset obiektów/,
- 2 - duże b.d. /rząd kilkudziesięciu tysięcy obiektów/,
- a - liczba grup q znana a priori,
- b - q nieznane i należy je określić,
- c - budowanie hierarchii na B ,

R, S, H - procedury odpowiednio: równoległe, szeregowo, hierarchiczne.

Opisy najważniejszych procedur R i S są zawarte m.in. w [1, 31, 67, 98, 138, 140]. Zadanie zastępcze II odpowiada podziałowi b.d. ze względu na parametry techniczne sprzętu cyfrowego /patrz p. 2.1 (b) / i nie będzie tu omawiane.

Przy podziale tematycznym b.d. może wystąpić zadanie I /liczba grup q nieznana/. $K(P^q, q)$ jest układem kryteriów, czyli dokonujemy polioptymalizacji. Jako kryteria cząstkowe zazwyczaj przyjmuje się:

$$K_r(P^q) \triangleq \left(\frac{1}{n} \sum_{i=1}^n \left(\frac{k(b_i)}{n} \right)^r \right)^{1/r} - \text{koncentracja obiektów w podziale } P^q / \text{nierosnąca funkcja liczby } q/$$

gdzie:
 $k(b_i)$ - liczność grupy zawierającej $b_i \in B$, $r \in \mathbb{R}_+$

$$R_r(P^q) \triangleq \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{n_i} \sum_{b \in P_i} \varrho^r(b_i, b) \right)^{1/r} - \text{rozrzut odległości w grupach /nierosnąca funkcja } q/$$

gdzie:
 P_i - grupa, do której należy b_i .

Praktyka pokazuje, że podział b.d. nie musi być zbyt dokładny, natomiast ze względu na liczność obiektów preferowane są metody o małej złożoności obliczeniowej. Warunki te spełniają m.in. heurystyczne metody niehierarchiczne - warstwowe /hierarchia jest natomiast pożądana przy grupowaniu deskryptorów/.

Na podstawie analizy kilkudziesięciu niehierarchicznych algorytmów podziału [138] autor stwierdził, że daje się w nich wyodrębnić następujące, specyficzne etapy:

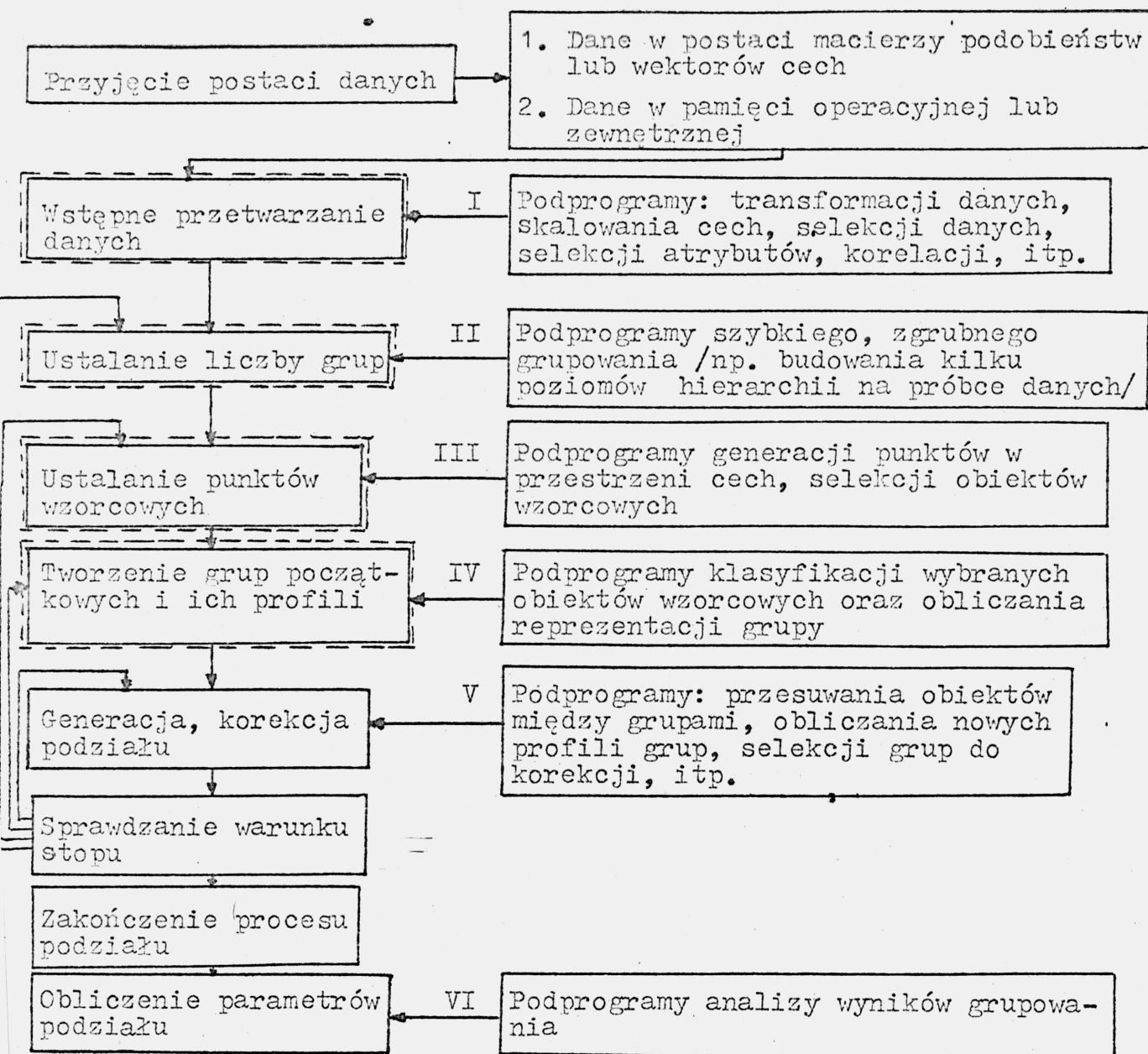
- 1/ faza wstępna /ustalenie liczby grup q , transformacje danych, selekcja danych, itp./,
- 2/ szukanie załączków grup /ustalenie punktów wzorcowych/,
- 3/ tworzenie grup początkowych,
- 4/ generacja i/lub korekcja podziału oraz modyfikacja profili jego elementów,
- 5/ sprawdzanie warunku stopu /stabilność podziału, optymalność kryterium, zadana liczba iteracji, itp./,
- 6/ zakończenie procesu podziału /ostateczne formowanie podziału, obliczanie parametrów, itp./.

Etapy te mogą się przeplatać /np. w algorytmie Bonniera czy Hyvärinena/, występować opcjonalnie /najczęściej brak któregoś z etapów: 1, 2, 5, 6/. Niektóre algorytmy mają też specyficzną realizację /komasowanie kilku etapów w jeden - np. Rocchia, Sonquista - Morgana/. Utrudnia to pełną systematyzację algorytmów grupowania.

Okazuje się, że każdy z tych etapów można zrealizować na kilka - kilkanaście sposobów /przedstawionych przez autora w pracy

[138]/, co nasuwa pomysł wykonania parametrycznego generatora programów grupowania. Znając złożoność obliczeniową, wymagany rozmiar pamięci i postać wyników każdego etapu możemy budować algorytmy o z góry określonej złożoności i własnościach.

Ogólną postać tego generatora przedstawiono schematycznie na rys. 2.1.



Rys. 2.1. Schemat blokowy generatora programów grupowania

W pracach [1, 67] częściowo wykonano ten zamiar tzn., podprogramy I, II, V i VI oraz kilkanaście najbardziej popularnych klasycznych algorytmów grupowania w językach Algol 68 [67] lub

Fortran/IBM, niestety, z błędami [1]. Natomiast przy pomocy w/w generatora można zbudować kilkaset istotnie różnych, niestatycznych algorytmów grupowania. Zgromadzone i przygotowane w odpowiedniej formie przez autora w [138] warianty realizacji poszczególnych etapów mogą być bezpośrednio użyte do oprogramowania takiego generatora.

Wyliczmy główne, ogólne cechy algorytmów niehierarchicznych:

- nie nadają się do grupowania atrybutów,
- nie wymagają jednoczesnego pamiętania wszystkich danych lecz jednocześnie wynik zależy od kolejności ich rozpatrywania,
- dzieli najczęściej zbiór B na z góry określoną liczbę grup /być może ustalaną w fazie wstępnej/,
- należy liczyć się z częstymi oscylacjami w pracy algorytmów /nie dotyczy to metod: Wisharta, Steinhausa, Mac Queena, Forgy'ego/,
- są mniej czasochłonne niż metody hierarchiczne,
- główną zasadą ich pracy jest iteracyjna korekcja podziału początkowego.

Przypadek B: Znane są tylko wzajemne podobieństwa między obiektami

Relacje podobieństwa między obiektami w b.d. można zadać w postaci obciążonego, prostego, skończonego grafu niezorientowanego $G = (V(G), E(G), \mu)$ /może być niespójnego/ na obiektach b.d. B jako wierzchołkach. Obciążenia są wartościami funkcji podobieństwa lub "odległości".

Tłumacząc warunek (* /p.2.2) na język teorii grafów można stwierdzić, że element podziału /grupa/ to odpowiednio silnie spójny podgraf grafu G .

Mamy możliwość budowania polihierarchii /w tym hierarchii/ lub tzw. podziałów warstwowych na grafie G . Metody grafowe wykorzystują pojęcie k -spójności / k - komponenta, k - blok, k - spojenie/. Zmiany parametru k modyfikują regułę decyzyjną grupowania od spójności zwykłej / $k = 1$, reguła 1-NN/ do spójności pełnej /reguła 1-FN/. Ponieważ sprawdzono, że klasyczne już algorytmy 1-NN i 1-FN mają wiele cech ujemnych, m.in. pierwszy zbyt szybko łączy obiekty w jedną grupę, a drugi zbyt wolno, ciekawe jest sprawdzenie, jakie własności posiadają algorytmy grupowania przy innych wartościach k i różnych określeniach grupy /komponenta, blok, koło, spojenie/.

Okazuje się, że na gruncie teorii grafów jest możliwe jednolite [78] ujęcie znacznej większości nieprobabilistycznych algorytmów podziału /grupowania/, jak również precyzyjniejsza ocena parametrów składających się na ich złożoność obliczeniową [140]. Ponadto algorytmy grafowe:

- a/ pozwalają wykrywać grupy o nieregularnych kształtach,
- b/ są prostsze obliczeniowo i pojęciowo,
- c/ ich praca nie zależy od przestrzeni cech.

Algorytmy znajdowania q -podziału P_q / q -nieznane/ b.d. B / q - podziału grafu G / przeważnie iteracyjnie obniżają rząd grafu G , stopniowo eliminując z G :

- a/ albo MPP_G lub k -bloki / k - BL_G /,
- b/ albo krawędzie niezgodne /po znalezieniu MDS_G lub NDH_G /,
- c/ albo k -gwiazdy / k - GW_G /,
- d/ albo k -komponenty spójne / k - KS_G / lub k -spojenia / k - SP_G /,
- e/ albo k -koła / k - K_G /.

bądź dokonują t -ekstrakcji grafu G , zaś uzyskane komponenty spójne $KS_i(t)$ są elementami szukanego podziału.

W przypadku grafów istnieją dodatkowe /w porównaniu z przypadkiem A/ możliwości określenia podobieństw między grupami /podgrafami/ $G_1, G_2 \subset G$; $V(G_1) \cap V(G_2) = \emptyset$, $E(G_1) \cap E(G_2) = \emptyset$; np.:

$$p(G_1, G_2) \triangleq \text{extremum}_{v_i \in V(G_1), v_j \in V(G_2)} \{ \mu(e_{ij}) \mid G^* = G_1 \cup G_2 \text{ ma własność } X, e_{ij} \in E(G) \}$$

gdzie X może być następującą własnością:

- a/ k -spójność wierzchołkowa lub krawędziowa grafu G^* ,
- b/ k -sieczność grafu G^* ,
- c/ promień grafu $G^* \leq h$,
- d/ k -pełność grafu G^* , itp.

Charakteryzacji grafowych metod konstrukcji polihierarchii wraz z analizą ich własności dokonał autor w pracy [140]. Zawiera ona m.in. oszacowania liczby poziomów polihierarchii - warstw, liczby grup w warstwie, złożoności obliczeniowej, itp.

W następnym rozdziale dokonamy zbiorczej, krytycznej analizy niehierarchicznych grafowych algorytmów tworzenia q -podziałów słabych /warstw/.

3. ANALIZA PORÓWNAWCZA WYBRANYCH GRAFOWYCH ALGORYTMÓW PODZIAŁU

W niniejszym rozdziale dokonamy krytycznej analizy porównawczej wybranych niehierarchicznych grafowych /lub grafowo-ekstremalnych/ algorytmów tworzenia warstw /q - podziałów słabych/ na zbiorze obiektów - wierzchołków. Część z tych algorytmów będzie wykorzystywana przy dzieleniu eksperymentalnej bazy danych.

Pełniejsza analiza metod grafowych /polihierarchie, złożoność obliczeniowa, porządkowanie metod ze względu na precyzję warstw, itp./ wykracza znacznie poza przyjęte ramy rozprawy. Znaleźć ją można w opracowaniu autora [140].

Grafowe metody grupowania można podzielić na dwie klasy:

- a/ metody wykorzystujące uporządkowanie zbioru krawędzi grafu G według obciążeń μ ,
- b/ metody wykorzystujące bezpośrednio obciążenia łuków.

Do pierwszego typu należą np.: MDS_G - Kruskal, KS_G , k- KS_G , k- BL_G , k- SP_G , MPP_G , a do drugiego typu można zaliczyć: MDS_G -Zahn, MDS_G [57], MPP_G -Gotlieb, KS_G -Gotlieb, NDH_G .

W przypadku a/ rolą funkcji $\mu: E(G) \rightarrow R$ jest wyłącznie uporządkowanie zbioru

$$E(G) = \{e_i\}_{i=1}^W \text{ w sekwencję: } \{e_1 \prec e_2 \prec \dots \prec e_W\} = \hat{E}(G), e_i = \{v_k, v_l\} \in E(G).$$

$$e_k \prec e_l \iff \begin{cases} \mu_d(e_k) \leq \mu_d(e_l), e_k = \{v_i, v_j\}, e_l = \{v_o, v_r\} \\ \text{lub } \mu_p(e_k) \geq \mu_p(e_l) \end{cases}$$

Wówczas graf $\hat{G} = (V(G), \hat{E}(G))$ z relacją porządkującą nazwiemy grafem $\prec$ - bliskości.

Dodając kolejno do grafu pustego $G = (V(G), \emptyset)$ krawędzie $e \in \hat{E}(G)$ otrzymujemy wstępujący ciąg grafów częściowych $\{\hat{G}^i | G_1 \subset G_2 \subset \dots \subset G_W\}$, w których szukamy zdefiniowanych struktur /spojen, KS, MPP, itd./. Uzyskujemy w wyniku ciąg grupowań warstwowych, które tworzą hierarchię na $V(G)$. Hierarchia ta jest hierarchią ścisłą dla k- KS , k- SP , natomiast dla metod k- BL , k- MPP jest to hierarchia słaba /nierozłączność grup w warstwie/.

3.1. Definicje teorii grafów

Niech $G = (V(G), E(G), \mu)$ - graf obciążony, skończony, nieskierowany, bez pętli /prosty/ gdzie:

$V(G) = \{v_1, \dots, v_n\} \triangleq \{b_1, b_2, \dots, b_n\} = B$ - zbiór wierzchołków /obiektów/,

$E(G) = \{e_{ij} = \{v_i, v_j\}, i \neq j\}$ - zbiór krawędzi /par nieuporządkowanych/,

$\text{card } E(G) = w \leq n(n-1)/2$,

$\mu: E(G) \rightarrow \mathbb{R}$, funkcja obciążeń - relacja porządkująca zbiór $E(G)$.

DF 3.1.1. $\forall (i \neq j) (e_{ij} \in E(G) \subset V^2) \Leftrightarrow$ /wierzchołki v_i, v_j łączy krawędź $e_{ij} = \{v_i, v_j\}$ /

Uwaga: $e_{ii} \notin E(G)$ oraz $e_{ij} \equiv e_{ji}$

DF 3.1.2. $\mu_p(e_{ij}) \triangleq p(b_i, b_j)$, $\mu_d(e_{ij}) \triangleq d(b_i, b_j)$ /p - podobieństwo, d - odległość/.

Wówczas graf G nazwiemy grafem podobieństwa /odległości/.

DF 3.1.3. Podgrafem G' grafu $G = (V(G), E(G), \mu)$ nazywamy graf $G' = (V(G'), E(G'), \mu)$, w którym $V(G') \subset V(G)$, $E(G') \subset E(G)$ oraz $\forall (i, j) ((e_{ij} \in E(G')) \Rightarrow (\{v_i, v_j\} \subset V(G')))$.

DF 3.1.4. Grafem częściowym G' grafu G nazywamy każdy podgraf $G' \subset G$, w którym $V(G') = V(G)$.

DF 3.1.5. Podgrafem indukowanym G' /przez zbiór wierzchołków $V(G')$ / grafu G nazywamy każdy podgraf G' , w którym:

1/ $V(G') \subset V(G)$, $E(G') \subset E(G)$

2/ $\forall (\{v_i, v_j\} \subset V(G')) ((e_{ij} \in E(G)) \Rightarrow (e_{ij} \in E(G')))$

DF 3.1.6. t - ekstrakcją $G(t) = (V(G), E_t(G), \mu)$ grafu G nazywamy graf częściowy $G(t)$ grafu G o własności $(e_{ij} \in E_t(G)) \Leftrightarrow (\mu_d(e_{ij}) \leq t, t \in \mathbb{R})$. Oczywiście $G(0) = (V(G), \emptyset)$ jest grafem pustym / $w = 0$ /.

DF 3.1.7. Jeśli $\{e_{ij}, e_{jk}\} \subset E(G)$ to:

- a/ wierzchołki v_i, v_j oraz v_j, v_k są sasiednie,
- b/ krawędzie e_{ij}, e_{jk} są sasiednie,
- c/ wierzchołki v_i, v_j są incydentne /przyległe/ krawędzi e_{ij} .

DF 3.1.8. Stopień /deg v_i / wierzchołka v_i - liczba krawędzi incydentnych z wierzchołkiem v_i .

DF 3.1.9. Droga $D(v_i, v_j)$ między wierzchołkami v_i, v_j w grafie G nazywamy ciąg /spójny/ krawędzi wzajemnie sąsiednich:

$$D(v_i, v_j) = \langle \{v_i, v_{i_1}\}, \{v_{i_1}, v_{i_2}\}, \dots, \{v_{i_{k-1}}, v_{i_k}\}, \{v_{i_k}, v_j\} \rangle.$$

Mówimy wówczas, że wierzchołek v_j jest osiągalny z wierzchołkiem v_i po drodze $D(v_i, v_j)$. Drogę $D(v_i, v_i)$ nazywamy cyklem.

Droga jest elementarna, gdy jej elementy /krawędzie/ są różne.

Droga jest prosta, gdy $\text{card}(\bigcup (\{v_k, v_l\} \in D(v_i, v_j))) = \text{card } D(v_i, v_j) + 1$ /wierzchołki przez które przechodzi są różne/.

Drogę elementarną przechodzącą przez wszystkie krawędzie nazywamy drogą Eulera.

Drogę prostą przechodzącą przez wszystkie wierzchołki nazywamy drogą Hamiltona.

DF 3.1.10. Długością $\tilde{D}$ drogi $D(v_i, v_j)$ nazywamy liczbę jej krawędzi

$$\tilde{D}(v_i, v_j) \triangleq \text{card } D(v_i, v_j)$$

DF 3.1.11. Obciążeniem D^Σ drogi $D(v_i, v_j)$ nazywamy sumę obciążeń jej krawędzi

$$D^{\Sigma}(v_i, v_j) \triangleq \sum_{e_{ij} \in D(v_i, v_j)} w(e_{ij})$$

DF 3.1.12. Najkrótszą drogą Hamiltona /NDH_G/ grafu G nazywamy drogę Hamiltona o najmniejszym obciążeniu D^{Σ} .

DF 3.1.13. Graf G jest spójny, jeśli każdy jego wierzchołek jest osiągalny/DF 3.1.9./z dowolnego innego jego wierzchołka. Relacja spójności jest zwrotna, symetryczna, przechodnia, więc jest relacją typu równoważności i wyznacza klasy abstrakcji - komponenty spójności.

DF 3.1.14. Komponenta spójna /KS_G/ grafu G nazywamy każdy maksymalny /w sensie zawierania/ podgraf spójny indukowany $G' = KS_G \subset G$. Każdy graf G prosty można przedstawić w postaci sumy mnogościowej komponent spójnych.

DF 3.1.15. Rząd grafu - liczba wierzchołków komponenty spójnej grafu G o największej liczbie wierzchołków.

DF 3.1.16. Promień grafu spójnego G:

$$r(G) \triangleq \min_{v_i \in V(G)} \max_{v_j \in V(G)} \tilde{D}(v_i, v_j)$$

DF 3.1.17. Wierzchołek $v^* \in V(G)$ nazywamy wierzchołkiem centralnym grafu spójnego G $\iff \max_{v_i \in V(G)} \tilde{D}(v_i, v^*) = r(G)$.

/v* może nie być wyznaczone jednoznacznie; dla grafów prostych nie jest znane nietrywialne oszacowanie ich liczby, dla drzew Jordan pokazał, że istnieją: jeden lub dwa sąsiednie wierzchołki centralne/.

DF 3.1.18. Średnica grafu spójnego G:

$$\text{diam } G \triangleq \max_{v_i, v_j \in V(G)} \tilde{D}(v_i, v_j)$$

Dla grafu pełnego G_n , $\text{diam } G_n = 1$.

- DF 3.1.19. 1-kołem grafu $/1-K_G/$ nazywamy każdy maksymalny spójny podgraf indukowany G' , dla którego: $\text{diam } G' \leq 1 \in \mathbb{N}$.
- DF 3.1.20. Indukowany podgraf G' grafu G jest k-spójny $/k \in \mathbb{N}/$ wierzchołkowo jeśli:
- 1/ $\text{card } V(G') \geq k+1$,
 - 2/ usunięcie nie więcej niż $k-1$ dowolnych wierzchołków /wraz z krawędziami incydującymi/ nie rozspaja G' , tzn.
 $\forall (G'' \subset G', \text{card } V(G'') \geq (\text{card } V(G') - k + 1) (G'' \text{ jest spójny})$
- DF 3.1.21. k-blokiem $/k-BL_G/$ grafu G nazywamy każdy maksymalny k-spójny indukowany podgraf G' grafu G .
- DF 3.1.22. k-spojeniem $/k-SP_G/$ grafu G nazywamy każdy maksymalny spójny podgraf indukowany G' , w którym:
- $$\forall (v_i \in V(G')) (\deg v_i \geq k, k \in \mathbb{N})$$
- DF 3.1.23. Grafem prostym pełnym nazywamy graf, w którym:
- $$\forall (v_i, v_j \in V(G)) (e_{ij} \in E(G)).$$
- Oczywiście:
 $w = \text{card } E(G) = n(n-1) / 2$. Oznaczmy go symbolem G_n
 $/G_{n_1, n_2, \dots, n_k}$ - dla grafów k-dzielnych - DF 3.1.35/.
- DF 3.1.24. Maksymalnym podgrafem pełnym /kliką/ $/MPP_G/$ nazywamy maksymalny /w sensie zawierania/ indukowany podgraf $G' \subset G$, który jest pełny.
- k-klika $/K-KL_G/$ nazywamy graf pełny rzędu co najmniej $k+1$.
- DF 3.1.25. k-gwiazdą $/k-GW_G/$ nazywamy spójny podgraf G' bez cykli o własnościach:
- 1/ $\text{card } V(G') = k+1$, $k = 2, 3, \dots$, $\text{card } E(G') = k$,
 - 2/ $r(G') = 1$ /DF 3.1.16/.
- DF 3.1.26. q-podziałem $/P_q/$ grafu G nazywamy każdy q-elementowy zbiór rozłącznych podgrafów spójnych $\{G_i, i = 1, 2, \dots, q\}$ pokrywających wierzchołkowo graf G , tzn.:
- 1/ $\forall (i = 1, 2, \dots, q) (V(G_i) \neq \emptyset)$

$$2/ \forall (i \neq j) (V(G_i) \cap V(G_j) = \emptyset)$$

$$3/ \bigcup_{i=1}^q V(G_i) = V(G) = B$$

Zastępując warunek 2/ warunkiem:

$$2a/ \forall (i \neq j) (V(G_i) \subset V(G_j)) \Rightarrow (V(G_i) \equiv V(G_j)) \text{ otrzymujemy } \underline{q\text{-podział słaby } P_q}$$

DF 3.1.27. Przekrojem /cieciem/ $C(G_1, G_2)$ grafu G nazywamy zbiór krawędzi spajających pewien dwu-podział $P_2 = \{G_1, G_2\}$ grafu G :

$$C(G_1, G_2) \triangleq \{e_{ij} \in E(G) | (v_i \in V(G_1)) \wedge (v_j \in V(G_2))\},$$

$$\text{ponadto: } G_1 \cup G_2 \cup C(G_1, G_2) \equiv G$$

DF 3.1.28. Grafem k-siecznym /k-spójnym krawędziowo/ nazywamy graf G w którym: $\text{card } C(G_1, G_2) \geq k$ dla każdego dwupodziału $P_2(G_1, G_2)$.

DF 3.1.29. k-komponenta spójna /k-KS_G/ nazywamy każdy maksymalny /w sensie zawierania/ indukowany podgraf G' , który jest k-sieczny.

DF 3.1.30. Szerokością $\nabla(G_1, G_2)$ przekroju $C(G_1, G_2)$ grafu G nazywamy liczbę:

$$\nabla(G_1, G_2) = \min_{e_{ij} \in C(G_1, G_2)} \mu_d(e_{ij})$$

DF 3.1.31. Wieża $\lambda(G_1, G_2)$ przekroju $C(G_1, G_2)$ grafu G jest zbiór $\lambda(G_1, G_2) \triangleq \{e_{ij} \in C(G_1, G_2) | \mu_d(e_{ij}) = \nabla(G_1, G_2)\}$ - krawędzie przekroju o obciążeniu równym szerokości przekroju.

DF 3.1.32. Drzewem grafu G nazywamy spójny graf częściowy bez cykli: $T_G = (V(G), E'(G), \mu) : \text{card } E'(G) \triangleq E(T_G) = \text{card } V(G) - 1 = \triangleq \text{card } V(T_G) - 1$.

Przez poddrzewo $T' \subset T_G$ rozumiemy podgraf T' , który jest drzewem.

DF 3.1.33. Minimalnym drzewem spinającym / MDS_G / grafu G nazywamy drzewo T_G o minimalnej sumie obciążeń krawędzi

$$MDS_G \triangleq \arg \min_{T_G} \sum_{e_{ij} \in E(T_G)} \mu_d(e_{ij})$$

DF 3.1.34. Krawędzia niezgodna nazywamy krawędź e^* , której obciążenie $\mu(e^*)$ istotnie /względem przyjętego kryterium/ różni się od obciążeń innych krawędzi $e \in E(G)$, np.

[149] :

$$\mu_d(e^*) \gg \sum_{e \in E(G)} \mu_d(e) / 2 \text{ card } E(G)$$

DF 3.1.35. Grafem k-dzielnym nazywamy graf $G = (V, E)$, w którym istnieje k-podział zbioru wierzchołków $V(G)$. Ponadto:
 $(e_{ij} = \{v_i, v_j\} \in E(G)) \Rightarrow (v_i, v_j \text{ należą do różnych elementów k-podziału})$.

3.2. Algorytmy podziału wykorzystujące MDS i NDH

Następujące fakty mają podstawowe znaczenie przy konstrukcji MDS lub NDH w grafie $G = (V(G), E(G), \mu)$ - prostym, skończonym, spójnym i obciążonym.

F 3.2.1. Jeśli $\mu: E(G) \rightarrow \mathbb{R}_+$ jest funkcją różnowartościową /obciążenia krawędzi są unikalne/ to MDS_G jest określone jednoznacznie.

F 3.2.2. Wszystkie krawędzie MDS_G / NDH_G / spajają pewne podziały grafu G .

Uwaga:

- a/ można podać przykład, gdy MDS_G jest jedyne, a obciążenia w G nie są różne,
- b/ dla NDH_G fakt F 3.2.1. nie jest prawdziwy w żadną stronę.

WN 3.2.1.

- a/ gdy μ -różnowartościowa, wynik algorytmu konstruującego MDS_G nie zależy od wierzchołka startowego i kolejności rozpatrywania krawędzi,

- b/ usuwając k krawędzi $/k < n-1/$ otrzymujemy $k+1$ poddrzew
 $MDS_G^i \subset MDS_G$ /poddróg $NDH_G^i \subset NDH_G/$ - być może zdegenero-
wanych do wierzchołków grafu G ,
c/ dla $k = n-1$, $MDS_G^i = NDH_G^i = \{v_i \in V(G)\}$, $i = 1, 2, \dots, n$.
d/ jeśli oznaczymy przez $MDS_G^i(k)/NDH_G^i(k)/$, $i = 1, 2, \dots, k+1$
podgrafy grafu G indukowane przez $V(MDS_G^i)/V(NDH_G^i)/$ po
usunięciu z $MDS_G/NDH_G/$ k krawędzi to:

$$\forall (MDS_G^i(k)/NDH_G^i(k)/, i = 1, 2, \dots, k+1, k \leq n-1) \exists (1 \leq k)$$

$$(MDS_G^i(k) \subset MDS_G^j(1), NDH_G^i(k) \subset NDH_G^j(1)),$$

Wnioski b/ - d/ umożliwiają na mocy F 3.2.2.

łatwe konstruowanie hierarchii na $V(G)$.

TW 3.2.1. Dowolna krawędź $e^* = \{v_i, v_j\} \in E(MDS_G)$ należy do pewnego
przekroju $C(G_i, G_j)$, $v_i \in V(G_i)$, $v_j \in V(G_j)$, grafu spój-
nego G , a w szczególności do więzi przekroju $\lambda(G_i, G_j)$.
Każda więź $\lambda(G_i, G_j)$ zawiera dokładnie jedną krawędź
 $e^* \in E(MDS_G)$, jeśli G_i, G_j - spójne

Dowód:

Z DF 3.1.32 wynika, że usunięcie $e^* = \{v_i, v_j\} \in$
 $E(MDS_G) \subset E(G)$ rozspaja drzewo na poddrzewa MDS_i ,
 MDS_j takie, że $v_i \in V(MDS_i)$, $v_j \in V(MDS_j)$. Niech G_i, G_j -
- rozłączne podgrafy indukowane przez $V(MDS_i)$, $V(MDS_j)$.
Z DF 3.1.27, $C(G_1, G_2) = G \setminus (G_1 \cup G_2)$, czyli

$e^* = \{v_i, v_j\} \in C(G_i, G_j)$. Załóżmy, że $e^* \notin \lambda(G_i, G_j)$, tzn.

$\exists (e \in \lambda(G_i, G_j)) (\mu(e) < \mu(e^*))$. A więc istnieje drzewo

$T = (V(G), E(MDS_G) \cup \{e\} \setminus \{e^*\})$ o obciążeniu mniejszym niż
obciążenie drzewa MDS_G , co przeczy minimalności MDS_G .
Jeśli zbiór krawędzi o jednakowych, minimalnych obciąże-
niach $\lambda(G_i, G_j)$ zawiera więcej niż jedną krawędź z
 $E(MDS_G)$ to w MDS_G występuje cykl, co przeczy temu, że
 MDS_G jest drzewem. Jeśli grafy G_i, G_j są niespójne, to
więź $\lambda(G_i, G_j)$ zawiera co najmniej dwie krawędzie e_i^* ,
 $e_j^* \in E(MDS_G)$ ■

WN 3.2.2. Każda krawędź $e^* \in E(MDS_G)$ jest krawędzią o minimalnym
obciążeniu spajającą pewien podział /DF 3.1.26/ $\{G_i, G_j\}$

grafu G .

O ile zagadnienie znajdowania MDS_G i aktualizacji [19, 123] można zawsze rozwiązać w czasie co najwyżej $O(n^3)$, o tyle problem charakteryzacji istnienia NDH_G nie został zadawalająco rozwiązany. Brak warunków koniecznych i dostatecznych, aby rozstrzygnąć w ogólnym przypadku, czy graf G jest grafem hamiltonowskim. Przytoczyć można jedynie dwa klasyczne twierdzenia Öre i Diraca:

F 3.2.3. Jeśli dla każdych dwu różnych i niesąsiednich wierzchołków v_1, v_2 grafu spójnego G zachodzi nierówność:

$$\deg v_1 + \deg v_2 \geq n-1, \quad n = \text{card } V(G), \quad v_1, v_2 \in V(G)$$

to G zawiera drogę Hamiltona.

F. 3.2.4. Jeśli w grafie spójnym G dla $\forall (v \in V(G)) (\deg v \geq \frac{1}{2} n)$, to G zawiera cykl Hamiltona.

WN 3.2.3. W grafie pełnym G_n zawsze istnieje droga Hamiltona.

Oczywiście, zawsze można rozstrzygnąć istnienie drogi Hamiltona w grafie prostym skończonym metodą przeglądu pełnego. Jej złożoność wynosi jednak:

$$n! \approx n^n e^{-n} \sqrt{2\pi n} \quad \text{/liczba permutacji zbioru wierzchołków/}.$$

Natomiast w grafach pełnych G_n pracochłonnym zadaniem jest znalezienie NDH_G wśród $\frac{1}{2}(n-1)!$ istniejących dróg Hamiltona.

Propozycje Spärck-Jones [120], Sobczaka i Maliny [117], czy też autorów prac [11, 113] wykorzystywania NDH_G są /dla dużych n / niepraktyczne; ponieważ, jak pokazano w [42] już dla grafów płaskich oraz kubicznych $\forall i, \deg v_i = 3$ /złożoność zagadnienia Hamiltona nie jest wielomianowo ograniczona z góry.

Próbami ominięcia tej trudności było opracowanie algorytmów znajdowania prawie- NDH_G . Wymienić można tu metody [7, 11, 34, 81] ponadto metody heurystyczne Karga-Raymonda-Helda, Rubina [104], Frencla [38]. Należy jednak dodać, że nie znalazły one szerszego zastosowania w systemach wyszukiwania informacji.

Z F 3.2.1 ÷ 2 i TW 3.2.1 wynika możliwość dokonywania podziału $P_q = \{P_1, P_2, \dots, P_q\}$ bazy danych B , w której relacje podobień-

stwa lub "odległości" obiektów przedstawione w postaci grafu $G = (V(G) \equiv B, E(G), \mu_d)$ na podstawie analizy MDS_G .

Algorytm podstawowy - AP1:

1. utwórz $MDS_G = (V(G), E(MDS_G), \mu_d)$. $E(MDS_G) = \{e_1, e_2, \dots, e_{n-1}\}$.
2. określ kryterium F niezgodności krawędzi /DF 3.1.34/.
3. uporządkuj krawędzie $e \in E(MDS_G)$ nierosnąco według wartości $F(e)$ w ciąg $\{e_j\}_{j=1}^{n-1}$.
4. liczba grup podziału q znana? Jeśli tak, idź do 7.
5. określ warunek stopu S procedury eliminacji krawędzi /PROC EL/.
6. wykonuj PROC EL dopóki nie zajdzie S . Idź do 8.
7. usuń z ciągu $\{e_j\}_{j=1}^{n-1}$ $q-1$ pierwszych krawędzi.
8. wierzchołki uzyskanych podgrafów /podrzew MDS_G / stanowią szukane grupy.
9. stop.

Omówimy metody realizacji poszczególnych kroków algorytmu AP1.

Krok 1.

Najpopularniejszymi metodami znajdowania MDS_G są metody: Prima /1957/, Dijkstry /1959/ o złożoności $O(n^2)$ i Kruskala-Dijkstry /1956/ o złożoności $O(w \log w)$. Znanych jest też wiele innych algorytmów o podobnej złożoności /np. [89, 103, 112, 134]/ oraz $O(w \log \log n)$ [8, 102, 146, 148] i $O(n+w)$ [56] algorytmy dla grafów G_n o wierzchołkach umieszczonych w R_+^r /dla $r = 2$ i metryki prostokątnej - $O(n \log n)$ /. Zupełnie niedawno [18], dzięki zastosowaniu metody tzw. poszukiwania w głąb /deep - first search/, uzyskano w ogólnym przypadku złożoności $O(w \log \log n)$, $O(n+w)$ i $O(n \log^* n + w)^*$, ponadto $O(w)$ dla grafów gęstych i $O(n)$ dla planarnych.

Krok 2.

Proponowano wiele kryteriów orzekających o niezgodności krawędzi e^* /DF 3.1.34./:

- a/ $e^* \triangleq \max \{ \mu_d(e) \mid e \in E(MDS_G) \}$,
- b/ $\{e^*\} \triangleq \{e \in E(MDS_G) \mid \mu_d(e^*) \gg \sum_{e \in E(MDS_G)} \mu_d(e) / \text{card } E(MDS_G)\}$,

) $\log^ n \triangleq \min \{ i \mid \underbrace{\log \log \dots \log n}_{i\text{-razy}} \leq 1 \}$

- c/ [149] $\{e^*\} \triangleq \{e' \in E(G) \mid \mu_d(e') > a \cdot \text{mean}\{\mu_d(e) \mid \{e\} \text{ - zbiór krawędzi sąsiednich krawędzi } e'\}, a - \text{współczynnik, mean } X - \text{średnia w zbiorze } X$
- d/ [149] $\{e^*\} \triangleq \{e' \in E(G) \mid \mu_d(e') > a \cdot \text{mean}\{(\mu_d(e) - \text{mean}\{\mu_d(e)\})^2 \mid \{e\} - \text{zbiór krawędzi p-sąsiednich krawędzi } e'\},$
 - a - współczynnik .

Propozycje c/ i d/ są wynikiem analizy statystycznej materiału empirycznego dotyczącego postrzegania przez człowieka figur płaskich. Nie znalazły one, jak dotąd, szerszego zastosowania w metodach klasyfikacji. Najczęściej stosuje się najprostszą obliczeniową propozycję a/. Wymaga ona $O(k \log k)$ operacji $/k = \text{card } E(\text{MDS}_G) /$, gdy sortujemy krawędzie drzewa według obciążeń. Nie jest to jednak konieczne. Można wykorzystać algorytmy selekcji:

- [96] /Reiser/ - wybór i -tej, ze względu na obciążenie, krawędzi w ciągu par $\{(e_j, \mu_d(e_j))\}_{j=1}^k$ o złożoności $O(k)$; $1 \leq i \leq k$,
- [35] /Floyd, Rivest/ - to samo, o złożoności: $k + \min\{i, k-1\} + O(k)$,
- [41] /Fussengger, Gabow/ - to samo, o złożoności nie niższej niż $k-i + \lceil \log \binom{k}{i-1} \rceil$ oraz określenie j największych elementów w ciągu k -elementowym /przydatne, gdy liczba grup q w podziale $\mathbb{P}_q$ jest znana/, o złożoności nie niższej niż:
 $k-j + \lceil \log \left(\frac{1}{j} \binom{k}{j-1} \right) \rceil$, $1 \leq j \leq k$.

Krok 3.

W tym kroku sortujemy krawędzie $\{e_j\}_{j=1}^k$ według ich obciążeń. Można tego dokonać na wiele sposobów /przegląd metod w [59] / o złożoności nie wyższej niż $O(k \log k)$ - Quicksort, a nawet średnio $O(k)$ - DPS [29].

Krok 5.

Jeśli nie jest znana liczba grup q w podziale $\mathbb{P}_q = \{P_1, P_2, \dots, \dots, P_q\}$ grafu $G = (V(G) \equiv B, E(G), \mu)$, należy opracować procedurę /najczęściej optymalizacyjną/ usuwania krawędzi spajających elementy P_i podziału $\mathbb{P}_q$.

Propozycje:

- a/ W posortowanej sekwencji $\{e_j\}_{j=1}^k$ krawędzi MDS_G znajdź krawędź e_r taką, że: $e_r = \arg \max_{1 \leq j \leq k-1} |\mu_d(e_j) - \mu_d(e_{j+1})|$. Podstaw

$q := r+1$, idź do kroku 7,

b/ Niech $G_i = (V(G_i) \equiv P_i, E(G_i), \mu_d)$, $i = 1, 2, \dots, q$; będą podgrafami grafu G indukowanymi przez zbiory wierzchołków P_i .

Oznaczmy $n_i = \text{card } P_i$, $n = \sum_{i=1}^q n_i$

$$q^* = \arg \min_{1 \leq q \leq n} \left(\frac{\alpha}{n} \sum_{i=1}^q \frac{1}{n_i} \sum_{e \in E(G_i)} \mu_d(e) + \frac{\beta \cdot n^2}{\sum_{i=1}^q n_i^2} \right)$$

gdzie: $\alpha, \beta \in \mathbb{R}^+$ - współczynniki,

c/ jak w b/ lecz $G_i = (V(\text{MDS}_G^i) \equiv P_i, E(\text{MDS}_G^i), \mu_d)$, co znacznie zmniejsza pracochłonność, ponieważ uwzględniamy tylko $n-1$ krawędzi z $\frac{n(n-1)}{2}$ możliwych,

d/ wykonać AP1 dla kilku, arbitralnie dobranych wartości q i po sprecyzowaniu kryterium jakości podziału K wybrać q dające najlepszą wartość K ,

e/ [57] niech: $\text{MDS}_G = (V(\text{MDS}_G), E(\text{MDS}_G), \mu_d)$, $P_q = \{P_1, \dots, P_q\}$ - q -podział wierzchołków $V(\text{MDS}_G)$, tzn. $P_i = (V(P_i), E(P_i), \mu_d)$, $i = 1, 2, \dots, q$; gdzie $V(P_i) \subset V(\text{MDS}_G)$, $E(P_i) \subset E(\text{MDS}_G)$ - poddrzewo MDS_G /być może zdegenerowane/.

Zdefiniujemy:

$E_\lambda = \{e_{\lambda_1}, \dots, e_{\lambda_{(q-1)}}\} \triangleq E(\text{MDS}_G) \setminus \bigcup_{i=1}^q E(P_i)$ - zbiór krawędzi usuwanych z MDS_G .

Ponadto:

$$u_1 \triangleq \frac{1}{q} \sum_{i=1}^q \frac{1}{n_i-1} \mu_d(e \mid e \in E(P_i)) \quad , \quad n_i = \text{card } V(P_i),$$

$$u_2 \triangleq \frac{1}{q-1} \sum_{i=1}^{q-1} \mu_d(e \mid e \in E_\lambda) \quad - \text{średnia "odległość" podgrafów od siebie,}$$

$$u_3 \triangleq \frac{1}{q-1} \sum_{i=1}^{q-1} \frac{w(e_{\lambda_i})}{\mu_d(e_{\lambda_i})} \quad - \text{średnia "ostrość" podziału,}$$

gdzie:

$$w(e_{\lambda_i}) \triangleq \max_e \{ \mu_d(e) \mid e \text{ - sąsiednie z } e_{\lambda_i} \text{ w } \text{MDS}_G \}$$

$$u_4 \triangleq q^q \prod_{i=1}^q \frac{n_i}{n} \quad - \text{"równomierność" podziału } \text{MDS}_G,$$

Szukamy takiego optymalnego zbioru E_{λ}^* , że:

$$E_{\lambda}^* = \arg \max_{E_{\lambda}} \max_{1 < q < k} \ln \frac{(u_2(E_{\lambda}))^a (u_4(E_{\lambda}))^b}{(1+u_1(E_{\lambda}))^c (1+u_3(E_{\lambda}))^d}, \quad a, b, c, d - \text{parametry}$$

Dla $a = b = c = d = 1$ stwierdzono dobrą zgodność wyników z eksperymentami. Oczywiście we wzorach na u_1, u_3 można sumować po krawędziach $e \in E(G)^*$, a nie $E(MDS_G)$.

Uwaga:

Autorzy tych procedur optymalizacyjnych zakładają milcząco, że graf G jest spójny i MDS_G jest jedyne. Jeśli tak nie jest obliczenia należy powtarzać dla każdego MDS_G , co znacznie zwiększa i tak już dużą złożoność obliczeniową procedur. Okazuje się, że /z wyjątkiem prostej propozycji a/ i ewentualnie c/ / procedury, które wykorzystują ścisłe metody optymalizacyjne już dla $n > 10^3$ pracują niedopuszczalnie długo.

Próbuje się ominąć trudności obliczeniowe przez generowanie quasi-MDS lub wykorzystanie do podziału grafu G algorytmów szukania tzw. mostów w grafie /np. Corneila - $O(n^2)$ /.

3.3. Algorytmy podziału wykorzystujące MPP i KS

Bardzo efektywne obliczeniowo są algorytmy wykorzystujące pojęcie t-ekstrakcji $G(t) = (V(G), E_t(G), \mu) / DF 3.1.6/$ do znajdowania komponent spójnych KS /DF 3.1.14/. Wynik tej metody nie zależy od kolejności rozpatrywania krawędzi grafu. Jest ona szczególnie efektywna, gdy krawędzie $e \in E(G)$, są posortowane rosnąco według ich obciążeń $\mu_d(e)$. Można wówczas wykorzystać algorytm Kruskala [66], lecz nie jest konieczna eliminacja cykli, co obniża złożoność do ok. $O(w)$.

Dla ekstrakcji $G(t)$ zachodzą następujące, łatwe do wykazania własności:

F. 3.3.1. a/ $\forall (t \in \mathbb{R}^+) (E_t(G) \subset E(G))$, a stąd $G(t) \subset G$,

b/ $G(0) = V(G) = B$, $G(t = \max_e \mu_d(e)) = G$,

c/ $(t_1 > t_2) \Rightarrow G(t_2) \subset G(t_1)$.

Graf $G(t)$ może być oczywiście niespójny. Liczba komponent $KS_i(t)$, które wyznaczają pewien podział $P(t) = \{KS_i(t)\}$, nie roś-

nie ze wzrostem t . Ponadto z własności c / wynika, że jeśli:

$$P(t_1) = \{KS_i(t_1) | i = 1, 2, \dots, a\}, P(t_2) = \{KS_j(t_2) | j = 1, 2, \dots, b\}, t_1 < t_2,$$

$$\text{to } \forall (i = 1, 2, \dots, a) \exists (j = 1, 2, \dots, b) (KS_i(t_1) \subset KS_j(t_2)).$$

Jest to cecha tzw. podziału mocnego.

Głównym zadaniem jest odpowiedni dobór liczby $t \in R^+$. W [4] proponuje się następujące wskazówki dotyczące wyboru t^* :

- 1/ liczba elementów podziału nie może być zbyt duża,
- 2/ podział P powinien być stabilny w pewnym otoczeniu t^* ,
- 3/ liczbę grup ustala się szukając największego skoku wartości gęstości $\Theta(t)$ grafu $G(t)$:

$$\Theta(t) = \frac{2 \text{ card } E_t(G) - n}{n^2}$$

Potwierdzony empirycznie, na rzeczywistym materiale /m.in. Spärck-Jones, Augustson, Patrick, Ogilvie [4, 84, 87]/, fakt dobrych własności podziałów dokonywanych na podstawie analizy MPP_G jest istotną zaletą tych metod /nawet gdy testy na danych generowanych [21] niezupełnie to potwierdzają/. Jednakże teoretycznie wykładnicza złożoność obliczeniowa jest ich poważną wadą. /Choć testy [16] pokazują, że uśredniona złożoność jest prawie liniowa/. Kompromisowe cechy mają algorytmy [4, 14, 45] oryginalnie wykorzystujące MPP_G , a będące wariantami następującego algorytmu szkieletowego:

AS

1. $i := 1, G_i := G(t), A = \{a_k\} := \emptyset$
2. $a_i := V(MPP_{G_i})$
3. $\text{card } V(MPP_{G_i}) \geq 8$, jeśli tak, idź do 6
4. $j = 1$?, jeśli nie, idź do 11
5. idź do 8
6. $j := 1, A := A \cup \{a_i\}, G_{i+1} := G_i \setminus MPP_{G_i}$ /wraz z krawędziami spajającymi MPP_{G_i} z G_i /,
7. $i := i+1$, idź do 2
8. $\forall (v \in V(G_i))$ (według K_1 ($a_k := a_k \cup \{v\}, V(G_i) := V(G_i) \setminus \{v\}$)).

9. czy wymagana korekcja podziału $P_i = \{a_1, a_2, \dots, a_i\} \cup \{v\}$,
 $v \in V(G_i)$?; jeśli nie idź do 11.
 10. połącz podobne /według K_2 / elementy podziału P_i
 11. stop /jeśli $j \neq 1$ należy zmodyfikować t lub ϵ i powtórzyć algorytm/.
- gdzie: $G(t) = (V(G), E_t(G))$ - t -ekstrakcja grafu $G = (V(G), E(G), \mu)$
 ϵ - zadana liczba naturalna określająca minimalny rząd
kliki w G , który wystarcza do jej eliminacji z $G(t)$,
 t - zadany próg obcięcia,
 K_1 - kryterium podobieństwa /odległości/: wierzchołek -
- podgraf
 K_2 - kryterium podobieństwa /odległości/: podgraf-podgraf.

Uwagi do AS:

- AS odpowiada tzw. I algorytmowi Bonniera,
- wyniki zależą od punktu startu /niejednoznaczność kroku 2/ oraz
kolejności rozpatrywania krawędzi,
- w kroku 6, usuwając MPP_G z aktualnego grafu tracimy część informa-
cji /kliki mogą się przekrywać/,
- ϵ , t są parametrami strojonymi,
- czas pracy algorytmu nie jest nadmiernie długi, algorytm jest
zbieżny /występuje tylko łączenie elementów podziału/, wynik jest
podziałem mocnym,
- w [14] przyjęto:

$$K_1 : a_k \triangleq \arg \min_{a_i \in A} \frac{1}{\text{card } a_i} \sum_{u \in a_i} \mu_d(v, u)$$

$$K_2 \rightarrow \min: d(a_i, a_j) \triangleq \frac{1}{\text{card } a_i + \text{card } a_j} \sum_{u \in a_i} \sum_{v \in a_j} \mu_d(v, u)$$

- metoda Dattoli-Murray'a generuje podziały słabe, bo:

$$K_1 : \deg v > \epsilon, \text{ w podgrafie indukowanym przez } V' = \{v\} \cup a_k \subset V(G)$$

$$K_2 : \text{card } C(a_i, a_j) > \eta, C - \text{przekrój /DF 3.1.27/ podgrafu indukowanego przez } V = a_i \cup a_j$$

ϵ, η - parametry strojone.

Ponadto, ograniczając z góry licznosc każdej grupy /możliwość rozszczepień/, dopuszcza się wystąpienie oscylacji.

- zastosowanie w AS /krok 2/ procedury Bierstone'a zamiast Bonniera znacznie skraca czas obliczeń [4] kosztem znacznego zwiększenia

niezbędnej pamięci operacyjnej,

- algorytm Gotlieba-Kumara /AGK/ [45] ma przebieg podobny do AS.

Wypunktujemy różnice:

- 1/ w kroku 6 nie usuwamy MPP_{G_i} z grafu $G(t)$ - algorytm generuje wszystkie różne MPP_G , o rzędzie większym niż δ^2 ,
- 2/ kroków 8 i 9 brak,
- 3/ w kroku 10 sukcesywnie łączy się MPP podobne /wg współczynnika Tanimoto przekrycia zbiorów wierzchołków $V(MPP_G)$ /.

Własności AGK:

- a/ konstruuje polihierarchię / wykorzystywany był z dobrym skutkiem do grupowania słów kluczowych/,
- b/ pewne obiekty mogą, jeśli nie zastosuje się dodatkowej procedury korekcyjnej, pozostać niezakwalifikowanymi - podział wynikowy nie pokryje $V(G_t)$,
- c/ duży czas obliczeń /k-klika ma $l = 2^k - 2 - k$ podgrafów pełnych/ oraz duża niezbędna pamięć,
- d/ niezależność od kolejności rozpatrywania danych i brak oscylacji.

3.4. Metody wykorzystujące pojęcie k-spójności

Zdefiniowane w p. 3.1 pojęcia k-spójności wierzchołkowej i krawędziowej oraz k-spojenia pozwalają na opracowanie szeregu warstwowych algorytmów grupowania. Grupę określają wówczas:

k-bloki /k-BL_G/, k-komponenty /k-KS_G/ i k-spojenia /k-SP_G/ [78].

Zauważmy, że mamy tu dwie możliwości: albo przy ustalonym $k \in \mathbb{N}$ dodawać do pustego grafu $\prec$ -bliskości $\hat{G} = (V(G), \phi)$ krawędzie $e \in \hat{E}(G)$ i po każdym dodaniu dzielić graf częściowy na k-KS, k-BL, k-SP /otrzymamy wówczas hierarchię H_k / bądź do grafu $G = (V(G), E(G))$, $\text{card } E(G) = w$ stosować k-metodę zmieniając k /uzyskamy wówczas hierarchię H_w /.

DF 3.4.1. Funkcją grafową $\mathcal{F}: \mathcal{G} \rightarrow \mathcal{V}$ nazywamy funkcję ze zbioru wszystkich grafów nieobciążonych G w rodzinę wszystkich podzbiorów zbioru wierzchołków, tzn. jeśli $G = (V, E) \in \mathcal{G}$, to $\mathcal{F}(G) \in 2^{V(G)}$.

Niech $G_t = (V(G), E_t(G)) \in \mathcal{G}$ będzie pewną t-ekstrakcją grafu $G = (V, E, \mu)$. Określmy następujące funkcje grafowe /przy ustalonym $k \in \{1, 2, \dots, n-1\} \subset \mathbb{N}$ /

DF 3.4.2 a/ $\alpha_k : \mathcal{G} \rightarrow \mathcal{V}$; funkcja grafowa

$\alpha_k(G_t) \triangleq \{G' \subset G_t \mid \max \text{ind } V(G') \text{ jest } k\text{-sieczny w } G_t\} \cup \{\{v\} \mid v \in V(G_t), \text{ jeśli } v \text{ nie należy do żadnej } k\text{-komponenty w } G_t\}$,

gdzie: $\max \text{ind } V(G')$ - maksymalny spójny podgraf indukowany przez $V(G')$.

a więc: $\alpha_k(G) = \{k\text{-KS}_G^i\} \cup \{\{v\} \mid v \in V(G) \setminus \bigcup_i V(k\text{-KS}_G^i)\}$.

b/ $\delta_k(G_t) \triangleq \{G' \subset G_t \mid \max \text{ind } V(G') \text{ jest } k\text{-spojeniem w } G_t\} \cup \{\{v\} \mid v \in V(G_t), \text{ jeśli } v \text{ nie należy do żadnego } k\text{-spojenia w } G_t\}$,

czyli: $\delta_k(G) = \{k\text{-SP}_G^i\} \cup \{\{v\} \mid v \in V(G) \setminus \bigcup_i V(k\text{-SP}_G^i)\}$.

c/ $\omega_k(G_t) \triangleq \{G' \subset G_t \mid \max \text{ind } V(G') \text{ jest } k\text{-kliką w } G_t\} \cup \{\{v\} \mid v \in V(G_t), \text{ jeśli } v \text{ nie należy do żadnej } k\text{-kliki w } G_t\}$,

a więc: $\omega_k(G) = \{k\text{-KL}_G^i\} \cup \{\{v\} \mid v \in V(G) \setminus \bigcup_i V(k\text{-KL}_G^i)\}$.

d/ $\kappa_k(G_t) = \{G' \subset G_t \mid \max \text{ind } V(G') \text{ jest } k\text{-blokiem w } G_t\} \cup \{\{v\} \mid v \in V(G_t), \text{ jeśli } v \text{ nie należy do żadnej } k\text{-kliki w } G_t\}$,

czyli: $\kappa_k(G) = \{k\text{-BL}_G^i\} \cup \{\{v\} \mid v \in V(G) \setminus \bigcup_i V(k\text{-BL}_G^i)\}$.

WN. 3.4.1. a/ Następujące funkcje grafowe są równoważne:

$\alpha_1 = \kappa_1 = \delta_1$ /wynika to z DF 3.1.14, 3.1.21, 3.1.22/

b/ $\alpha_1(G) = V(G)$

F 3.4.1. Maksymalna spójność wierzchołkowa, jaką można osiągnąć w grafie G , w którym $\text{card } V(G) = n$, $\text{card } E(G) = w$ jest całkowitą częścią liczby $2w/n$ [28].

WN 3.4.2. Dla $k > \left\lceil \frac{2w}{n} \right\rceil$, $k \in \mathbb{N}$ zachodzi: $\kappa_k(G) = \{\{v\} \mid v \in V(G)\}$

F 3.4.2.[28] a/ Każda para wierzchołków $v_1, v_2 \in V(k\text{-BL}_G)$ jest wzajemnie osiągalna po co najmniej k drogach wierzchołkowo rozłącznych,

b/ Każda para wierzchołków $v_1, v_2 \in V(k\text{-KS}_G)$ jest połączona przez co najmniej k różnych dróg krawędziowo rozłącznych.

TW. 3.4.1. Dla $\forall (G \in \mathcal{G}) \forall (k \in \mathbb{N})$ każdy indukowany podgraf w następującej sekwencji $S_1 = \langle k\text{-KL}_G, k\text{-BL}_G, k\text{-KS}_G, k\text{-SP}_G, 1\text{-KS}_G \rangle$ jest podgrafem swego następnika /o ile go posiada/.

Dowód:

Zauważmy, że każda $k\text{-KL}_G, k\text{-BL}_G, k\text{-KS}_G, k\text{-SP}_G$ ma co najmniej $k+1$ wierzchołków. Ponadto [28] stopień spójności wierzchołkowej nie może przewyższyć stopnia spójności krawędziowej $\Rightarrow k\text{-BL}_G \subset k\text{-KS}_G$ i $k\text{-KL}_G \subset k\text{-BL}_G$.

Każda para wierzchołków $v_1, v_2 \in V(k\text{-KS}_G)$ jest połączona przez co najmniej k różnych dróg krawędziowo rozłącznych /DF 3.1.29/ $\Rightarrow \forall (v_i \in V(k\text{-KS}_G)) (\deg v_i \geq k) \Rightarrow k\text{-KS}_G \subset k\text{-SP}_G$. Z DF 3.1.22 $k\text{-SP}_G$ jest spójny $\Rightarrow k\text{-SP}_G \subset 1\text{-KS}_G$ ■

WN. 3.4.3. Sekwencja $S_1 = \langle k\text{-KL}_G, k\text{-BL}_G, k\text{-KS}_G, k\text{-SP}_G, 1\text{-KS}_G \rangle$ odzwierciedla precyzję grafowych metod grupowania, tzn. w sekwencji $S_1' = \langle \omega_k, \alpha_k, \delta_k, \alpha_1 \rangle$ każda metoda jest precyzyjniejsza od swego następnika /daje drobniejsze podziały wierzchołków $V(G)$ /.

F. 3.4.3. Grafowe metody $\alpha_1, \alpha_k, \delta_k$ /przy ustalonym $k \in \{1, 2, \dots, n-1\}$ i dodawaniu do grafu $\hat{G} = (V(G), \emptyset)$ kolejnych krawędzi $e \in \hat{E}(G)$ dają ciąg podziałów tworzących hierarchię $\mathbb{H}_G$.

Dowód w [78].

WN. 3.4.4. Dla każdej metody $\mathcal{F} \in \{\alpha_1, \alpha_k, \delta_k\}$ spełnione są następujące warunki:

- 1/ niezmienniczość: $(G' \text{ izomorficzny z } G) \Rightarrow (\mathcal{F}(G') \text{ izomorficzne z } \mathcal{F}(G))$,
 - 2/ niezawieranie: $(C, C' \in \mathcal{F}(G)) \wedge (C \subset C') \Rightarrow (C \equiv C')$,
 - 3/ izolacji: $(E(G) = \emptyset) \Rightarrow \mathcal{F}(G) = \{\{v_i\} | v \in V(G)\}$,
 - 4/ pochłanianie: jeśli $G' = (V(G), E(G'))$ - podgraf częściowy grafu $G = (V(G), E(G))$, $E(G') \subset E(G)$, to
- $$(C' \in \mathcal{F}(G)) \Rightarrow (\exists (C \in \mathcal{F}(G)) (C' \subset C)).$$

TW. 3.4.2. $1\text{-KS}_G, k\text{-KS}_G, k\text{-SP}_G$ generowane odpowiednio przez metody $\alpha_1, \alpha_k, \delta_k$ są rozłączne.

Dowód:

Dla 1-KS_G jest to na mocy DF 3.1.14 oczywiste. Dla metody δ_k : niech $V(k-SP^1) \cap V(k-SP^2) \neq \emptyset \Rightarrow G = (V(k-SP^1) \cap V(k-SP^2) = V, E)$ jest spójny z DF 3.1.22 i $\forall (v_i \in V)$ ($\deg v_i \geq k$) $\Rightarrow G = (V, E)$ jest k-spojeniem.

/Metoda α_k /: Niech: $k-KS^1, k-KS^2$ - dwa podgrafy k-sieczne i takie, że $V(k-KS^1) \cap V(k-KS^2) = A \neq \emptyset$.

Niech $v_1 \in V(k-KS^1), v_2 \in V(k-KS^2), v \in A$. Z F.3.4.2 b/ wynika, że $\text{card}\{D(v_1, v)\} \geq k$ i $\text{card}\{D(v_2, v)\} \geq k$ przy czym są to drogi krawędziowo rozłączne. Wobec tego $\text{card}\{D(v_1, v_2)\} \geq k$ i graf $G = (V(k-KS^1) \cup V(k-KS^2), E)$ jest k-sieczny, czyli jest k-komponentą ■

Powyższe twierdzenia, wnioski i fakty umożliwiają opracowanie algorytmów grafowych metod $\alpha_1, \delta_k, \alpha'_k, x_k, \omega_k$ o złożoności odpowiednio: $O(n^2), O(n^3), O(n^2w), O(n^2w(\frac{n}{k})), O(2^{\frac{n}{k}})$, np. [127].

Własności niektórych z omówionych metod przetestowano na empirycznej dokumentacyjnej bazie danych w rozdziale 4.

3.5. Podsumowanie

Teoria grafów umożliwia zunifikowanie i jednorodne ujęcie większości algorytmów podziału i grupowania /szczególnie gdy jedyną informacją jest macierz podobieństw lub odległości obiektów/. Definicje skupień mają intuicyjnie jasną interpretację. Ponadto umożliwia ona bardziej szczegółową analizę własności algorytmów [78, 140]: ich złożoności obliczeniowej, liczby poziomów hierarchii, liczby grup na poziomie, jakości podziału ostatecznego, itp. O zainteresowaniu grafowymi metodami grupowania świadczą statystyczne badania [4, 21, 45, 51] testujące ich pracę.

Z drugiej strony, znaczna większość klasycznych już algorytmów generowania podziałów monotematycznych, które wykorzystywano w eksperymentalnych s.i.d. /np. SMART/, to algorytmy heurystyczne, testowane w dodatku na różnych zbiorach danych [100, 106, 109, 116] Dlatego porównywanie własności tych metod /które są odpowiednią kombinacją procedur omówionych w pracy [138] / może mieć tylko orientacyjny/jakościowy/charakter. I choć możliwe jest przedstawienie niektórych z nich na gruncie teorii grafów /np. w algorytmie

Rocchia grupami początkowymi są $\langle k, l \rangle$ - gwiazdy, o odpowiednio dużych k i l /, nie wydaje się to celowe.

Próba jakościowego porównania algorytmów grafowych i heurystycznych jest TABELA 1, w której: na podstawie literatury zestawiono własności wybranych algorytmów tworzenia grup tematycznych.

Zestaw cech /numery kolumn/:

- I - niezależność wyników od kolejności rozpatrywania danych,
- II - dopuszczalność podziałów słabych,
- III - złożoność obliczeniowa,
- IV - pracochłonność aktualizacji,
- V - typ wyniku: H - hierarchia, P - polihierarchia, W - warstwa/,
- VI - rozmiar wymaganej pamięci /w zasadzie zależy od konkretnej realizacji a nie typu metody/,
- VII - wrażliwość wyniku na małe zmiany danych,
- VIII - regularny kształt elementów podziału /elipsoidalny, kubiczny/,
- IX - występowanie oscylacji,
- X - niezmienniczość na przekształcenia liniowe danych,
- XI - postać danych: MP - macierz podobieństw, OW - opis wektorowy.

Skala ocen:

1 - bardzo mały, 2 - mały, 3 - średni, 4 - duży, 5 - bardzo duży,
N - nie występuje, T - występuje, ? - nie ma sensu.

Na podstawie TABELI 1 można stwierdzić, że część z tych algorytmów nie może mieć szerszego zastosowania w s.i.d. do grupowania obiektów ze względu na: złożoność obliczeniową /MPP - Gotlieb, B_k - Sibson/, niejednoznaczność i niedopuszczanie podziałów słabych /MDS-Zahn, NDH/, oscylacyjność pracy /Dattola/, konieczność opisywania obiektu wektorem wartości cech /Rocchio, Yu, Mc'Queen/.

Najważniejszą wadą s.i.d. z tezaurem słów kluczowych jest brak samokorygowania się podziałów tematycznych b.d. Przegrupowania odbywają się skokowo, przy okazji wprowadzania większych zmian w tezaurysie - obiekty stare powinny wówczas być przecechowane, co jest operacją kosztowną.

Wady tej nie mają s.i.d., które wykorzystują cytowania prac publikowanych. Do opracowania takich danych mogą być wykorzystane tylko niektóre z w/w algorytmów /1 ÷ 6, 8, 13, 14/ jak również grafowe algorytmy [78, 140], w których wykorzystano pojęcie k -spójności. Wyniki dawane przez te algorytmy przebadano w rozdzia-

TABELA 1 Porównanie algorytmów grupowania

Nazwa algorytmu	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	Uwagi
MDS-Zahn	T	N	$2; 0(n^2)$	$2, 0(n^2)$	H, W	2	1	N	N	T	MP	Najpopularniejszy, prosty ale niejednoznaczny
NDH - Zimmerman	T	N	4	3	H, W	4	4	N	N	T	MP	Prosty dla grafów pełnych
MPP - Bonner	1!	N, T	4	3	P, W	4	4	T	N	T	MP	
MPP - Gotlieb	T	T	5!	5!	P	5!	1	T	N	T	MP	Grupowanie słów kluczowych
MPP - Needham	3	T	$3 \div 4$	3	P, W	3	3	T	N	T	MP	Grupowanie słów kluczowych
MPP - pochłaniający	2	T	4	3	W	3	3	T	N	T	MP	Własności mało zbadane
Forel-2-promieniowy	T	T	4	4	P, W	4	3	T	T	N	OW	
KS - Augustson	T	N	$2; 0(n^2)$	2	H, W	2	1	N	N	T	MP	Jeden z najprostszych lecz wyniki często absurdalne
B_k - Sibson	T	T	4	4	P, W	4	4	T	N	T	MP	Jednoznaczny lecz skomplikowany
Yu	T	T	$2; 0(mn)$	2	W	3	3	T	N	N	OW	Grupowanie pytań - nieefektywny dla zmieniających się profili
Rocchio	4	T	$2; 0\left(\frac{n^2}{\log n}\right)$	2	W	3	3	T	N	N	OW	Test "gęstości" - statyczna b.d.
Dattola	3	T	$2; 0(n \log n)$	3	W	4	3	T	T	N	OW	Start z grup początkowych, statyczna b.d.

c.d. TABELI 1

Nazwa algorytmu	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	Uwagi
Schiminovich	T	N	3 iteracje	3	W	4	?	?	T	?	OW	Daje duże grupy tematyczne - grupowanie na podstawie bibliografii
Eaton, Bichteler	T	N	2	2	W	4	?	?	N	?	OW	Prosty, lecz wyniki dobre tylko dla dziedzin uporządkowanych /fizyka, matematyka/
McQueen /Žitkov/	1!	N	2	2	W	2	3	T	T	N	OW	Daje często zaskakujące, absurdalne wyniki
Johnson /single-pass/	1!	N	2; $O(n \log n)$	3	W	2	3	T	T	N	OW	Modyfikacja metody Mc'Queena wymagająca tylko 1 przeglądnięcia danych

le 4.

Zasygnalizować na koniec warto naszą nową koncepcję wykorzystania MPP w tzw. algorytmie ściągającym wierzchołki. Aby nie znajdować wszystkich PP /czas obliczeń!/ a jednocześnie zmniejszyć zależność wyników AS od kolejności danych, przy skromnych wymaganiach co do rozmiaru pamięci, po każdorazowym znalezieniu w grafie G MPP_G o rzędzie $\geq \delta$ zastępuje się go pojedynczym wierzchołkiem /tzw. ściąganie wierzchołków/. Następnie, po redukcji krawędzi wielokrotnych, uzyskujemy graf prosty o rzędzie obniżonym o $\delta - 1$ /jeśli G był spójny/. Wierzchołki ściągnięte stanowią jedną z grup. W grafie zredukowanym znów szuka się MPP, itd. W etapie końcowym łączy się grupy podobne /miarą podobieństwa jest max min obciążenie drogi w G między centrami /DF 3.1.17/ grafów tych grup/. Do znajdowania centrów pierwotnie zastosowano algorytm Hakimiego [58] o złożoności $O(n^4)$ - znane są i inne [83] okazało się jednak, że prawie takie same podziały uzyskuje się, korzystając z centrów drzew zbalansowanych /o prawie - minimalnej średnicy/ zbudowanych na grafie grupy. Zyskujemy na prostocie obliczeń $O(n)$ - twierdzenie Jordana/ i zmniejszamy /do dwóch/ możliwą liczbę centrów.

Wiarygodnych wniosków dotyczących własności tego algorytmu z braku wystarczającego materiału empirycznego nie można podać. Wiadomo na razie, że dla niedużych zadań czas obliczeń nie jest długi a wyniki są dobre, choć zależą od punktu startu. Zagadnienie wymaga jednak dalszych, szczegółowych badań.

4. OPIS EKSPERYMENTÓW I ANALIZA WYNIKÓW ALGORYTMÓW GRUPOWANIA

4.1. Plan eksperymentów

Dla wykazania prawdziwości tezy przeprowadzono szereg eksperymentów na dobranej dokumentacyjnej bazie danych. Bazę tę utworzono z artykułów z lat 1970÷1980. Wybierano je z 15 czasopism fachowych z zakresu informatyki. Dla zmniejszenia pracochłonności obliczeń, jak również z powodu trudności w doborze ekspertów oceniających rezultaty, ograniczono rozmiary b.d. dokonując manualnej selekcji obiektów w sposób opisany w Dodatku 2. Eksperymentalna b.d. /785 obiektów/ w dużym przybliżeniu odpowiada grupie tematycznej, którą wyodrębniłby algorytm Schiminowicha [110], co ilustruje

jak małą precyzją cechuje się ten algorytm.

Przyjęto, że opisem obiektu jest jego bibliografia /informacja niezniesztakowana, obiektywna i nie zmieniająca się w czasie/. Długość opisu wpływa na rozmiar wymaganej pamięci i czas obliczeń. Ze względów praktycznych należało zredukować długość opisu, nawet jeśli miałyby to powodować pogorszenie jakości podziału b.d. Autor zdecydował się na usuwanie z opisu prac starych oraz trudno-dostępnych /nieperiodycznych/ i monografii. Przyjęty sposób redukcji może mieć wpływ na wyniki grupowań.

Ekspertami ExpIII i ExpI byli odpowiednio: autorzy 13 prac przeglądowych o różnych aspektach s.w.i. /bibliografie o ponad 50 pozycjach/ oraz autor.

Na podstawie struktury cytowań prac określono 4 współczynniki podobieństwa między obiektami /artykułami i czasopismami/.

Przyjęto następujący plan eksperymentów:

- 1/ sprawdzenie czy obiektów b.d. nie dobrane tendencyjnie - w dostatecznie dużej i różnorodnej b.d. powinny zachodzić znane prawidłowości bibliometryczne. Zagadnienie to przebadano i omówiono w Dodatku 2. Przy okazji sprawdzono jaki wpływ na te prawidłowości mają zastosowane redukcje opisów obiektów,
- 2/ sprawdzenie czy czasopisma z wybranego zestawu są ze sobą odpowiednio silnie związane tematycznie - gdyby tak było, to w praktycznych realizacjach algorytmów podziału tematycznego należałoby, dla zmniejszenia pracochłonności rozważyć możliwość ich dwuetapowej pracy: wyodrębnianie grup czasopisma następnie podział każdej z nich na monotematyczne zbiory obiektów. Temu zagadnieniu poświęcono rozdział 4.3.
- 3/ zbadanie przydatności przyjętego opisu obiektów dla automatyzowanych systemów wyszukiwania informacji - poprzez sprawdzenie hipotez, że podobieństwo bibliografii implikuje podobieństwo tematyczne obiektów oraz, że obiekty grupują się w przyjętej przestrzeni cech. Omówiono to w rozdziałach 4.4 i 4.5,
- 4/ przebadanie własności zdefiniowanych współczynników podobieństwa /rozdział 4.4/,
- 5/ przebadanie własności podziałów i dobrane parametrów w grafowych algorytmach podziału tematycznego b.d.

W zakres tego zagadnienia wchodzi zbadanie wpływu doboru funkcji podobieństwa i sposobu redukcji danych na podziały wynikowe

uzyskiwane różnymi algorytmami grupowania oraz zbadania wpływu aktualizacji podziału na jego stabilność. Porównano również grafowe algorytmy grupowania wykorzystujące pojęcie k-spójności. Zagadnienie to przeanalizowano w rozdz. 4.5.

Artykuły z lat 1976-77 grupowano na podstawie związków między ich zredukowanymi opisami wykorzystując różne definicje współczynników podobieństwa i różne algorytmy podziału. Obliczono korelacje współczynników podobieństwa oraz ich wpływ na wynik grupowania. Wyniki podziału porównano /dla każdego pytania/ z ocenami ExpI i ExpII. Prace z lat 1978-1980 stanowiły zbiór aktualizujący podział uzyskany w poprzednim etapie w celu sprawdzenia dynamiki zmian w podziale oraz częściowego zbadania procedur aktualizujących. Decyduje to bowiem o własnościach aplikacyjnych s.w.i. z dzieloną bazą danych.

4.2. Pojęcia wstępne

4.2.1. Podobieństwo między obiektami

Zgodnie z rozdziałem 2 należy określić współczynniki podobieństwa między obiektami. Na podstawie analizy struktury cytowań /Dodatek 1/ proponujemy następujące współczynniki podobieństwa:

DF 4.2.1.1. k-retrospektywne:

$$p_r(d_i, d_j) \triangleq p(\Omega^k(d_i), \Omega^k(d_j))$$

gdzie: p - funkcja podobieństwa /symetryczna, zwrotna, nieujemna/,

Ω^k - relacja k-cytowania /patrz Dodatek 1/.

Przykładem podobieństwa k-retrospektywnego może być:

$$p_r(d_i, d_j) = \frac{\text{card} \{ \Omega^k(d_i) \cap \Omega^k(d_j) \}}{\text{card} \{ \Omega^k(d_i) \cup \Omega^k(d_j) \}}, \text{ k-zadane}$$

DF 4.2.1.2. k-prospektywne:

$$p_p(b_i, b_j) \triangleq p(\Gamma_k(b_i), \Gamma_k(b_j)), \text{ k-zadane}$$

gdzie:

$\Gamma_k(b_i) \triangleq \{d_s ; b_i \in \Omega^k(d_s)\}$ - zbiór dokumentów d_s

k-cytujących pracę b_i .

Okazuje się, że w/w k-podobieństwa, choć atrakcyjne obliczeniowo, są niewygodne w stosowaniu; mianowicie:

- a/ podobieństwo k-prospektywne zmienia się w dynamicznej, dokumentacyjnej b.d.,
- b/ nie różnicuje się ważności cytowanych prac zależnie od ich wieku,
- c/ w praktyce, nie tyle ważna jest liczba poziomów poddrzewa cytowań, co jego wysokość mierzona różnicą lat na osi czasu /zapewnienie równego "zagłębienia" cytowań/,
- d/ wartości podobieństw między pracami, których daty opublikowania znacznie się różnią są z reguły bliskie zeru, mimo silnych związków między tymi pracami.

Wypadałoby różnicować podobieństwa w zależności od stopni wierzchołków w hierarchicznym grafie cytowań. Wydaje się, że można wykorzystać tu metodę określania odległości między grafami hierarchicznymi zaproponowaną w [126]:

Jeśli obiekt d_i / $\bar{d}_j$ / jest korzeniem w indukowanym podgrafie cytowań $T_i = (V_i, E_i)$ / $T_j = (V_j, E_j)$ / /patrz Dodatek 1/ to:

$$DF\ 4.2.1.3. \quad \varphi_{T_i, T_j}(d_i, d_j) \triangleq L(T_i \rightarrow T_j)$$

gdzie:

$L(T_i \rightarrow T_j)$ - minimalna liczba nietrywialnych operacji elementarnych konieczna do przekształcenia grafu skierowanego T_i w graf T_j ,

/elementarna operacja $\equiv$ dodanie, skasowanie pojedynczego łuku/,

$T_i, T_j \subset \mathcal{G}_C$ - zorientowany graf cytowań reprezentujący macierz cytowań C

$$V_i = \{d_i\} \cup \bigcup_{k=1}^l \Gamma_k^{(-1)}(d_i), \quad V_j = \{d_j\} \cup \bigcup_{k=1}^p \Gamma_k^{(-1)}(d_j)$$

Własności odległości φ /DF 4.4.1.2.3/ są następujące:

- a/ nietrudno pokazać, że φ_{T_i, T_j} spełnia aksjomaty metryki w zbiorze wszystkich grafów hierarchicznych zakorzenionych z dokładnością do izomorfizmu,
- b/ obliczenie odległości $\varphi_{T_i, T_j}(d_i, d_j)$ wymaga $O(n_i n_j l_i^2 l_j^2)$ operacji,

gdzie: $n_x = \text{card } V_x$, l_x - długość maksymalnej drogi w $T_x = (V_x, E_x)$,

c/ DF 4.2.1.3 daje możliwość obliczania podobieństw między pracami o znacznie różniących się datach publikacji,

d/ w DF 4.2.1.3 nie uwzględnia się różnicy lat między opublikowaniem pracy cytującej a cytowaniami.

Następująca propozycja uwzględnia wpływ czasu na wartości podobieństwa: w zakorzenionych $/d_i, d_j/$ poddrzewach T_i, T_j należy obciążyć łuki cytowań wagami μ odwrotnie proporcjonalnymi do różnicy lat między pracą cytującą a cytowaną. Wagi te dla prostoty dalszych obliczeń należy unormować tak, aby suma obciążeń łuków wychodzących z każdego wierzchołka była jedynką.

Pracę $d_i / d_j /$ możemy przedstawić wówczas jako wektor

$\tilde{d}_i = \langle w_{i_1}, w_{i_2}, \dots, w_{i_s} \rangle$ / $\tilde{d}_j = \langle w_{j_1}, w_{j_2}, \dots, w_{j_r} \rangle$ /, w którym atrybutami $B_i = \langle B_{i_1}, B_{i_2}, \dots, B_{i_s} \rangle$ / $B_j = \langle B_{j_1}, B_{j_2}, \dots, B_{j_r} \rangle$ są prace cytowane /nawet pośrednio/ przez $d_i / d_j /$, gdzie:

$$w_{i_k} \triangleq \begin{cases} \sum_D \prod_{\tau \in \mathbb{Z}_k} \mu(\tau) / \sum_{k=1}^s w_{i_k} & ; \text{jeśli istnieje droga z } d_i \text{ do } d_k, \\ 0 & ; \text{jeśli nie istnieje droga z } d_i \text{ do } d_k. \end{cases}$$

$D = \{\mathbb{Z}_k\}$ - zbiór wszystkich różnych zorientowanych dróg $\mathbb{Z}_k = D(d_i, d_k)$ z d_i do d_k

τ - łuk drogi $\mathbb{Z}_k$

Wówczas, po ujednoliceniu opisu dokumentów d_i, d_j według wspólnych atrybutów $B = B_i \cup B_j$:

$$\text{DF 4.2.1.4. } p(d_i, d_j) \triangleq \sum_{k=1}^l w_{i_k} w_{j_k}, \quad l = \text{card}(B_i \cup B_j)$$

Osobną sprawą jest dobór wag μ . Można zastosować tu wagi geometryczne lub harmoniczne używane w ekonometrii [5] do uwzględniania wpływu czasu, np.

$$\mu(t, n) \triangleq \frac{1}{n-1} \sum_{i=2}^t \frac{1}{n-i+1}; \quad t = 2, 3, \dots, n, \quad n - \text{ustalone}$$

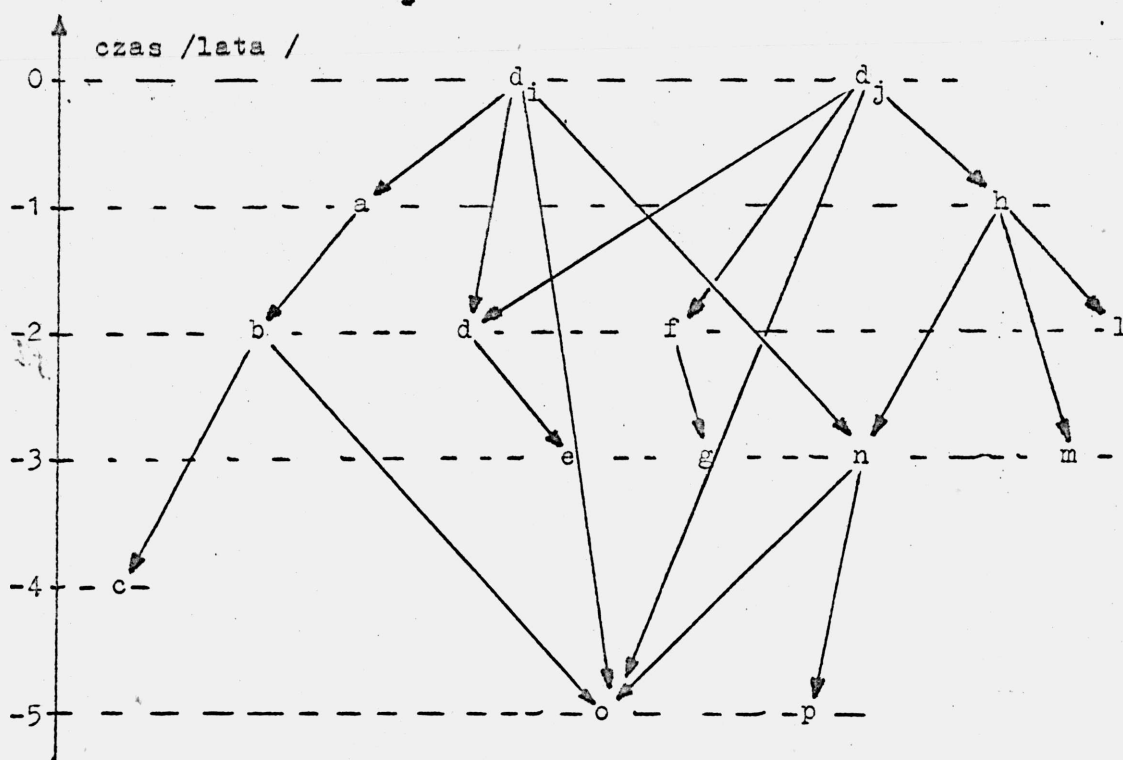
o własności:

$$\sum_{t=2}^n \mu(t, n) = 1$$

Metoda ta jest pracochłonna /wymaga zapamiętania wszystkich

dróg w grafie cytowań/ tylko w fazie inicjacji bazy danych. Cechowanie napływających, nowych dokumentów jest bardzo proste /wykorzystujemy dokumenty już ocechowane/. Zachodzi tylko konieczność ograniczania wysokości drzewa cytowań, bowiem wektory wag opisu stają się nadmiernie długie.

Obliczanie podobieństw zilustrujemy przykładem /rys. 4.1/.



Rys. 4.1. Przykładowy graf cytowań

Obciążenia /harmoniczne/ łuków podaje TABELA 2.

Podobieństwo między dokumentami d_i, d_j wyniesie odpowiednio:

a/ 1-retrospektywne /nieważone/: $p_r(d_i, d_j) \Big|_{k=1} = \frac{1}{3} \approx 0,333$

b/ k-retrospektywne /nieważone/, $k \leq 3$ $p_r(d_i, d_j) = 0,3846$

c/ podobieństwo grafów /DF 4.2.1.3/: $p_g(d_i, d_j) \triangleq \frac{1}{1+g_{T_i, T_j}(d_i, d_j)} = 0,091$

d/ retrospektywne ważne /DF 4.2.1.4/: $p(d_i, d_j) = 0,0395$

Jak widać, podobieństwa mogą się znacznie różnić między sobą. Ma to oczywiście wpływ na wyniki algorytmów grupowania. Zależności między współczynnikami przetestowano a wyniki opisano w p. 4.4.

TABELA 2. Macierz opisu struktury z rys. 4.1.

	a	b	c	d	e	f	g	h	l	m	n	o	p
$\tilde{d}_i$	0,502			0,282							0,172	0,044	
$\tilde{d}_j$				0,254		0,254		0,452				0,04	
a		1											
b			0,621									0,379	
d					1								
f							1						
h									0,47	0,265	0,265		
n												0,5	0,5
$\tilde{d}_i$	0,204	0,204	0,127	0,115	0,115	0	0	0	0	0	0,07	0,13	0,035
$\tilde{d}_j$	0	0	0	0,122	0,122	0,122	0,122	0,217	0,102	0,06	0,06	0,048	0,029

$$/\mu(t,n) = \mu(t,5)/$$

4.2.2. Podobieństwo między czasopismami

Na podstawie macierzy odsyłań między czasopismami /Tabela II - Dodatek 2/ można określić współczynniki podobieństwa $p(i,j)$ między czasopismami i oraz j dla ustalonej bazy danych wzorem:

$$DF\ 4.2.2.1. \quad p(i,j) = \begin{cases} 1/2 \frac{n_i}{N_i} \frac{a_{ij}}{\sum_j} + \frac{n_j}{N_j} \frac{a_{ji}}{\sum_i} & ; i = 1, 2, \dots, 15; \\ & j = i+1, \dots, 15. \\ 1/2 \frac{n_i}{N_i} \frac{a_{i,16}}{\sum_j} & ; i = 1, 2, \dots, 15 ; j = 16 \end{cases}$$

gdzie: n_i - liczba artykułów z czasopisma i w bazie danych,

N_i - całkowita liczba przeglądniętych artykułów czasopisma i ,

$a_{ij} \neq a_{ji}$ - liczba cytowań czasopisma j przez czasopismo i ,

$$\sum_j = \sum_{k=1}^{16} a_{k_j} \quad , \quad \sum_i = \sum_{k=1}^{15} a_{ik}.$$

Wyniki obliczeń przedstawiono w Tabeli 3 /p.4.3/. Na podstawie wartości podobieństw dokonano grupowania czasopism.

4.2.3. Podobieństwo między grupowaniami

Aby porównywać wyniki grupowań należy zdefiniować miarę podobieństwa tych grupowań:

a/ porównywanie grupowań mocnych /podziałów/

Przyjmijmy trzy, naturalne w tym przypadku, założenia:

- 1/ dyskretność podziału /każdy obiekt należy do jednej i tylko jednej grupy/,
- 2/ grupa jest dobrze definiowana przez obiekty, które do niej należą, jak też przez obiekty, które do niej nie należą,
- 3/ wszystkie obiekty mają identyczną wagę w określaniu grup.

W eksperymentach będzie wykorzystywana miara Randa [95] podobieństwa między podziałami $P' = \{P'_1, P'_2, \dots, P'_q\}$,
 $P'' = \{P''_1, P''_2, \dots, P''_r\}$:

$$\text{DF 4.2.3.1. } p^*(P', P'') \triangleq \left[\binom{n}{2} - \frac{1}{2} \left(\sum_{i=1}^q \left(\sum_{j=1}^r n_{ij} \right)^2 + \sum_{j=1}^r \left(\sum_{i=1}^q n_{ij} \right)^2 \right) - \sum_{i=1}^q \sum_{j=1}^r n_{ij}^2 \right] / \binom{n}{2}$$

gdzie:

$$n = \text{card} \bigcup_{i=1}^q P_i' = \text{card} \bigcup_{j=1}^r P_j'' - \text{liczba grupowanych obiektów,}$$

$$n_{ij} = \text{card} (P_i' \cap P_j'').$$

Własności miary $p^*(P', P'')$:

a/ $p^*(P', P'') \in [0, 1] \subset \mathbb{R}$,

b/ $p^*(P', P'') = 0 \Leftrightarrow P' = \{\bigcup_i P_i'\}, P'' = \{P_1'', P_2'', \dots, P_n''\}, P_i'' = \{b_i\}$

$p^*(P', P'') = 1 \Leftrightarrow \forall (P_i' \in P') \exists! (P_j'' \in P'') (P_i' \equiv P_j'')$,

c/ $p^*(P', P'') = p^*(P'', P')$,

d/ miara $g^*(P', P'') = 1 - p^*(P', P'')$ jest metryką w zbiorze wszystkich podziałów danego zbioru obiektów.

b/ porównywanie podziałów słabych

W przypadku podziałów słabych nie jest spełnione założenie 1/ i niekiedy 3/ z poprzedniego podpunktu, a sprawa definicji adekwatnej miary podobieństwa komplikuje się.

W podziale słabym $P' = \{P_1', P_2', \dots, P_q'\} / P_i' \cap P_j' \neq \emptyset /$ dla każdej z $s = \binom{n}{2}$ możliwych, różnych nieuporządkowanych par $\alpha'_k / \forall (1 \leq k \leq s) (\alpha'_k = \{d_{i_k}, d_{j_k}\}; d_{i_k}, d_{j_k} \in B = \bigcup_{i=1}^q P_i' /$ określimy zmienną binarną γ'_k wzorem:

$$\gamma'_k = \begin{cases} 1, & \exists (P_i' \in P') (\{d_{i_k}, d_{j_k}\} \subset P_i') , \\ 0, & \text{przeciwnie.} \end{cases}$$

Wówczas:

$$\text{DF 4.2.3.2. } p(P', P'') \triangleq \frac{1}{s} \sum_{k=1}^s [\gamma'_k \gamma''_k + (1 - \gamma'_k)(1 - \gamma''_k)]$$

gdzie: $\bigcup P_i' = \bigcup P_j''$

Własności podobieństwa $p^*(P', P'')$:

$$1/ p^*(P', P'') \in [0, 1] \subset \mathbb{R},$$

$$2/ (p^*(P', P'') = 1) \Leftrightarrow (P' \text{ izomorficzne z } P''),$$

$$3/ p^*(P', P'') = p^*(P'', P'),$$

4/ $g = 1 - p^*(P', P'')$ nie jest metryką w zbiorze podziałów słabych.

Przykładem są podziały: $P_1 = \{\{1, 2, \dots, 9\}\}$, $P_2 = \{\{1, 2, 3, 4\}, \{4, 5, 6, 7, 8\}, \{7, 8, 9\}\}$, $P_3 = \{\{1, 7, 8, 9\}, \{1, 2, 3, 4\}, \{4, 5, 6, 7, 8\}\}$, dla których nie jest spełniona nierówność trójkąta.

5/ $p^*(P', P'') = p^*(P'', P')$, jeśli P', P'' - podziały mocne.

Ponieważ mała wartość $p^*(P', P'')$ może być skutkiem małej liczności P' bądź małej zgodności z P'' celowym wydaje się zdefiniowanie miary średniej precyzji podziału /odpowiednik makro-precyzji przy wyszukiwaniu informacji/:

$$DF \ 4.2.3.3. \ \Psi(P', P'') \triangleq \frac{1}{q} \sum_{i=1}^q \frac{\text{card}(P'_i \cap P''_i)}{\text{card}(P'_i \cup P''_i)} \in [0, 1], \ P'_i \in P', \ P''_i \in P''_q$$

gdzie przyporządkowania $P'_i \leftrightarrow P''_i$ dobrano tak, aby maksymalizować $\Psi(P', P'')$. Łączna miara zgodności podziałów P', P'' będzie:

$$DF \ 4.2.3.4. \ \Theta(P', P'') \triangleq p^*(P', P'') \cdot \Psi(P', P'')$$

Miary te wykorzystano w rozdziale 4.5.

4.3. Analiza związków tematycznych między czasopismami

Na podstawie definicji 4.2.2.1 i Tabeli II /Dodatek 2/ obliczono /kalkulator/ współczynniki podobieństw między czasopismami. Tabela 3 zawiera wyniki tych obliczeń. Posłużyły one do grupowania tematycznego czasopism za pomocą grafowych algorytmów: MDS/1-NN/, 1-FN, MPP opisanych w rozdziale 3. Wyniki przedstawiono w postaci dendrogramów na rys. 4.2 ÷ 4.4.

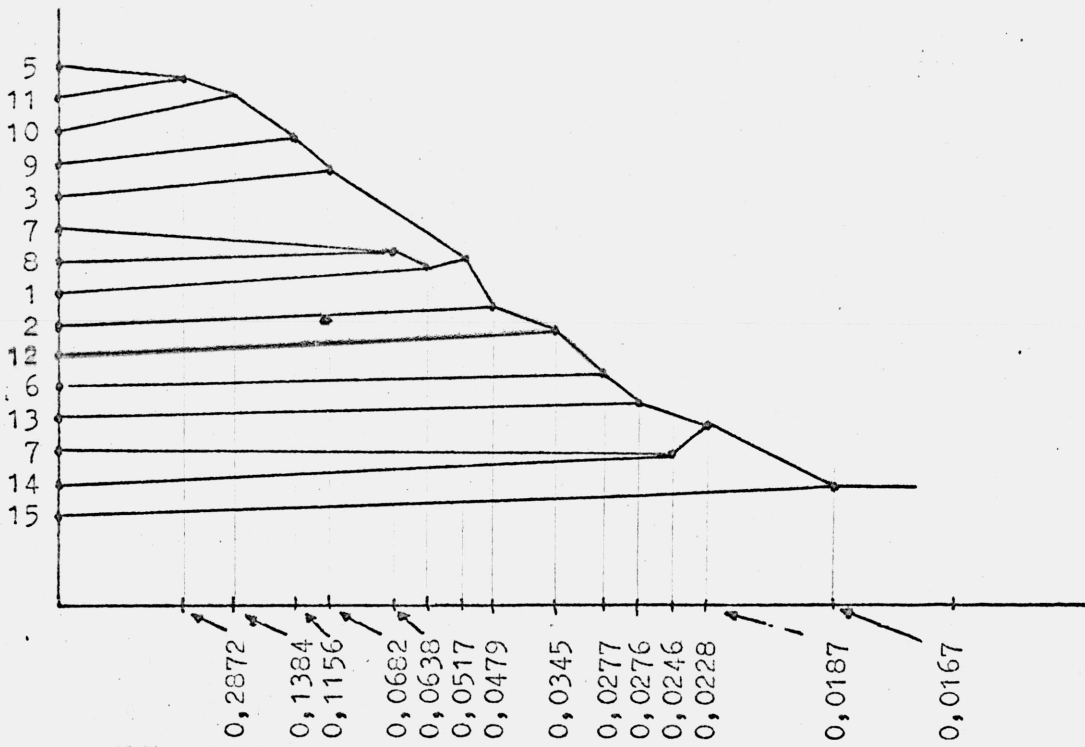
Z wyników algorytmu 1-NN można sądzić, że analizowana grupa czasopism jest jednorodna tematycznie /z wyjątkiem być może IBM J. of Res. and Develop./ Algorytm 1-FN wyróżnia 5 grup, natomiast algorytm MPP odrzuca Pattern Recogn. i Software Eng. jako mało tematycznie podobne do reszty.

TABELA 3. Odesłania między czasopismami

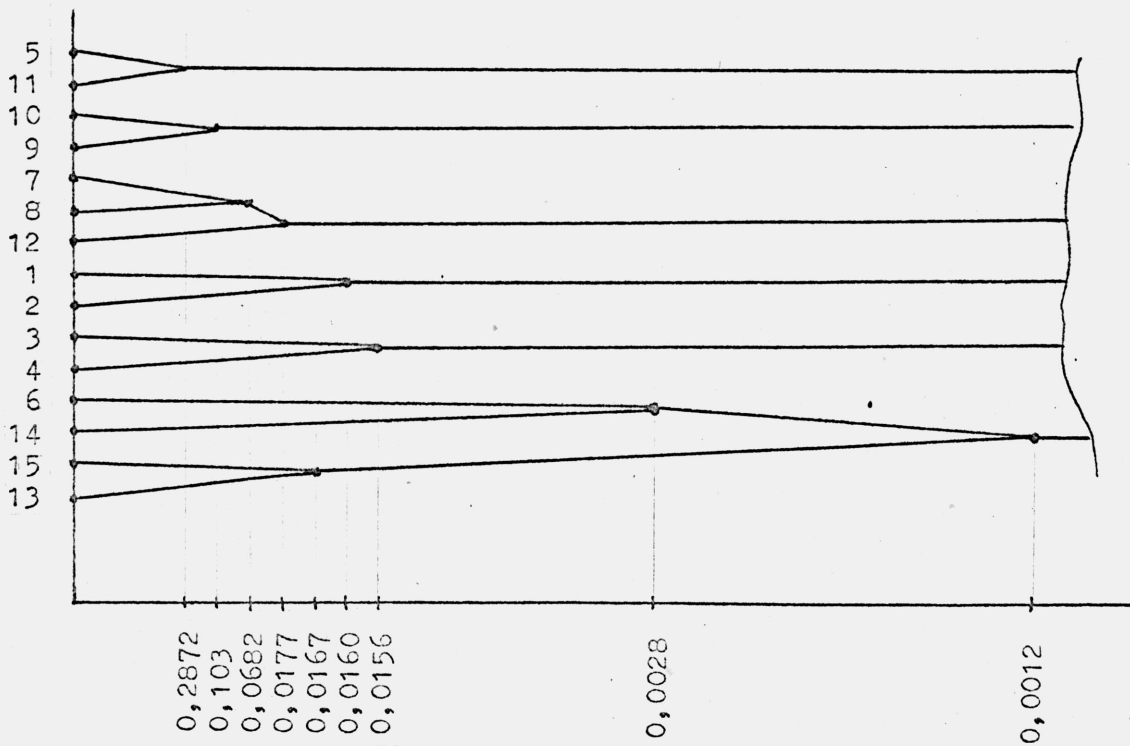
	1	2	3	4	5	6	7	8	9
1	73 0,0292	5 0,016		0,0001		6 0,0139	0,0007	0,0517	1 0,0038
2	7	22 0,034	2 0,02	3 0,0152			2 0,0137	0,0123	
3		1	1 0,012	3 0,0156		0,007			
4	1	1	2						
5									
6	88		1			64 0,0483	1 0,002	0,0105	0,0105
7	10	2				7	11 0,0201		
8	78	2				4		8 0,1582	8 0,0139
9	12					9			108 0,0846
10	20	2	1			8		4	125
11	67	3	2		2	10	5	2	119
12	31	8				2	5		25
13	30					1		3	
14	31					3	6		
15	15	4				3			
$\sum^1$	463	50	9	6	2	117	30	17	386
$\frac{x_i}{y_i}$	0,0764	0,0084	0,0153	0,1304	0,1823	0,1261	0,0613	0,6133	0,2727

TABELA 3 /c.d./

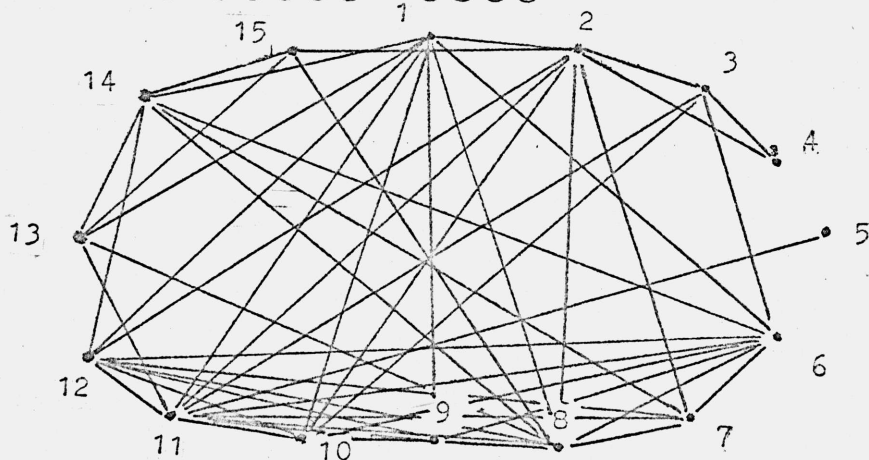
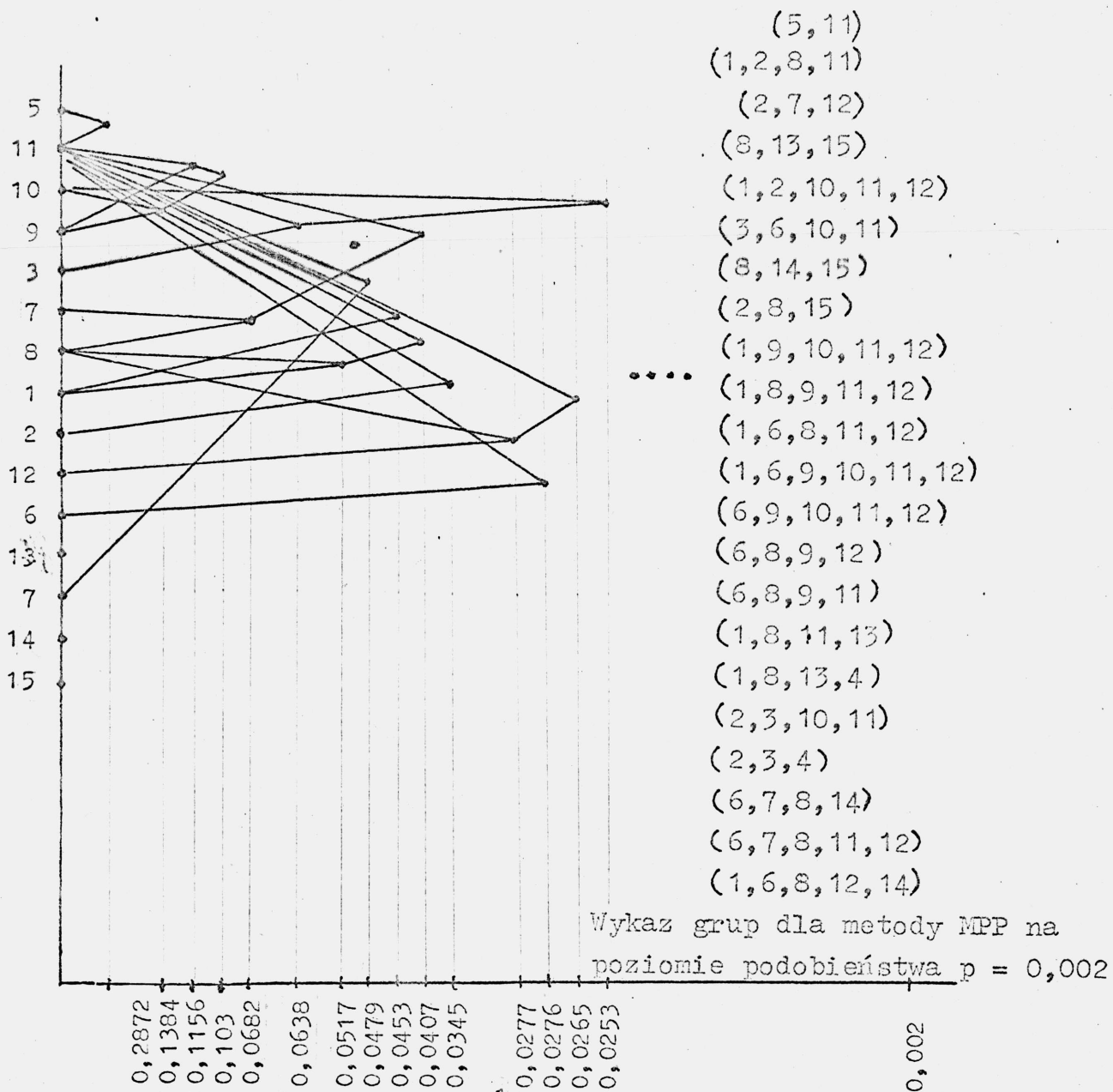
	10	11	12	13	14	15	16	Σ^2
1	5 0,0119	12 0,0453	10 0,0087	1 0,003	2 0,008	1 0,0019	5	121
2	0,0095	0,0345	3 0,0134	1 0,0001		1 0,0041	13	59
3	0,0265	0,0638	1 0,0004			2 0,0007	13	21
4								4
5		0,2872						0
6	3 0,0169	14 0,0276	10 0,0036	4 0,0012	0,0028	3 0,0019	104	292
7		0,0479	5 0,0177		1 0,0228		2	38
8	7 0,0682	4 0,0407	16 0,0277	10 0,0246	5 0,0087	7 0,0121	28	177
9	60 0,103	62 0,1156	31 0,0187	2 0,0009			32	317
10	89 0,1469	85 0,1384	21 0,0145	1 0,0006		2 0,0013	14	372
11	69	144 0,2072	43 0,0253	5 0,0029	1 0,0016	3 0,0017	28	503
12	3	6	36 0,035	1 0,0006	0,0020	0,0018	21	138
13				14 0,0179	1 0,0025	2 0,0167	15	72
14		3	4	1	11 0,0667	2 0,0069	24	85
15			8	18	2	16 0,0283	11	77
Σ^1	236	331	219	58	23	45	315	
$\frac{x_i}{y_i}$	0,4766	0,5744	0,1645	0,0822	0,2196	0,1004		



Rys. 4.21. Grupowanie metodą MDS /1-NN/



Rys. 4.3. Grupowanie metodą 1-FN



Rys. 4.4. Grupowanie metodą MPP / $p \geq 0,0020$ / - polihierarchia i odpowiadający graf $G(0,0020)$

Jak widać, nawet dla tej samej funkcji podobieństwa, algorytmy mogą dawać różne wyniki. Wyniki te potwierdzają, że wybrany zestaw czasopism /za wyjątkiem być może Software Eng./ jest dość jednolity tematycznie /mała liczba odesłań "na zewnątrz" grupy/. Pokazują one również możliwość obniżenia złożoności obliczeniowej algorytmów grupowania obiektów /artykułów/ przez zastosowanie dwuetapowej realizacji grupowania; najpierw wyodrębnianie grup czasopism tematycznie związanych a następnie grupowanie obiektów w obrębie tych czasopism. Miałoby to duże znaczenie przy ewentualnej praktycznej realizacji s.w.i. z dzieloną bazą danych.

4.4. Analiza proponowanych funkcji podobieństwa

Do eksperymentu wybrano z baz danych obiekty z lat 1978-1980, ponieważ mają one największą "głębokość" cytowań pośrednich k-retrospektywnych / $k \leq 4$, DF 4.2.1.1 i DF 4.2.1.4/. Liczbą tych prac wyniosła: 277 /b.d.I/ i 133 /b.d.II/.

A. Wpływ redukcji bibliografii na wartości podobieństwa

Obliczono współczynniki podobieństwa dla każdej z $\binom{133}{2}$ par obiektów z b.d.II na podstawie ich opisu:

- a/ zredukowanego do prac sprzed co najwyżej 10 lat, które pochodzą z wybranego zestawu czasopism bądź są monografiami /BBD I/,
- b/ jak w a/, dodatkowo zredukowanego o monografię /BBD II/.

Współczynniki podobieństwa obliczano według:

- 1/ definicji DF 4.2.1.1 ,
- 2/ definicji DF 4.2.1.4 ,
- 3/ definicji DF 4.2.1.3 .

Sześć oddzielnych zbiorów ponumerowanych par posortowano przy pomocy podprogramów standardowych według:

- ich numerów /dla obliczenia korelacji/,
- ich współczynników podobieństwa /dla obliczenia korelacji rangowej - jest to istotna miara dla badania działania algorytmów grupowania, które zależą tylko od uporządkowania współczynników podobieństwa/, oraz obliczono wartość współczynników korelacji.

Wyniki tych obliczeń zebrano w TABELI 4.

TABELA 4. Macierz współczynników korelacji

	a-1*)	b-1	a-2	b-2	a-3	b-3
**) a-1	1,000	0,564	0,341	0,283	0,201	0,152
	1,000	0,623	0,437	0,395	0,253	0,207
b-1		1,000	0,255	0,231	0,300	0,199
		1,000	0,283	0,411	0,308	0,272
a-2			1,000	0,748	0,332	0,331
			1,000	0,803	0,378	0,342
b-2				1,000	0,297	0,312
				1,000	0,322	0,432
a-3					1,000	0,345
					1,000	0,473
b-3						1,000
						1,000

Symetrycznie

*) Kod "a-1" oznacza: typ redukcji opisu - a, definicja podobieństwa
- 1 /patrz str. 61 /

**) Umowa dla TABEL 4 i 5

- pierwszy wiersz - współczynnik korelacji zwykłej,
- drugi wiersz - współczynnik korelacji rangowej.

Wnioski:

- 1/ wartości korelacji rangowej są większe od korelacji zwykłej.
Wynika to raczej z pewnej niewrażliwości współczynnika Kendalla lub Spearmana na zmiany wartości podobieństw, niż z cech danych,
- 2/ generalnie biorąc, wartości korelacji nie są duże; szczególnie niskokorelowane z innymi są podobieństwa liczone na podstawie DF 4.2.1.1. Jedną z przyczyn jest fakt, że wartość podobieństwa liczona z DF 4.2.1.3 nigdy nie jest równa zeru, podczas, gdy dla DF 4.2.1.1 znaczna liczba par miała podobieństwa zerowe. Odwrotnie, niezerowe wartości podobieństw par /wg DF 4.2.1.1/ były zwykle istotnie większe od podobieństw tych par liczonych według innych definicji. Wartość średnia i rozrzut wartości podobieństw teorio-grafowych były zaskakująco niskie,
- 3/ redukcja nie zmienia w zasadniczy sposób powyższych zależności /dla b.d. II/,

- 4/ zwraca uwagę dużą wrażliwość teorio-grafowej miary podobieństwa na redukcję bibliografii oraz małą wrażliwość miary $2/DF$ 4.2.1.4/,
- 5/ oczywiście, jeszcze słabszych powiązań należy się spodziewać w b.d. I /277 obiektów/, gdzie są obiekty, które cytują wspólnie tylko jedną pozycję i jest nią monografia,
- 6/ usunięcie monografii z bibliografii nie wywołuje radykalnych zmian zależności dla prac o wystarczająco dużej głębokości cytowań /w naszym przypadku 8 lat/.

B. Wpływ głębokości cytowań

W definicji DF 4.2.1.1 mamy możliwość uwzględnienia wpływu głębokości cytowań /długości łańcucha cytowań pośrednich/ na wartość współczynnika podobieństwa k -retrospektywnego, poprzez zmianę liczby naturalnej k .

W naszym przypadku maksymalna długość łańcucha cytowań wynosiła 4, stąd: $1 \leq k \leq 4$. W TABELI 5 przedstawiono współczynniki korelacji rangowej dla wariantów z TABELI 4 przy różnych k . Wzorcem porównawczym jest odpowiedni wariant przy $k=4$, co widać z pierwszego wiersza tabeli.

TABELA 5. Wpływ głębokości cytowań

variant k \	a-1	a-2	a-3	b-1	b-2	b-3
k=4	1,000	1,000	1,000	1,000	1,000	1,000
k=3	0,697	0,815	0,365	0,678	0,821	0,301
k=2	0,386	0,296	0,102	0,209	0,332	0,083
k=1	0,032	0,057	0,001	0,053	0,105	0,000

Wnioski:

- 1/ głębokość cytowania ma zasadniczy wpływ na wartości obliczanych współczynników podobieństwa /por. wiersze dla $k=4$ i $k=1$ /,
- 2/ o ile dla współczynnika ważonego obserwuje się zjawisko "nasyca-
nia", tzn. zwiększanie głębokości cytowania powyżej $k=3$ mają w zasadzie marginalne znaczenie, o tyle dla pozostałych współ-
czynników podobieństwa takiego efektu nie zaobserwowano.

Szczególnie wrażliwy jest współczynnik grafowy - ponadto ma tę wadę, że wartości podobieństw w większości przypadków rosły z maleniem k .

- 3/ z maleniem k zaobserwowano tendencję do grupowania się wartości podobieństw wokół kilku liczb $/0, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots /$ - szczególnie dla podobieństw: grafowego i k -cytowań nieważonych,
- 4/ widać, że w pewnym sensie "optymalna" długość łańcucha cytowań jest różna dla różnych definicji podobieństwa /łańcuch nie może być za długi, bo rośnie złożoność obliczeniowa, ani za krótki bowiem zniekształcają się wartości podobieństw - precyzyjniejsze określenie tego faktu wymagałoby przeprowadzenia bardzo kosztownych obliczeń na dostatecznie dużym zbiorze danych/.

Ponadto wydaje się, że pożądana długość łańcucha cytowań jest różna dla różnych dziedzin /wydłuża się dla dziedzin o szybkim rozwoju lub stagnacji/.

4.5. Analiza wyników zastosowania algorytmów grupowania

Celem eksperymentu było wykazanie, że na podstawie proponowanych funkcji podobieństw /DF 4.2.1.1, 4.2.1.3 ÷ 4/ można w sposób automatyczny wydzielić monotematyczne grupy obiektów przy pomocy grafowych algorytmów podziału. Zbadano ponadto wpływ szeregu parametrów na wynik podziału.

Eksperyment składał się z dwóch zasadniczych części. Najpierw poddano grupowaniu obiekty z lat 1976 ÷ 1977 na podstawie podobieństwa ich opisów. Natomiast zbiór prac z lat 1978-80 służył jako zbiór aktualizujący otrzymany podział. W drugim etapie poddano grupowaniu prace z lat 1976 - 80 w celu zbadania wpływu strategii aktualizacji na stabilność podziału. Zależy bowiem od niej częstość dokonywania całkowitej reorganizacji b.d.

Przetestowano wpływ na wyniki grupowania różnych:

- A - definicji podobieństwa i typów redukcji /p. 4.4.A/ opisów obiektów,
- B - głębokość cytowań Ω^n ,
- C - stopnia k -spójności w metodach grafowych.

Sprawdzono również dynamikę zmian podziałów wynikowych /D/.

A. Wpływ różnych definicji podobieństw i redukcji opisów obiektów

Wyniki grupowań oceniano na podstawie grupowań wzorcowych ekspertów Exp I, Exp II. Grupy wzorcowe to; odpowiednio zredukowane /lata, wykaz czasopism/ bibliografie przeglądowych prac ekspertów Exp II oraz wynik grupowania Exp I. Podziały dokonane przez ekspertów są podziałami skąbymi oraz na różnych zbiorach obiektów.

Tabela 6 zawiera wyniki ocen grupowań dla różnych grafowych algorytmów grupowania /opisanych w p. 3/, różnych definicji funkcji podobieństwa /p. 4.2.1/, różnych typów redukcji opisów.

Wnioski:

- 1/ współczynnik podobieństwa $p_r(n=1)$ /propozycja Eatona [12] nie nadaje się do wykorzystywania w automatycznym grupowaniu - daje bardzo dużo drobnych grup /z wyjątkiem metody α^* /, co powoduje, że p^* przyjmuje duże wartości. Dodanie monografii niewiele poprawia wyniki.
- 2/ współczynnik grafowy p_g daje dobre wyniki dla metod ω^* , ω^* , MPP-Gotlieb. Dla innych daje mało, zbyt licznych grup. Ponadto daje zaskakujące grupy wynikowe i wyniki wrażliwe są na redukcję opisów,
- 3/ współczynnik ważony p daje dobre wyniki prawie dla wszystkich metod. Wyniki są dobre dla różnych redukcji danych. Być może jest to wpływ przyjęcia dużej głębokości cytowań / $n = 4$ /. Wymaga to szczegółowego zbadania /TABELA 7/.
- 4/ w metodzie MPP-Bonnera uśredniono wyniki trzech grupowań /różny punkt startu/, dla częściowego uniezależnienia się od efektu kolejnościowego. Mimo tego wyniki są bardzo różne.
- 5/ dla metody MPP-Gotlieba najlepszy podział osiągnięto dla wartości parametru obcięcia $t = 0,4$ /a nie $t = 0,6$ - jak polecają autorzy przy grupowaniu słów kluczowych [45] /,
- 6/ metody ω^* i ω^* nie dają dobrych wyników w grupowaniu dokumentów - dają wiele, małowlicznych grup /szczególnie dla wysokich progów obcięcia/; ponadto grupy znacznie różnią się rozmiarami,
- 7/ ze zwiększeniem głębokości cytowań, dla uzyskania lepszych wyników należy stosować metody wykorzystujące pojęcie k-spójności / $k \geq 2$ /,
- 8/ w miarę wzrostu gęstości grafu podobieństwa coraz skuteczniejsze są metody bazujące na MPP /za wyjątkiem metody Bonnera/,

Podobieństwo	Wzorzec	Algorytm Miara	α^* -Kruskal		$1/\alpha^*_{k=2}$		$\mathcal{J}^*_{k=2}$		$\omega^*_{k=10}$		ω^*_n	
			b.d.I	b.d.II	b.d.I	b.d.II	b.d.I	b.d.II	b.d.I	b.d.II	b.d.I	b.d.II
$p_r(n=1)$ DF 4.2.1.1	Exp. I 76÷77	p^*	0,301	0,21	0,25	0,203	0,295	0,208	0,296	0,201	0,296	0,201
		Ψ	0,093	0,062	0,073	0,053	0,086	0,059	0,066	0,055	0,061	0,048
	Exp. II 78÷80	p^*	0,313	0,225	0,263	0,211	0,3	0,221	0,283	0,204	0,283	0,204
		Ψ	0,091	0,072	0,071	0,056	0,083	0,061	0,059	0,045	0,060	0,048
$p_g(n=4)$ DF 4.2.1.3	Exp. I 76÷77	p^*	0,51	0,49	0,52	0,503	0,65	0,68	0,412	0,397	0,383	0,38
		Ψ	0,421	0,42	0,43	0,435	0,444	0,445	0,41	0,401	0,385	0,371
	Exp. II 78÷80	p^*	0,495	0,48	0,511	0,515	0,583	0,601	0,4	0,382	0,37	0,371
		Ψ	0,401	0,411	0,386	0,406	0,42	0,422	0,385	0,361	0,374	0,362
$p(n \leq 5)$ DF 4.2.1.4	Exp. I 76÷77	p^*	0,632	0,691	0,788	0,803	0,724	0,738	0,461	0,453	0,487	0,47
		Ψ	0,603	0,621	0,743	0,772	0,711	0,721	0,537	0,542	0,532	0,502
	Exp. II 78÷80	p^*	0,609	0,689	0,789	0,811	0,704	0,739	0,442	0,466	0,493	0,478
		Ψ	0,591	0,602	0,741	0,781	0,705	0,719	0,539	0,542	0,54	0,491

TABELA 6. Zbiorcze zestawienie wyników eksperymentów

TABLEA 6 /c.d.d./

1/ metody α^* , α_2^* , δ_2^* , ω_{10}^* dla t-ekstrakcji grafu / $t = 0,1$

2/ uśrednione dla trzech przebiegów / różne punkty startu/

Podobieństwo	Wzorzec	Algorytm Miara	MPP-Gottlieb ($t=0,4$)		MPP-Bonner ^{2/}		MDS - Zahn	
			b.d. I	b.d. II	b.d. I	b.d. II	b.d. I	b.d. II
$p_r(n=1)$ DF 4.2.1.1	Exp. I 76÷77	p	0,296	0,201	0,296	0,201	0,291	0,197
		γ	0,068	0,050	0,068	0,043	0,091	0,059
	Exp. II 78÷80	p	0,285	0,207	0,282	0,204	0,302	0,2
		γ	0,058	0,047	0,052	0,036	0,09	0,064
$p_g(n=4)$ DF 4.2.1.3	Exp. I 76÷77	p	0,441	0,416	0,396	0,395	0,597	0,601
		γ	0,451	0,402	0,382	0,409	0,503	0,522
	Exp. II 78÷80	p	0,427	0,421	0,301	0,298	0,566	0,567
		γ	0,399	0,398	0,208	0,331	0,4	0,423
$p(n \leq 5)$ DF 4.2.1.4	Exp. I 76÷77	p	0,588	0,561	0,486	0,498	0,666	0,672
		γ	0,578	0,566	0,405	0,613	0,643	0,608
	Exp. II 78÷80	p	0,601	0,572	0,501	0,472	0,652	0,639
		γ	0,589	0,568	0,451	0,503	0,645	0,591

TABELA 7. Jakość uzyskanych podziałów przy zmianie głębokości cytowań

Podobieństwo	n	Algorytm	α - Kruskal		α_2		$\mathcal{J}_2$		MPP-Gottlieb /t = 0,4/	
			Miara		b.d.I	b.d.II	b.d.I	b.d.II	b.d.I	b.d.II
				b.d.I	b.d.II	b.d.I	b.d.II	b.d.I	b.d.II	b.d.I
$p_r(n)$	1	Ψ	0,091	0,072	0,071	0,056	0,083	0,061	0,058	0,047
		p°	0,313	0,325	0,263	0,311	0,3	0,321	0,285	0,307
	2	Ψ	0,210	0,113	0,102	0,077	0,194	0,098	0,081	0,063
		p°	0,2	0,222	0,253	0,266	0,22	0,25	0,266	0,269
	3	Ψ	0,501	0,475	0,381	0,303	0,482	0,421	0,221	0,139
		p°	0,492	0,412	0,3	0,301	0,352	0,383	0,23	0,2
	4	Ψ	0,463	0,548	0,476	0,385	0,513	0,601	0,478	0,352
		p°	0,413	0,52	0,421	0,365	0,49	0,51	0,37	0,35
$p(n)$	1	Ψ	0,084	0,056	0,066	0,053	0,078	0,055	0,052	0,042
		p°	0,3	0,322	0,254	0,32	0,283	0,299	0,331	0,387
	2	Ψ	0,152	0,11	0,098	0,078	0,112	0,09	0,076	0,048
		p°	0,19	0,25	0,26	0,302	0,22	0,26	0,283	0,321
	3	Ψ	0,538	0,441	0,404	0,3	0,521	0,519	0,31	0,28
		p°	0,385	0,366	0,386	0,4	0,48	0,511	0,341	0,325
	4	Ψ	0,585	0,552	0,653	0,697	0,688	0,702	0,486	0,421
		p°	0,502	0,52	0,599	0,605	0,6	0,64	0,45	0,382
	5	Ψ	0,591	0,602	0,741	0,781	0,705	0,719	0,589	0,568
		p°	0,609	0,689	0,789	0,811	0,704	0,739	0,601	0,572

- 9/ generalnie: dla grupowań b.d. I uzyskiwano większą pełność i średnią precyzję, a dla grupowań b.d. II - większą precyzję a średnią pełność odpowiedzi.
- 10/ dodanie monografii polepsza wyniki dla $p_r/n=1/$ bądź dla metod ω_k , natomiast pogarsza dla $p(n \leq 5)$ i $p_s(n=4)$ dla niektórych metod $/\alpha^*, \alpha_2^*, \delta_2^*, \text{MPP-Bonner}/$,
- 11/ z TABELI 7 ponadto widać, że gdy uwzględniamy dużą głębokość cytowań i większy stopień spójności to usunięcie monografii z prac cytowanych polepsza wyniki grupowania, natomiast dodanie monografii jest konieczne przy głębokościach mniejszych $/n \leq 3/$ - jednak poprawa wyników nie jest znaczna.

B. Wpływ głębokości cytowań Ω^n

Jak już częściowo widać z TABELI 6 jakość grupowań zależy od uwzględnianej /przy obliczaniu współczynnika podobieństwa/ głębokości cytowań Ω^n . W Tabeli 7 pokazano ten wpływ dla współczynnika $p_r(n)$ i $p(n)$ przy ustalonych algorytmach: $\alpha, \alpha_2, \delta_2, \text{MPP-Gotlieb}$. Danymi były prace lat 1978÷80 wraz z opisami /ocena wg Exp. II/.

Wnioski:

- 1/ widać wyraźnie wzrost jakości podziału przy zwiększaniu głębokości cytowań /szczególnie duże przyrosty dla $n = 2,3/$,
- 2/ wartość Ψ dla współczynnika $p(n)$ rośnie monotonicznie ze wzrostem n , przy czym przyrosty Ψ są coraz mniejsze /zjawisko nasycania/,
- 3/ dla współczynnika $p_r(n)$ występuje wyraźne ekstremum zgodności /różne dla różnych algorytmów/ przy czym ma ono niższą wartość niż najlepsze wartości dla $p(n)$. Ekstremum to występuje dla różnych głębokości /w zależności od metody grupowania/,
- 4/ dla rozpatrywanych współczynników podobieństwa głębokość cytowań równa 2 nie jest wystarczająca dla celów grupowania,
- 5/ im większa głębokość cytowań tym bardziej precyzyjną metodę należy stosować przy grupowaniu obiektów /u nas: dla $n=3$ i I typu redukcji danych najlepsze wyniki daje α^* , ale dla $n=4$ najlepsza jest metoda δ_2^* - II typ redukcji/ oraz stosować ostrzejszą redukcję opisów,
- 6/ grupowania wykorzystujące współczynnik $p(n)$, $n > 3$ są lepsze od odpowiednich grupowań dla $p_r(n)$.

C. Wpływ stopnia k-spójności w grafowych metodach grupowania

Jak widać z TABELI 6 i 7, w celu osiągnięcia dobrych wyników ze wzrostem n /od pewnej wartości progowej, u nas równej 3/ należy zwiększać k lub stosować ostrzejszą redukcję danych. Każda z metod k -spójnych osiąga najlepszy podział dla ściśle określonego n_{opt} /przy ustalonym współczynniku podobieństwa i typie redukcji/, przy czym ze względu na posiadane dane, nie można było wyznaczyć n_{opt} dla $k > 2$. Wynioskowana zależność nie jest ścisła. Precyzyjniejsze jej wyznaczenie jest pracochłonne i niecelowe /dla innych dziedzin prawdopodobnie n_{opt} będzie miało inną wartość/.

Ważne jest, że dobre podziały można uzyskiwać grafowymi metodami k -spójnościowymi przy niskim k /ważne ze względu na złożoność obliczeniową/.

D. Dynamika zmian podziałów wynikowych

Dla sprawdzenia, jak zmieniają się podziały w czasie, przeprowadzono eksperyment polegający na:

- a/ pogrupowaniu prac z lat 1976 ÷ 77 /metoda α_2^* , głębokość cytowań $n \leq 5$, współczynnik podobieństwa - $p(n)$, różne typy selekcji - najlepsza zgodność z ocenami ekspertów/,
- b/ chronologicznej aktualizacji otrzymanego podziału przez prace z lat 1978 ÷ 1980 - na zasadzie 1-NN,
- c/ porównaniu zaktualizowanego podziału z podziałem /jak w punkcie a/ / prac z lat 1978 ÷ 1980 przy pomocy miar p^* i ψ .

Wyniki zawarto w TABELI 8.

TABELA 8. Zmiany podziału przy aktualizacji

Typ redukcji	b.d. I	b.d. II
p^*	0,802	0,835
ψ	0,873	0,881

Okazuje się, że zgodność jest bardzo duża a wybraliśmy do badań dziedzinę /s.w.i./ o szybkim rozwoju w ostatnich latach.

Upoważnia to do sformułowania następującego wniosku:

- przy wymienionych wyżej założeniach powtórnego dzielenia b.d.

nie trzeba dokonywać częściej niż co 4 ÷ 5 lat. — ○

Niewątpliwie, gdyby zaproponować lepszą regułę aktualizacyjną /p-kt b/, można wydłużyć okresy między reorganizacjami b.d.

Reguła 1-NN jest za sztywna /podziały b.d. są podziałami słabymi/, natomiast reguła 2-NN /dająca podobne wyniki jak 1-NN/ za bardzo zwiększa stopień przecięcia między grupami. Być może, najstosowniejsza byłaby reguła oparta na progowym podobieństwie obiektu do grupy, bądź granicznej różnicy między uporządkowanymi podobieństwami do grup. Zagadnienie to wymaga dalszych badań.

E. Porównanie selektywności podziałów

Przy grupowaniu obiektów na podstawie ich cytowań okazało się, że zbiory słów kluczowych z charakterystyki bibliograficznej dokumentów z wyodrębnionych grup przekrywały się w dużym stopniu. Świadczy to o tym, że:

- większość wyodrębnionych grup byłaby nierozróżnialna w s.w.i. z tezauresem deskryptorów /gorsza selektywność i precyzja odpowiedzi/,
- różne wydawnictwa stosują różne reguły indeksowania /tezaurus, głębokość indeksowania, liczba słów kluczowych/,
- część deskryptorów zmienilo swój zakres pojęciowy, a także uległo rozwojowi słownictwo.

To zagadnienie również wymaga dalszych, precyzyjniejszych badań.

Tym niemniej wymieniony zestaw słów kluczowych stanowi tematyczną charakterystykę grupy uzyskaną za darmo, przy okazji dzielenia b.d. Charakterystykę tę można wykorzystać do obsługi pytań, które nie mają formy zestawu bibliografii.

5. ZAKOŃCZENIE

W przedstawionej pracy autor starał się rozwinąć pomysł podziału tematycznego bazy danych dokumentacyjnego systemu informacyjnego na gruncie teorii grafów z wykorzystywaniem struktury cytowań dla obliczania podobieństwa między obiektami. Stwierdzenie możliwości automatycznego wyodrębniania adekwatnych, tematycznych grup obiektów miałoby duże znaczenie dla polepszenia parametrów

eksploatacyjnych s.w.i. Przy takim podejściu można by wyeliminować bądź ograniczyć manualne indeksowanie dokumentów deskryptorami /słowami kluczowymi, symbolami klasyfikacyjnymi, itp./. Ponadto, przy takiej grupowej strukturze danych skróceniu uległby czas wyszukiwania informacji z b.d., ponieważ pytanie porównywane byłoby tylko z reprezentantami grup. Aby propozycja wykorzystania cytowań i algorytmów grafowych miała sens praktyczny trzeba, by automatycznie wydzielane grupy były stabilne w wystarczająco długim okresie czasu. Zapewnia to, że reorganizacja /powtórne grupowania/ nie będą zbyt częste. Wymagana jest również mała pracochłonność dokonywania podziału. Metody grafowe przy niskim stopniu spójności k gwarantują, że złożoność obliczeniowa będzie wielomianowo ograniczona od góry [140].

Należy zaznaczyć, że ze względu na przyjętą formę opisu obiektu, heurystyczne metody grupowania dokumentów opisanych ciągiem słów kluczowych okazały się w tym przypadku nieprzydatne. Cytowania, co prawda wykorzystywano wcześniej /m.in. w pracach [2, 6, 43, 44, 114, 128, 139, 143, 144] / lecz głównie w celach naukometrycznych, socjologicznych lub bibliograficznych. Próby przeprowadzone przez autora i przedstawione w rozdziale 4 wykazały, że cytowania mogą być również przydatne w wyszukiwaniu informacji.

W szczególności eksperymenty te potwierdzają, że:

- a/ celowym jest uprzednie wyodrębnianie grup tematycznych wydawnictw. Można to łatwo wykonać na podstawie ich wzajemnych cytowań,
- b/ stosowanie współczynnika k-retrospektywnego w grupowaniu jest niecelowe, bowiem lepsze wyniki osiąga się przy stosowaniu współczynnika ważonego,
- c/ zastosowane redukcje opisów obiektów nie pogarszają /a często polepszają/ wyniki grupowania a pozwalają znacznie zmniejszyć wymaganą pamięć i pracochłonność obliczania podobieństw. Redukcje te nie zmierzają w rażący sposób własności bibliometrycznych b.d.,
- d/ zasadniczą sprawą jest wybór najodpowiedniejszej głębokości grafu cytowań tzn. głębokości odpowiadającej maksimum współczynnika Ψ . Dla różnych metod grupowania i różnych wag jest ona różna,
- e/ im mniejszą wagę przypisujemy najstarszym pracom cytowanym

- /przy odpowiedniej głębokości cytowań/ tym niższy stopień spójności k można przyjmować w grafowych algorytmach podziału,
- f/ metody grafowe o dużym stopniu spójności, mimo dużej złożoności obliczeniowej, nie dają z reguły zadowalających wyników /chyba, że pytanie dotyczyłoby prac bardzo silnie związanych/,
 - g/ zmiany grupowania w czasie nie są duże /mimo przyrostu nowych informacji/, tzn. nie trzeba często dokonywać reorganizacji b.d.
 - h/ stosując przyjęty sposób opisu obiektów osiągamy lepszą selektywność s.w.i.

Powyższe fakty są potwierdzeniem tezy pracy, że przy pomocy grafowych metod grupowania oraz korzystając ze zdefiniowanej funkcji podobieństwa między obiektami można adekwatnie podzielić dokumentacyjną bazę danych na grupy tematyczne. Podział taki może mieć praktyczne zastosowanie przy usprawnieniu procesu indeksowania obiektów i ich wyszukiwaniu w s.i.d. Tego typu organizacja b.d. jest bardzo naturalna dla sposobu zadawania pytań, tym bardziej, że nawet pojedyncze błędne zaklasyfikowania nie mają praktycznie wpływu na jakość działania systemu informacyjnego takiego typu ponieważ przeciętny "czas życia" artykułu naukowo-technicznego szacuje się na $3 \div 8$ lat.

Oczywiście, dokonanie podziału b.d. musi być usprawiedliwione warunkami w jakich działa s.w.i. ponieważ komplikują się czynności przed wyszukiwaniem. Generalnie można stwierdzić, że grupowa struktura danych będzie nieefektywna dla b.d. o małej liczbie danych, małej liczbie wyszukiwań lub bardzo dużej częstości zmian zawartych informacji [109, 116, 145]. Dokumentacyjne systemy informacyjne tych cech z reguły nie posiadają [106, 131, 139].

Wydaje się, że zagadnienia dotyczące przedmiotu pracy nasuwają szereg dalszych problemów badawczych. W szczególności należałoby zbadać kwestie odnoszące się do struktury i wykorzystania cytowań naukowych, a mianowicie:

- 1/ czy osiągnięte wnioski są słuszne dla b.d. o innej tematyce,
- 2/ jak zależą wyniki podziału od doboru wag oraz innych sposobów redukcji opisów - być może wystarczy obliczać podobieństwa tylko na podstawie cytowanych czasopism bądź nazwisk autorów,
- 3/ jakie prawidłowości mogą ujawnić się przy korzystaniu z perspektywnego współczynnika podobieństwa obiektów.

Zastosowania praktyczne wymagają przeanalizowania i rozwiąza-

zania następujących problemów:

- 1/ wyboru najlepszej strategii aktualizacji grup,
- 2/ wyboru najlepszej strategii wyszukiwania informacji w grupowej organizacji b.d.,
- 3/ określania momentu czasu, w którym reorganizacja b.d. staje się niezbędna,
- 4/ sposobu "strojenia" pytań informacyjnych,
- 5/ określenia najlepszej organizacji wyszukiwań bieżących i retrospektywnych /być może należy stosować również grupowanie pytań/.

Odnosnie badań teoretycznych nasuwają się następujące propozycje:

- 1/ badania tych własności współczynników podobieństwa, które są istotne dla procesu grupowania na podstawie np. funkcji monotoniczności podobieństw,
- 2/ dalsza analiza algorytmów i polepszenie oszacowań parametrów grupowań uzyskiwanych algorytmami grafowymi - być może na gruncie grafów losowych.

Autor pragnie wyrazić wdzięczność promotorowi prof. dr hab. Adamowi Sielickiemu za stałą opiekę merytoryczną, zachętę do kontynuacji tematu oraz cenne propozycje i uwagi. Wszystko to pozwoliło nadać pracy jej obecną formę i treść.

6. BIBLIOGRAFIA

- [1] Anderberg M.R.: Cluster analysis for applications, Acad. Press, New York, 1973, 353p. / 160 ref. /
- [2] Arms W.Y., Arms C.R.: Cluster analysis used on social science journal citations, J. Docum., 1978, vol. 34, no. 1, p. 1-11
- [3] Arnovick G.N., Gee L.G.: Design and evaluation of information systems, Inf. Proc. Manage., 1978, vol. 14, no. 6, p. 369-380
- [4] Augustson J.G., Minker J.: An analysis of some graph theoretical cluster techniques, J. ACM, 1970, vol. 17, no. 4, p. 571-588
- [5] Bartosiewicz S.: Ekonometria, PWE, 1978, 231 s.
- [6] Bekisz S.: Das Klassifizieren von Dokumenten auf Grundlage der bibliographischen Analysen, XI koll. Inf. Dok., 1979, heft 45, s. 261-272
- [7] Bellmore M., Neumhauser G.L.: The travelling salesman problem: a survey, Oper. Res., 1968, vol. 16, no. 3, p. 538-558
- [8] Bentley J.L., Friedman J.H.: Fast algorithms for constructing minimal spanning trees in coordinate spaces, IEEE Comp., 1978, no. 2, p. 97-105
- [9] Berman G., Colijn A.W.: A modified list technique allowing binary search, J. ACM, vol. 21, 1974, no. 2, p. 227-233
- [10] Berztiss A.T.: Data structures; theory and practice, Acad. Press, 1971, 442p.
- [11] Bhat M.V., Haupt A.: An efficient clustering algorithm, IEEE Syst. Man., 1976, vol. SMC-6, no. 1, p. 61-64
- [12] Bichteler J., Eaton E.A.: Comparing two algorithms for document retrieval using citation links, JASIS, 1977, vol. 28, no. 4, p. 192-195
- [13] Bichteler J., Parsons R.G.: Document retrieval by means of an automatic classification algorithm for citations, Inf. Stor. Retr., 1974, no. 7/8
- [14] Bonner R.E.: On some clustering techniques, IBM J., 1964, vol. 22, no. 1, p. 22-32
- [15] Bookstein A.: On the perils of merging boolean and weighted retrieval systems, JASIS, 1978, vol. 29, no. 3, p. 156-158
- [16] Bron C., Kerbosch J.: Finding all cliques of an undirected graph, CACM, 1973, vol. 16, no. 9, p. 575-577
- [17] Brookes B.C.: The Shannon model of IR systems, J. Docum., 1972, vol. 28, p. 160-162
- [18] Cheriton D., Tarjan R.E.: Finding minimum spanning trees, SIAM J. Comp., 1976, vol. 5, no. 4, p. 724-742
- [19] Chin F., Houck D.: Algorithms for updating minimal spanning trees, J. Comp. Sys. Sci., 1978, vol. 16, no. 4, p. 333-344
- [20] Cooper W.S., Maron M.E.: Foundations of probabilistic and utility-theoretic indexing, JASIS, 1978, vol. 25, no. 1, p. 67-80
- [21] Corneil D.G., Woodward M.E.: A comparison and evaluation of graph theoretical clustering techniques, INFOR, 1978, vol. 16, no. 1, p. 74-89
- [22] Croft W.B.: Clustering large files of documents using the single-link method, JASIS, 1977, vol. 28, no. 6, p. 341-344
- [23] Croft W.B.: Theoretical model of clustering searching, J. Inf., 1978, vol. 2, no. 1, p. 78-81
- [24] Crouch D.B.: A file organization and maintenance procedure for dynamic document collections, Inf. Proc. Manage., 1975, vol. 11, no. 1, p. 11-21
- [25] Czaja L.: Heurystyczne lokowanie macierzy rzadkich, Prace CO PAN, 1973, nr 115, 13s.
- [26] Dąbrowski M., Laus K.: A model of information retrieval process for hierarchical set of descriptors, Inf. Stor. Retr., 1974, vol. 10, no. 7/8, p. 261-265
- [27] Dąbrowski M., Laus K.: Metody wyszukiwania i klasyfikacji informacji, WNT, Warszawa, 1978, 140s. / bibl. 116 poz. /
- [28] Deo N.: Teoria grafów i jej zastosowania w technice i informatyce, PWN, 1980, Warszawa, 607s.
- [29] Dobosiewicz Wł.: Sorting by distributive partitioning, Inf. Proc. Letters, 1978, vol. 7, no. 1, p. 1-6
- [30] Dobrov G.M., Korennoj A.A.: Nauka; informacija i upravlenije, Sov. Radio, 1977, Moskva, 256s.
- [31] Duran B., Odell P.: Klasternyj analiz, Statistika, Moskva, 1977, 128s. / 409 ref. /

- [32] Ehrlich H.D., Lipski W.: On the storage space requirement of consecutive retrieval with redundancy, *Inf. Proc. Lett.*, 1976, vol. 5, no. 4, p. 101-104
- [33] Elchesen D.R.: Cost-effectiveness comparison of manual and on-line retrospective bibliographic searching, *JASIS*, 1978, vol. 29, no. 2, p. 56-67
- [34] Fisher M.L., Neumaier G.L., Wolsey L.A.: An analysis of approximations for finding a maximum weight Hamiltonian circuit, *Oper. Res.*, 1979, vol. 27, no. 4, p. 799-809
- [35] Floyd R.W., Rivest R.T.: Expected time bounds for selection /Algorithm 489/, *CACM*, 1975, vol. 18, no. 3, p. 165-173
- [36] Flynn T., Holohan P.A., Magson M.S., Munro J.D.: Cost effectiveness comparison of on-line and manual bibliographic information retrieval, *J. Inf. Sci. Princ. Practice*, 1979, vol. 1, no. 2, p. 77-84
- [37] Fredman M.L.: New bounds on the complexity of the shortest path problem, *SIAM J. Comp.*, 1976, vol. 5, no. 1, 83-89
- [38] Frencl Z.: Routing problem/Algorithm 456/, *CACM*, 1973, vol. 16, no. 9, p. 572-574
- [39] Fritzsche M.: Automatic clustering techniques in information retrieval, *Lux. Comm. of Europ. Communities*, 1974, 145p.
- [40] Fromm F.R., Northouse R.A.: CLASS-a nonparametric clustering algorithm, *Patt. Recogn.* 1976, vol. 8, no. 2, p. 107-114
- [41] Fussengger F., Gabow H.N.: A counting approach to lower bounds for selection problems, *JACM*, 1979, vol. 26, no. 2, p. 227-238
- [42] Garey M.R., Johnson D.S., Tarjan R.E.: The planar Hamiltonian circuit problem is NP-complete, *SIAM J. Comp.*, 1976, vol. 5, no. 4, p. 704-719
- [43] Garfield E.: Citation analysis as a tool in journal evaluation, *Science*, 1972, no. 178 /4060/, p. 471-479
- [44] Garfield E., Sher I.H.: New factors in the evaluation of scientific literature through citation indexing, *Am. Doc.*, 1963, vol. 14, no. 2, p. 195-201
- [45] Gotlieb C.C., Kumar S.: Semantic clustering of index terms, *J. ACM*, 1968, vol. 15, no. 4, p. 493-513
- [46] Gowda K.Ch., Krishna G.: Agglomerative clustering using the concept of mutual nearest neighbourhood, *Patt. Recogn.*, 1978, vol. 10, no. 1, p. 105-112
- [47] Gubankov V.N.: O niekotorykh vremennyykh sootnosheniyakh v zakonomiernostiakh izmeneniya količestva citirujemykh naučnykh dokumentov, *NTI*, 1975, ser. 2, nr 3, s. 3-6
- [48] Hahn B.: A new technique for compression and storage of data, *CACM*, 1974, vol. 17, no. 8, p. 434-436
- [49] Hawkins D.T.: Multiple database searching: techniques and pitfalls, *Online*, 1978, vol. 2, no. 2, p. 9-15
- [50] Heine M.H.: The Swets theory, lect. to Wrocław Tech. Univ., July 1975
- [51] Hubert L.J., Baker F.B.: r-diameter hierarchical clustering, in: *Classification and clustering*, ed. J. Van Ryzin, Acad. Press, 1977, New York, 467p.
- [52] Informacja w badaniach naukowych. Praca zbiorowa, WKiŁ, Warszawa, 1972, 275s.
- [53] Jaegerman M.: Information storage and retrieval systems-mathematical foundations. Part IV-systems with incomplete information, *CC PAS Rep.*, 1975, no. 215, 78p.
- [54] Jardine N., Rijsbergen C.J.: The use of hierarchic clustering in information retrieval, *Inf. Stor. Retr.*, 1971, vol. 7, no. 3, p. 217-240
- [55] Jensen R.E.: A dynamic programming algorithm for cluster analysis, *Oper. Res.* 1969, vol. 12, no. 6, p. 1034-1057
- [56] Johnson D.B.: Priority queues with update and finding minimum spanning trees, *Inf. Proc. Letters*, 1975, vol. 4, no. 1, p. 53-57
- [57] Jokkina V.N., Zagorujko N.G.: Kaličestviennyje kriterii kačestva taksonomii i ich ispolzovanie v processie prinjatia rešenij, *Nauka*, 1969, Novosibirsk, sb. Vyčislitelnyje Sistemy, vyp. 36
- [58] Kariv O., Hakimi S.L.: An algorithmic approach to network location problems. I-the p-centers, II-the p-medians, *SIAM J. Appl. Math.*, 1979, vol. 37, no. 3, p. 513-538, 539-560
- [59] Knuth D.: The art of computer programming, Addison-Wesley, 1973, vol. 3 /sorting and searching/. Takže: Knut D.: Iskusstvo programmirovania, 1978, Mir, t. 3/sortirovka i poisk/, 845 s.

- [60] Kofman A.:Vviedienie v prikladnuju kombinatoriku,Nauka,1975,Moskva,479 s.
- [61] Konieczna D.:Badania w zakresie nieformalnych procesów komunikacji nauko-
wej,APID,1979,nr 4,s.16-20
- [62] Kozačkov L.C.:Informacionnyje sistiey s jererchičeskoj strukturoj,NTI,
1978,ser.2,nr 8,s.15-25
- [63] Kraft D.H.:A threshold rule applied to the retrieval decision model,JASIS,
1978,vol.29,no.2,p.77-80
- [64] Kraft D.H.:Information retrieval-a decision theory approach,JASIS,1973,
vol.24,no.5,p.368-376
- [65] Kraft D.H.,Bookstein A.:Evaluation of information retrieval systems:a de-
cision theory approach,JASIS,1978,vol.29,no.1,p.31-40
- [66] Kristofides H.:Teoria grafov-algoritmičeskij podchod,Mir,Moskva,1978,432s.
- [67] Kucharczyk J.:Metody analizy skupień i ich algorytmy w języku ALGOL 60,
PWN,1981,144 s./ w druku /
- [68] Kwok K.L.:The use of titles and cited titles as document representation
for automatic classification,Inf.Proc.Manage.,1975,vol.11,no.2,201-206
- [69] Lance G.N.,Williams W.T.:A general theory of classificatory sorting stra-
tegies-hierarchical systems,Comp.J.,1967,vol.9,no.4,p.373-380
- [70] Li Y.K.:Evaluating information storage and retrieval system-a decision
theory approach,Inf.Stor.Reptr.,1973,vol.9,no.5,p.281-293
- [71] Liou J.H.,Yao S.B.:Multi-dimensional clustering for data base organiza-
tions,Inform.Syst.,1977,vol.2,no.2,p.187-198
- [72] Lipski W.:Kombinatoryczne aspekty teorii wyszukiwania informacji,PWN,1975,
Warszawa,51 s.
- [73] Lipski W.:On strings containing all subsets as substrings,ICS PAS Rep.,
Warsaw,1977,no.282,15 p.
- [74] Loizou G.:Graph models of complex information sources,Inf.Proc.Manage.,
1979,no.3,vol.15,p.127-132
- [75] Lyon G.:Packed scatter tables,CACM,1978,vol.21,no.10,p.857-865
- [76] Mann S.H.:Least-cost decision rules for dynamic information management,
Inf.Stor.Reptr.,1971,vol.7,no.2,p.111-121
- [77] Marek W.,Pawlak Z.:Information storage and retrieval systems:mathematical
foundation,Theoret.Comp.Sci.,1976,no.4,p.331-355
- [78] Matula D.W.:Graph theoretic techniques for cluster analysis algorithms,in:
Classification and clustering,Acad.Press,1977,New York,467p./p.95-129/
- [79] Matula D.W.:k-components,clusters and slicings in graphs,SIAM J.Appl.Math.,
1972,vol.22,no.3,p.459-480
- [80] Mazur Z.:Properties of a model of information retrieval system based on
thesaurus with weight,Inf.Proc.Manage.,1979,vol.15,no.3,p.145-155
- [81] Mc Cormick W.T.,Schweitzer P.J.,White T.W.:Problem decomposition and data
by a clustering technique,Oper.Res.,1972,vol.20,p.993-1009
- [82] Michajłov A.I.,Černyj A.I.,Gilarevskij R.S.:Naučnyje kommunikacii i in-
formatika,Nauka,Moskva,1976,435 s.
- [83] Minieka E.:The centers and medians of a graph,Oper.Res.,1977,vol.25,p.641-50
- [84] Minker J.,Wilson G.A.,Zimmerman B.H.:An evaluation of query expansion by
the addition of clustered terms for a document retrieval system,Inf.
Stor.Reptr.,1972,vol.8,no.5,p.329-348
- [85] Nalimow W.W.,Muliczenko Z.M.:Naukometria,WNT,Warszawa,1971,s.61-110
- [86] Negoita C.V.,Flondor P.:On fuzziness in information retrieval,Int.J.Man-
-Mach.Study,1976,vol.8,no.6,p.711-716
- [87] Ogilvie J.C.,Olson C.L.:On the use of complete subgraphs in cluster analy-
sis,Inf.Proc.Letters,1972,vol.1,no.3,p.76-79
- [88] Oppenheim Ch.,Penn S.P.:Highly cited old papers and the reasons why they
continue to be cited,JASIS,1978,vol.29,no.5,p.225-231
- [89] Page R.L.:A minimal spanning tree clustering method / Algorithm 479/,CACM,
1974,vol.17,no.6,p.321-323
- [90] Paluar E.,Rahini M.A.,Robinson R.W.:Efficiency of a binary comparison sto-
rage technique,J.ACM,1974,vol.21,no.3,p.376-385
- [91] Parish C.C.M.:On the application of queueing theory to analysis on-line
computing systems,Comp.J.,1975,vol.18,no.2,p.117-121

- [92] Pawlak Z.: Mathematical foundations of information retrieval, CC PAS Rep., Warsaw, 1973, no. 101, 8 p. Information systems, CC PAS Rep., Warsaw, 1978, no. 338, 15 p.
- [93] Radecki T.: Fuzzy set theoretical approach to document retrieval, Inf. Proc. Manage., 1979, vol. 15, no. 5, p. 247-259
- [94] Radecki T.: Mathematical model of information retrieval system based on the concept of fuzzy thesaurus. II New approach to the problem of information system effectiveness evaluation, Inf. Proc. Manage., 1976, vol. 12, no. 5, p. 313-318, 319-326
- [95] Rand W.M.: Objective criteria for the evaluation of clustering methods, J. ASA, 1971, vol. 66, no. 336, p. 846-850
- [96] Reiser A.: A linear selection algorithm for sets of elements with weights, Inf. Proc. Letters, 1978, vol. 7, no. 3, p. 159-161
- [97] Reisinger L.: On fuzzy thesauri, Proc. Comp. Statist., Vienna, Physica-Verlag, 1974, p. 119-127
- [98] Riebrova M.P.: Avtomatičeskoje kłassificirovanije obiekto v avtomatizirovannyh informacionno-poiskovyh sistiemach, NTI, 1976, ser. 2, nr 12, s. 8-14
- [99] Rijsbergen C.J.: A theoretical basis for the use of co-occurrence data in information retrieval, J. Docum., 1977, vol. 33, no. 2, p. 106-120
- [000] Rijsbergen C.J., Sparck Jones K.: Test collections, J. Docum., 1976, vol. 32, no. 1, p. 62-75
- [001] Robertson S.E.: On the nature of fuzz: A diatribe, JASIS, 1978, vol. 29, no. 6, p. 304-307
- [002] Rohlf F.J.: A probabilistic minimum spanning tree algorithm, Inf. Proc. Lett., 1978, vol. 7, no. 1, p. 44-48
- [003] Rohlf F.J.: Hierarchical clustering using the minimum spanning tree /Algorithm 76/, Comp. J., 1973, vol. 16, no. 1, p. 93-95
- [004] Rubin F.: A search procedure for Hamiltonian paths and circuits, J. ACM, 1974, vol. 21, no. 4, p. 576-580
- [005] Sachs W.M.: An approach to associative retrieval through the theory of fuzzy sets, JASIS, 1976, no. 2, p. 85-87
- [006] Salton G.: Automatic information organization and retrieval, Mc Graw-Hill, New York, 1968
- [007] Salton G.: Mathematics and information retrieval, J. Docum., 1979, vol. 35, no. 1, p. 1-29
- [008] Salton G., Wong A., Yang C.S.: A vector space model for automatic indexing, CACM, 1975, vol. 18, no. 11, p. 613-620
- [009] Solton Dž.: Dinamičeski je bibliotečno-informacionny je sistiemy, Mir, Moskva, 1979, 557 s. / tčum. z ang. /
- [010] Schiminovich S.: Automatic classification and retrieval documents by means of a bibliographic pattern discovery algorithm, Inf. Stor. Retr., 1971, vol. 6, p. 417-435
- [011] Sgall P., Panenova J., Máchova H.S.: A linguistic approach to information retrieval-II, Inf. Proc. Manage., 1975, vol. 11, no. 5/7, p. 147-153
- [012] Sibson R.: SLINK-an optimally efficient algorithm for the single-link cluster method, Comp. J., 1973, vol. 16, no. 1, p. 30-34
- [013] Slagle J.R., Chang C.L., Heller S.R.: A clustering and date-reorganizing algorithm, IEEE Trans. Syst.-Man., 1975, vol. SMC-5, no. 1, p. 125-128
- [014] Small H.: Co-citation in the scientific literature: a new measure of the relationship between two documents, JASIS, 1973, vol. 24, no. 4, p. 265-269
- [015] Small H.G.: The relationship between citation indexing and word indexing; study of co-occurrences of title words and cited references, Proc. ASIS, 1974, vol. 10, p. 217-218
- [016] SMART-automatyczny system wyszukiwania informacji, WNT, Warszawa, 1975, 585 s.
- [017] Sobczak W., Malina W.: Metody selekcji informacji, WNT, Warszawa, 1978, 158 s.
- [018] Soergel D.: Mathematical analysis of documentation systems, Inf. Stor. Retr., 1963, vol. 3, p. 129-173 / 65 ref. /
- [019] Sparck Jones K.: Automatic indexing, J. Docum., 1974, vol. 30, no. 4, p. 393-432
- [020] Sparck Jones K., Jackson D.M.: Some experiments in the use of automatically obtained term clusters for retrieval. In: Mechanized information storage,

- retrieval and dissemination, IFIP Proc. 1967, North-Holland, 1968, p. 203-212
- [121] Sparck Jones K., Jackson D.M.: The use of automatically obtained keyword classification for information retrieval, Inf. Stor. Retr., 1970, vol. 5, no. 2, p. 175-201
 - [122] Sparck Jones K., Kay M.: Linguistics and information science, Acad. Press, New York, 1973, 244 p.
 - [123] Spira P.M., Pan A.: On finding and updating spanning trees and shortest paths, SIAM J. Comp., 1975, vol. 4, no. 4, p. 375-380
 - [124] Tague J.M., Farradane J.: Estimation and reliability of retrieval effectiveness measures, Inf. Proc. Manage., 1978, vol. 14, no. 1, p. 1-16
 - [125] Tahani V.: A fuzzy model of document retrieval system, Inf. Proc. Manage., 1976, vol. 12, no. 2, p. 177-187
 - [126] Tai K.Ch.: The tree-to-tree correction problem, JACM, 1979, vol. 26, no. 3, p. 422-33
 - [127] Tarjan R.E.: A new algorithm for finding weak components, Inf. Proc. Letters, 1974, vol. 3, no. 1, p. 13-15
 - [128] Vatter E.K.F.: O strukturie naučnych ssyžok, NTI, 1975, ser. 2, no. 5, s. 17-20
 - [129] Verhoeff J., Goffman W., Belzer F.: Inefficiency of use of boolean functions for information retrieval systems, CACM, 1961, vol. 4, no. 6, p. 557-558
 - [130] Vinod H.D.: Integer programming and the theory of grouping, J. Am. Stat. Assoc., 1969, vol. 64, no. 6, p. 506-519
 - [131] Voprosy informacionnoj teorii i praktiki, VINITI, Moskva, 1975, 513 s.
 - [132] Waksman A., Green M.W.: On the consecutive retrieval property in file organization, IEEE Trans. Comp., 1974, no. 2, p. 173-174
 - [133] Ward H.: Hierarchical grouping to optimize an objective function, J. Am. Stat. Assoc., 1963, vol. 58, no. 301, p. 236-244
 - [134] Whitney V.K.M.: Minimal spanning tree/Alg. 422/, CACM, 1972, vol. 15, no. 4, p. 273-4
 - [135] Williams P.W.: Criteria for choosing subsets to obtain maximum relative entropy, Comp. J., 1978, vol. 21, no. 1, p. 57-62
 - [136] Wnuk W.: Algorytm tworzenia monotematycznych grup dokumentów wykorzystujący cytowania bibliograficzne, Kom. Bibl. Gł. i OINT PWr., nr. 67, ser. A, 1977, Wrocław, 20 s.
 - [137] Wnuk W.: Algorytmy podziału w zastosowaniu do bazy danych dokumentacyjnego systemu informacyjnego, Rap. Inst. Cybern. PWr., Wrocław, 1979, ser. Preprinty, nr 8, 49 s. /37 ref./
 - [138] Wnuk W.: Grupowanie obiektów - analiza zagadnienia, metody i algorytmy grupowania, Praca magist. Inst. Mat. UWr., Wrocław, 1981, 55 s. /niepublikowana/
 - [139] Wnuk W.: Celowość realizacji systemu informacyjnego wykorzystującego cytowania naukowe, Kom. BGŁ. i OINT PWr., Wrocław, 1977, ser. A, nr 70, 18 s.
 - [140] Wnuk W.: Charakterystyka grafowych metod grupowania obiektów, Rap. Inst. Cyb. PWr., ser. Preprinty, nr 25, Wrocław, 1981, 52 s.
 - [141] Wnuk W.: Die Bewertung von Informationsrecherchesystemen, XI Koll. Inf. Dok., 1979, heft 45, s. 102-115
 - [142] Wnuk W.: Eksploatacja posładowatelno-serijnej bazy danych, Proc. μP'79, Budapest, 1979, vol. 2, s. 595-602
 - [143] Wnuk W.: Optymalizacja selekcji wydawnictw ciągłych, Prace Nauk. BGŁ. i OINT PWr., ser. Studia i materiały, nr 2, Wrocław, 1977, s. 17-23
 - [144] Wnuk W., Bekisz S.: Możliwości wykorzystania indeksów cytowanego piśmiennictwa, Prace Nauk. BGŁ. i OINT PWr., 1977, Wrocław, nr 2, s. 3-15
 - [145] Wnuk W.: System informacyjny z dzieloną bazą danych, Kom. Inst. Cybern. PWr., Wrocław, 1977, nr 596, s. 20
 - [146] Yao A.Ch.-Ch.: An $O(E \log V)$ algorithm for finding minimum spanning trees, Inf. Proc. Letters, 1975, vol. 4, no. 1, p. 21-23
 - [147] Yu C.T., Salton G., Siu M.K.: Effective automatic indexing using term addition and deletion, J. ACM, 1978, vol. 25, no. 2, p. 210-225
 - [148] Yuval G.: Finding near neighbors in k-dimensional space, Inf. Proc. Letters, 1973, vol. 3, no. 2, p. 113-114
 - [149] Zahn Ch.T.: Graph-theoretical methods for detecting and describing gestalt clusters, IEEE Trans. Comp., 1971, vol. C-20, no. 1, p. 68-86

Dodatek 1. Opis struktury cytowań

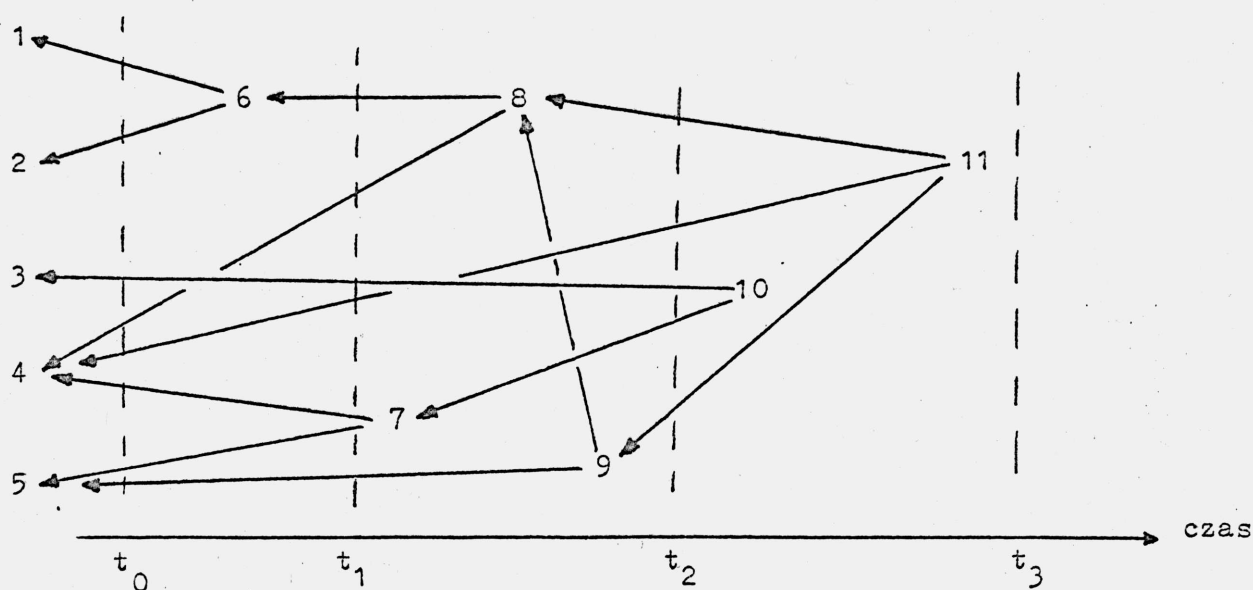
Założmy, że posiadamy zbiór dokumentów /artykułów/
 $D = \{d_i \mid i = 1, 2, \dots, n\}$ oraz zbiór $B = \{b_i \mid i = 1, 2, \dots, m\}$ będący
 ich skomasowaną bibliografią /przy analizowaniu odpowiednio
 długiego okresu czasu zazwyczaj $D \cap B \neq \emptyset$ /. Określmy macierzą
 C odwzorowanie $\Omega : D \rightarrow 2^B$ /być może nieróżnowartościowe/
 w następujący sposób:

$$C \triangleq D \times B = (c_{ij})_{n \times m}$$

gdzie:

$$c_{ij} \triangleq f(d_i, b_j) = \begin{cases} b & , \text{ jeśli } d_i \text{ cytuję } b_j ; \text{ /zazwyczaj } b=1/, \\ a & , \text{ jeśli } d_i \text{ nie cytuję } b_j ; \text{ /zazwyczaj } a=0/. \end{cases}$$

Ilustracja:



$B_{t_3} = \{1, 2, 3, \dots, 10, 11\}$ - bibliografia dla chwili t_3

$B_p \subset B : B_p = \{1, 2, 3, 4, 5\}$ - prabibliografia dla t_0

$D_{t_1, t_2} = \{7, 8, 9\}$ - zbiór dokumentów opublikowanych w okresie $[t_1, t_2]$

$D_{t_2, t_3} = \{10, 11\}$ - zbiór dokumentów opublikowanych w okresie $[t_2, t_3]$

$$C_{t_3} = D \times B = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \end{matrix} \\ \begin{matrix} 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \end{matrix} & \begin{bmatrix} 1 & 1 & & & & & & & & & \\ & & & 1 & 1 & & & & & & \\ & & & 1 & & 1 & & & & & \\ & & & & 1 & & & 1 & & & \\ & & & & & & & & 1 & & \\ & & 1 & & & & 1 & & & & \\ & & & 1 & & & & & 1 & 1 & \end{bmatrix} \end{matrix}$$

Określmy macierz C_{t_3} ; $C_{t_3} \triangleq A \times A = (c_{ij})_{p \times p}$; gdzie:
 $A = D \cup B$, $p = \text{card } A$.
 Wówczas macierz: $C_{t_3}^k \triangleq C_{t_3}(C_{t_3}^{k-1})$, $/k - \text{naturalne}$, $C_{t_3}^0 \equiv C_{t_3}$ /
 określa nam relację k -cytowania $/d_i$ cytuje b_j łańcuchem długości k /
 $\Omega^k : D \rightarrow 2^B$.

Relacja cytowania jest niesymetryczna, przechodnia i nie jest
 zwrotna.

Przyjmując za prawdziwe następujące fakty:

- autor publikacji cytuje dokumenty, których treść jest związana z treścią artykułu,
- prace, w których cytuje się podobne bibliografie są tematycznie związane,
- autorzy prac z wybranej dziedziny czerpią cytaty ze wspólnej puli prac wcześniejszych dotyczącej tej dziedziny,
- cytowana bibliografia zawiera prace związane tematycznie, można określić współczynniki podobieństwa między artykułami na podstawie grafu cytowań.

Dodatek 2: Własności empirycznej bazy danych

A. Materiał doświadczalny i selekcja danych

Na podstawie wykazu czasopism przeglądanych w ISI /przy tworzeniu indeksu cytowań SCI/ oraz listy INSPEC'a i doświadczenia autora wybrano 15 najważniejszych czasopism anglosaskich z zakresu s.w.i. Z czasopism tych /głównie z "Journal of Documentation" - dział: Progress in documentation/ wyselekcjonowano 13 prac podsumowujących /bibliografia dłuższa niż 50 pozycji/ o wybranych aspektach s.w.i. Autorzy tych prac /m.in. G. Salton, K. Spärck Jones, G.D. Knott, S.E. Robertson, C.J. Van Rijsbergen, A. Bookstein/ będą dalej oznaczani wspólnie mianem Exp I. Tytuł każdej z tych prac będzie tematem hipotetycznego pytania informacyjnego dla s.w.i., natomiast jej bibliografia będzie zbiorem prac relewantnych dla tego pytania /wg oceny autora artykułu - Exp II/. Takie podejście umożliwia, jak się okaże, redukcję rozmiarów eksperymentalnej bazy danych /b.d./ a ponadto ocenę rezultatów algorytmów podziału.

W b.d. wydzielono dwa zbiory: OBD - obiekty b.d., BBD - bibliografia b.d., /OBD $\subset$ BBD/. Poczynając od najstarszych dostępnych roczników wybranych czasopism dokonano /metodą przeglądu pełnego/ kwalifikacji artykułów /wraz z ich słowami kluczowymi do OBD, a ich zredukowanych bibliografii /wraz z pracą cytującą/ do BBD.

Do b.d., przyrastającej w czasie, przypisywano prace:

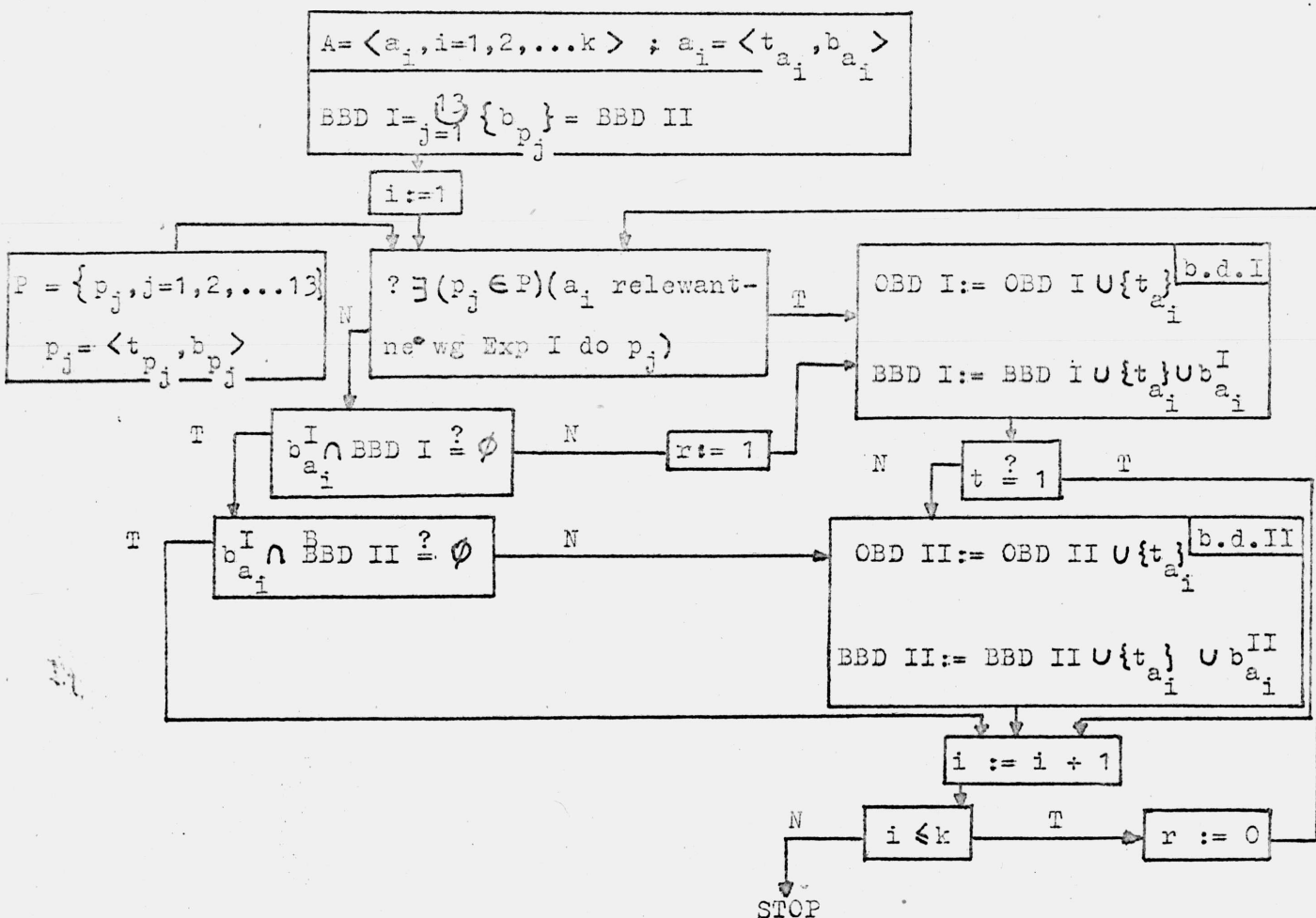
- relewantne /wg Exp.I/ dla któregośkolwiek z pytań informacyjnych lub
- cytujące co najmniej jeden obiekt z aktualnej BBD.

Przebieg tworzenia baz danych dla eksperymentów i reguły selekcji obiektów przedstawiono na rys. D2.1

W wyniku selekcji powstały dwie eksperymentalne bazy danych b.d.I i b.d.II /oczywiście, b.d.II zawiera się całkowicie w b.d.I/.
Rezultat selekcji przedstawia TABELA I.

Ogółem przeglądnięto 6322 artykułów. Liczność baz danych jest następująca:

- b.d.II : card OBD II = 419 , card BBD II = 832 ,
- b.d.I : card OBD I = 785 , card BBD I = 2131.



A - zbiór wszystkich publikacji z wybranych czasopism uporządkowany chronologicznie,

t_{a_i} - opis publikacji a_i ,

b_{a_i} - bibliografia publikacji a_i ,

P - zbiór pytań informacyjnych /wybranych publikacji/, $P \subset A$,

$b_{a_i}^I$ - bibliografia zredukowana publikacji a_i /przez usunięcie cytowań sprzed 1970 r. lub cytowań, które nie są monografiami albo odesłaniami do ustalonego zestawu czasopism/,

$b_{a_i}^{II}$ - bibliografia powtórnie zredukowana publikacji a_i / $b_{a_i}^{II} = b_{a_i}^I \setminus \{\text{monografie}\}$ /.

Rys. D2.1. Schemat selekcji artykułów

TABELA I. Wykaz ilościowy wyselekcjonowanych prac

Czasopismo		1970	1971	1972	1973	1974	1975	1976	1977	1978	1979 +1980
1/ Communication of ACM *)	1				114	103	96	85	112	113	71 ^{1/}
	2				8	3	5	3	6	6	4
	3				10	4	5	6	10	12	6
2/ IEEE Trans. on Computers	1			212	186	210	181	186	202	201	55
	2			1	2	2	-	2	3	2	-
	3			1	2	2	1	3	4	2	1
3/ IEEE Trans. on System, Man and Cybernetic	1				109	97	82	133	121	146	47 ^{2/}
	2				-	-	1	1	-	-	-
	3				1	-	2	5	-	1	3
4/ Pattern Recognition	1					17	23	31	27	43	43
	2					1	-	-	-	-	-
	3					2	4	8	5	4	1
5/ IEEE Trans. on Software Engineering	1						33	40	43	50	37
	2						-	-	-	-	-
	3						4	6	8	10	9
6/ The Computer Journal	1			65	65	63	68	71	67	66	82 ^{3/}
	2			4	5	3	5	1	4	7	4
	3			4	6	5	12	10	13	11	8
7/ Information Processing Letters /od 1971 r/.	1			9	-	-	44	96 ^{9/}	56	77	142
	2			4	-	-	1	3	2	3	8
	3			4	-	-	1	3	3	3	12

TABELA I cd.

Czasopismo \ Rok		1970	1971	1972	1973	1974	1975	1976	1977	1978	1979 +1980
8/ Information Systems /od 1975r/	1						14	14		21	26
	2						7	8		5	14
	3						7	9		10	20
9/ Journal of Am. Soc. Inf. Sciences /dawniej American Document/	1				63	49	38 ^{4/}	46	46	67	43 ^{5/}
	2				8	5	7	6	8	9	10
	3				10	12	10	18	11	15	20
10/ Journal of Documentation	1				19	24	22	21	19	4 ^{6/}	19
	2				3	8	8	2	5	1	7
	3				4	12	13	11	8	2	11
11/ Information Processing Management /dawniej Inf. Storage and Retrieval/	1	34	3 ^{7/}	25	49	8 ^{8/}	23	32	35	59	29
	2	7	1	7	29	-	12	10	22	12	17
	3	9	2	10	35	-	16	21	23	31	19
12/ Journal of ACM	1				51	62	50	64	49	56	57
	2				4	4	3	3	6	8	8
	3				5	8	12	10	7	12	10
13/ IBM Systems Journal	1				20	16	21	18	20	22	29
	2				1	1	-	1	3	-	-
	3				2	3	-	1	3	2	1
14/ Acta Informatica /od 1971 r./	1		5	8 ^{8/}	23	48	10	53	41 ^{11/}	25	50 ^{12/}
	2		2	-	1	4	1	2	1	2	3
	3		2	-	3	8	1	17	4	7	14
15/ Journal of Research and Development	1			65	48	53	60	56	57	64	65
	2			1	4	3	3	2	1	2	1
	3			1	6	3	8	2	7	9	11

TABELA II. Macierz wzajemnych cytowań między czasopismami

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16 inne
1	73(25)	5(3)	-	-	-	6(2)	-	-	1(0)	5(0)	12(0)	10(3)	1(0)	2(0)	1(1)	5
2	7(4)	22(3)	2(0)	3(0)	-	-	2(1)	-	-	-	-	3(1)	1(0)	-	1(0)	18
3	-	1(0)	1(0)	3(0)	-	-	-	-	-	-	-	1(0)	-	-	2(0)	13
4	1(0)	1(0)	2(0)	-	-	-	-	-	-	-	-	-	-	-	-	-
5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6	88(21)	-	1(1)	-	-	64(15)	1(1)	-	-	3(1)	14(11)	10(3)	4(2)	-	3(1)	104
7	10(6)	2(1)	-	-	-	7(4)	11(9)	-	-	-	-	5(1)	-	1(0)	-	2
8	78(51)	2(0)	-	-	-	4(0)	-	8(3)	8(0)	7(0)	4(2)	16(5)	10(4)	5(5)	7(3)	28
9	12(5)	-	-	-	-	9(5)	-	-	108(38)	60(10)	63(17)	31(10)	2(2)	-	-	32
10	20(5)	2(1)	1(1)	-	-	8(3)	-	4(4)	125(46)	89(32)	85(28)	21(7)	1(0)	-	2(1)	14
11	67(14)	3(0)	2(1)	-	2(1)	10(1)	5(5)	2(2)	119(39)	69(18)	144(46)	43(12)	5(0)	1(0)	3(1)	28
12	31(18)	8(2)	-	-	-	2(1)	5(3)	-	25(12)	3(1)	6(2)	36(17)	1(0)	-	-	21
13	30(0)	-	-	-	-	1(0)	-	3(0)	-	-	-	-	14(0)	1(0)	8(0)	15
14	31(2)	-	-	-	-	3(0)	6(4)	-	-	-	3(0)	4(0)	1(0)	11(4)	2(1)	24
15	15(4)	4(0)	-	-	-	3(1)	-	-	-	-	-	8(1)	18(4)	2(0)	16(3)	11

*) numery czasopism odpowiadają numerom z TABELI I

**) pierwsza liczba - łączna liczba cytowań, liczba w nawiasie - cytowania dla prac z 1978-1980 r.

Wykaz uwag do TABELI I:

- */ wiersz: 1 - liczba wszystkich prac w wolumenie,
2 - liczba prac zakwalifikowanych do b.d. II,
3 - liczba prac zakwalifikowanych do b.d. I.
- 1/ do numeru 6, 1979 r., 7/ tylko numer 6, 1971 r.,
2/ do numeru 5, 1979 r., 8/ rocznik niedostępny /zaginął/,
3/ do numeru 1, 1980 r., 9/ bez numeru 3, 1976 r.,
4/ bez numeru 3, 1975 r., 10/ tylko vol. 1, fasc. 1,
5/ do numeru 5, 1979 r., 11/ bez vol. 9, fasc. 4,
6/ tylko numer 2, 1978 r., 12/ z 1980 r. tylko fasc. 1.

Ogółem przglądnięto 6322 artykułów. Do b.d. I zakwalifikowano łącznie 785 prac, a do b.d. II 419.

B. Cytowania i związki między wybranymi czasopismami

W TABELI II przedstawiono macierz /niesymetryczną/ wzajemnych cytowań między wybranymi czasopismami na podstawie bibliografii niezredukowanej prac.

Czasopisma inne to m.in.: Computer Surveys, Computer Science Theory, Journal of Applied Statistic, Psychometrika, IEEE Information Theory, Int. J. of Computer Information Science, ACM Comput. Surveys, ACM Trans. on Database Management, Journal of Chemical Documentation, Theoretical Computer Science, Information Science, Operations Research, ACM Trans. on Data Base Systems, SIAM J. Appl. Math.

B. Prawidłowości bibliograficzne

B.1. Liczba współautorów artykułu

W TABELI III pokazano jednowymiarowy szereg rozdzielczy liczby współautorów prac z b.d. I.

TABELA III. Szereg rozdzielczy liczby współautorów

liczba autorów	1	2	3	4	5
liczba artykułów	627	150	41	11	3
% liczby artykułów	75,36	18,029	4,928	1,322	0,361
P	75,188	18,655	4,63	1,148	0,338

Okazuje się, że histogram ten jest dobrze przybliżany przez obcięty i przesunięty rozkład Pascala $p(k) = \frac{a^k}{(a+1)^{k+1}}$, $k=0,1,\dots,4/$

z parametrem $a = 0,39$ /wiersz P tabeli III/, lecz przetestowanie tej hipotezy, ze względu na licznosc próby i dyskretność rozkładu, standardowymi testami jest niemożliwe.

Potwierdza się prawo geometrycznej zależności:

$$p(i) = \frac{c}{b^i}, i = 1,2,\dots,5; a, c - \text{stałe}$$

Przyjmując, że $c = 41 a^3$ /zerowy błąd dla $i=3/$ metodą najmniejszych kwadratów obliczono parametry: $a = 3,91$, $c = 2450,835$.

B.2. Wydajność autorów artykułów

Uszeregowanie autorów artykułów według liczby opublikowanych przez nich prac przedstawiono w TABELI IV.

TABELA IV. Wydajność naukowa autorów

numer autora	1	2	3	4,5	6,7	8-10	11-14	15-21	22-38	39-60	61-102	103-174	175-620
liczba prac	28	18	15	13	11	10	8	7	5	4	3	2	1

/1 = G. Salton, 2 = K. Sparck Jones, 3 = C.J. Van Rijsbergen, 4 = S.E. Robertson, 5 = C.T. Yu, G = A. Bookstein, 7 = W.S. Cooper/.

Wydajność autorów maleje w zasadzie wykładniczo lecz istnieją duże odstępstwa od prawa Zipfa-Lotki. Wobec czego z danych z b.d.II nie można wnioskować o cechach s.w.i., w którym podobieństwa określone byłyby na podstawie związków nazwisk /propozycja Cleverdona nie poparta analizą statystyczną/.

Spowodowane jest to wieloma przyczynami:

- b.d. nie ujmuje całokształtu działalności naukowej autorów /innych ich publikacji/,
- b.d. dotyczy zagadnień o krótkiej historii,
- być może jest to wynikiem redukcji bibliografii.

B.3. Częstość cytowania

TABELA V przedstawia:

- szereg rozdzielczy częstości cytowań prac należących do EBD II przez publikacje z OBD II,

TABELA V. Częstość cytowań między pracami

liczba cytowań	0	1	2	3	4	5	6	7	8
liczba prac zaobserwowana	201	338	123	55	32	26	14	11	7
liczba prac oczekiwana	188,04	338,125	138,97	62,73	33,87	23,474	12,485	9,6576	6,827
χ^2_i	0,8932	0,00005	1,8352	0,9525	0,103	0,272	0,1838	0,1866	0,0044
przedziały dla y	0÷0,5 I	0,5÷1,5 II	1,5÷2,5 III	2,5÷3,5 IV	3,5÷4,5 V	4,5÷5,5 VI	5,5÷6,5 VII	6,5÷7,5 VIII	7,5÷8,5 IX

liczba cytowań	9	10	11	12	13	14	16	19	25
liczba prac zaobserwowana	7	5	3	3	3	1	1	1	1
liczba prac oczekiwana	4,58	3,58	4,914	4,74					
χ^2_i	1,28	0,5632	0,24	1,077					
przedziały dla y	8,5÷9,5 X	9,5÷10,5 XI	10,5÷12,5 XII	12,5÷∞ XIII					

- histogram częstości cytowań,
- histogram skupkowy lognormalnego :

$$f_Y(y) = \frac{1}{y\sqrt{2\pi}\sigma_{\ln Y}} \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma_{\ln Y}} \ln \frac{y}{\bar{m}_Y}\right)^2\right] \approx \text{LN}(\bar{m}_Y, \sigma_{\ln Y})$$

- wartości statystyki χ^2 .

Uzasadnieniem przyjęcia rozkładu lognormalnego jest multiplikacyjny /w czasie/, charakter procesu cytowania [88, 109].

Nie można obserwacji podzielić na klasy o w przybliżeniu równej liczności /choć jest to własność pożądana dla testu χ^2 /, bo byłoby

ich dwie. Najmniej liczne klasy połączono w celu uniknięcia obciążenia w obliczaniu χ^2 .

Estymacja parametrów metodą największej wiarygodności wymaga rozwiązania metodami numerycznymi układu równań nieliniowych, wobec czego zastosowano metodę momentów /estymacja punktowa/.

Parametry rozkładu wynoszą:

$$EY = m_Y = 1,83059, D^2Y = 9,35697 - m_Y^2 = 6,006134, \sigma_Y = 2,4507415 \\ \sigma_{\ln Y} = 0,989, \bar{m}_Y = 1,19.$$

Hipotezy: $\begin{cases} H_0 : Y \approx \text{LN}(1,19, 0,989) \\ H_1 : Y \neq \text{LN}(1,19, 0,989) \end{cases}$ /nie daje się określić błędu II rodzaju/.

Z tablic: $\chi^2_{0,05;10} = 18,25$, $\chi^2_{0,15;10} = 3,9$ /moda ≈ 8 /. Ponieważ

XIII
 $\sum_{i=1}^X \chi^2_i = 7,59 \in (3,9; 18,25)$ na 10 % poziomie istotności nie ma podstaw do odrzucenia H_0 /test dwustronny, symetryczny/.

Wnioski:

- skumulowany wokół wartości całkowitych rozkład lognormalny bardzo dobrze przybliżył histogram dyskretnego rozkładu częstości cytowań /przypuszczenie G.Saltona [109] potwierdziło się/,
- szansa dużej liczby cytowań w rozpatrywanym 10-letnim okresie szybko maleje - selektywność s.w.i. powinna być duża /podobieństwa między artykułami obliczane na podstawie ich bibliografii powinny przyjmować znacznie różniące się wartości/,
- pośrednio potwierdza się hipoteza, że średni "czas życia" artykułów wynosi od 4,65 ÷ 12 lat, co uzasadnia przyjętą w pracy granicę wieku /10 lat/,
- dla b.d.I /uwzględnienie cytowań monografii/ rozkład mocno się zniekształca /pojawia się długi "ogon", np. monografię D. Knutha cytowano 109 razy, a G. Saltona 75 razy/,

- współczynnik zmienności $V_Y = \sigma_Y / m_Y = 1,34$ a więc jest duży,
- współczynnik asymetrii $\gamma_1 = \mu_Y^{(3)} / \sigma_Y^3 = 3,25$, rozkład $f_Y(y)$ jest mocno, dodatnio asymetryczny,
- przyjmując, że opis jest przesuniętym rozkładem lognormalnym da się na pewno /dodatkowy parametr/ zapewnić jeszcze większą zgodność opisu z danymi empirycznymi.

B.4. Długość bibliografii, wiek i średni wiek bibliografii

Histogram rozkładu długości bibliografii z OBD II /po redukcji/ przedstawiono w TABELI VII, natomiast rozkład wieku bibliografii przedstawia TABELA IX, a średniego wieku bibliografii TABELA XI.

B.5. Wpływ redukcji bibliografii na parametry bibliograficzne

B.5.1. Długość bibliografii

Histogramy rozkładu długości bibliografii przed i po redukcji /OBD II/ przedstawiają TABELĘ VI i VII:

TABELA VI. Histogram długości bibliografii przed redukcją

długość bibliografii	0	1	2	3	4	5	6	7	8	9	10	11	12	13
liczba prac	4	2	12	13	13	26	31	32	31	32	31	29	18	18

14	15	16	17	18	19	20	1	2	3	4	5	6	7	8	9	30
16	14	14	6	4	4	7	8	6	7	3	1	1	2	2	2	2

31	2	3	4	7	8	9	41	46	47	49	50	52	64	71	93	105	119	121	140
1	2	1	2	1	1	1	2	1	1	2	1	2	1	1	1	1	1	1	1

152	203	328
1	1	1

$EX = 14,642 / 10,02$ przy pominięciu prac przeglądowych
o liczbie cytowań powyżej 34/

$$EX^2 = 146,047$$

$$D^2X = 45,5698 ; \sigma = 6,7505$$

Wyraźnie uwidacznia się dwumodalność histogramu, co nie potwierdza przypuszczeń G. Saltona o rozkładzie wykładniczym.

Daje się wyróżnić trzy grupy publikacji:

- o długości bibliografii $2 \div 16$,
- o długości bibliografii w pobliżu $20 / 19 \div 24 /$,
- wszystkie prace o bibliografii dłuższej niż 37 okazały się pracami przeglądowymi.

Jeśli przyjąć, że rozkład graniczny ekstremalnych długości bibliografii /bibliografie prac przeglądowych/ można przybliżyć rozkładem Gumbela II rodzaju /ograniczonym z lewa przez zero/, to można oszacować rozkład zmiennej Y , $Y = \max\{X_i\} : F_Y(y) = e^{-(n/y)^k}$, $y \geq 0$. Wówczas: $V_Y^2 = \left(\frac{\sigma_Y}{m_Y}\right)^2 = \frac{\Gamma(1-2/k)}{\Gamma^2(1-1/k)} - 1$, $k > 2$. Dla naszych danych: $m_Y = 90,2857$, $\sigma_Y = 65,542$, wobec tego $V_Y = 0,7259$, skąd $k \approx 2$. Czyli $n = \frac{m_Y}{\Gamma(\frac{1}{2})} = 50,94$. Korzystając z tablic rozkładu F_Y znajdujemy że prawdopodobieństwo, iż maksymalna długość bibliografii przekroczy 376 pozycji, jest mniejsze niż 0.02.

Powyższe rozważania mają charakter orientacyjnych szacunków /mała próba/ i raczej ilustruje tylko możliwość przewidzenia maksymalnego rozmiaru rekordu w zbiorze danych.

TABELA VII. Histogram długości bibliografii po redukcji

długość bibliografii	0	1	2	3	4	5	6
liczba prac	14.	23	49	70	75	66	48
oczekiwana liczba prac	5,95	25,00	52,51	73,51	77,19	64,84	45,38
χ^2	1,186	0,2346	0,1676	0,062	0,02	0,1513	

długość bibliografii	7	8	9	10	11	12	13	14	18	24
liczba prac	28	12	8	*) 2 (1)	2 (2)	(3)	(4)	(6)	(5)	(1)
oczekiwana liczba prac	27,23	14,30	6,67	2,803	1,07					
χ^2	0,55									

*) Dalsze obliczenia przeprowadzono z pominięciem prac przeglądowych /liczby w nawiasach/.

Znany w literaturze jest fakt, że w dostatecznie licznej b.d. rozkład długości bibliografii jest dobrze przybliżany przez roz-

kład Poissona.

Chcemy więc przetestować zgodność danych $EX = 4,1687 \approx \lambda$ z rozkładem Poissona z parametrem $\lambda = 4,2$ przy pomocy testu χ^2 . W celu uniknięcia obciążenia i wpływu doboru przedziałów zmienności podzielono dane na grupy o w przybliżeniu równej liczności /obniżamy tym samym liczbę stopni swobody $\sqrt{}$ /.

Hipotezy:
$$\begin{cases} H_0 : X \approx \text{Poiss}(4,2) \\ H_1 : X \approx \text{Poiss}(4,2) \end{cases}$$

Wartość krytyczna $P\{\chi^2 > \chi^2_{(0,05;5)}\} = 0,05$, wynosi $\chi^2_{(0,05;5)} = 11,07$.

Ponieważ $\sum \chi^2_i = 2,3690705 < 11,07$, nie ma podstaw, aby odrzucić na 5 % poziomie istotności hipotezę H_0 .

/dla $\sqrt{v} = 9$, $\chi^2_{(0,005;9)} = 16,919$, a $\sum \chi^2_i = 12,65 < 16,919$ -

- czyli też brak podstaw do odrzucenia H_0 /.

Wnioski:

- redukcja spowodowała, że 10 artykułów /2,4%/ pozbawiono bibliografii - artykuły te nie będą wydawane w odpowiedzi na żadne pytanie, chyba że będą zacytowane przez inne prace,
- radykalnie zmienił się charakter rozkładu /jednomodalność, znaczne skrócenie bibliografii - co jest ważne dla ekonomicznego przechowywania opisów, rozkład Poissona - możliwość przewidywania wzrostu BBD w czasie eksploatacji/,
- powtarzając, jak dla TABELI VI, szacowanie maksymalnej długości bibliografii zredukowanej dla prac przeglądowych otrzymujemy, że prawdopodobieństwo, iż maksymalna długość bibliografii przekroczy 29, wynosi mniej niż 0,02.

Zaobserwowano ponadto, że większość prac odrzuconych z bibliografii prac przeglądowych była: trudnodostępna /raporty, materiały konferencyjne, preprinty/, cytowana dla kilku zdań - a więc o znikomej przydatności. Ponadto część prac, cytowana jako raporty, była już opublikowana w czasopiśmie przed ukazaniem się pracy przeglądowej.

B.5.2. Rozkład wieku cytowań

Histogramy rozkładu wieku cytowań w b.d.II przed i po redukcji pokazują TABELI VIII i IX /odpowiednio/.

Umowa:

► wiek cytowania = rok opublikowania pracy cytującej - rok opublikowania pracy cytowanej ◀

TABELA VIII. Histogram wieku cytowań przed redukcją

wiek	*)	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13
liczba cytowań		8	134	582	839	930	681	653	403	290	259	201	153	126	91	73

14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
84	65	51	34	32	25	24	18	11	7	4	12	9	8	5	9	8	3	2	1	2	2

36	37	38	39	40	41	43	44	45	47	48	49	50	51	52	53	61	74	79	105	254	258
1	1	3	2	1	1	2	2	1	2	1	2	1	1	1	1	1	1	1	1	1	1

*) wiek cytowania ujemny oznacza cytowanie "w przód" - prace, które będą opublikowane.

Sumaryczna liczba cytowań wynosi 5908, a powinna /na podstawie TABELI VI/ wynosić 5972. Spowodowane to jest tym, że 64 prace cytowane nie były datowane.

Rozkład wieku cytowań nie daje się opisać żadnym ze standardowych rozkładów obciętych. Można stwierdzić, że:

- najczęściej cytuje się prace 2-4 letnie,
- od wielkości modalnej /ok. 3/ częstości szeregu rozdzielczego maleją w zasadzie hiperbolicznie i dosyć wolno,
- potwierdza się hipoteza [88], że prace starsze niż 25 lat są cytowane ze względów pozanaukowych i często są wynikiem cytowań wtórnych,

- rozkład z TABELI VIII jest najlepiej aproksymowany /średniokwadratowo/ przez obcięty do przedziału [0,24] rozkład $\Gamma(\lambda, k)$

dla $\lambda = \frac{EX}{D^2X} = 0,3076$ i $k = m_x \lambda = 1,61$ lecz autor nie znalazł uzasadnienia dla tego faktu w rodzaju np. mechanizmu rządzącego wiekiem cytowania; a klasą rozkładów Γ da się opisać większość różnych rozkładów empirycznych.

TABELA IX. Histogram wieku cytowań po redukcji

wiek cytowań	0	1	2	3	4	5	6	7	8	9	10
liczba cytowań	52	112	245	248	210	275	172	88	108	93	52
liczba cyt. przewidyw.	20,3	89,4	198,7	288,5	317,3	279,2	204,8	128,7	70,8	34,6	15,2
χ^2_i	49,5	5,7	11,8	5,7	36,3	0,06	5,25	12,9	19,5	98,6	89,1

Znów niezmiernie trudno określić charakter rozkładu, W TABELI IX pokazano przykładowo jak duże wartości przybiera statystyka $\sum x_i^2$ dla rozkładu Poissona z $\lambda = 4,4$ /EX = 4,4897, po uwzględnieniu obcięcia rozkładu $\lambda \approx 4,4/$. Otóż $x^2 = 334,4$, a $\chi^2_{0,05;9} = 23,589 \ll x^2$.

Stwierdzić można natomiast, że jeśli rozkład z TABELI VIII obciąć do przedziały $[0, 10]$ to okaże się, że średni wiek cytowania wyniesie 4,08 lat. A więc, niestety, redukcja wydłużyła średni wiek pojedynczego cytowania /do 4.49 lat/, choć różnica nie jest duża. Analiza danych pokazała, że główną przyczyną tego zjawiska jest cytowanie materiałów konferencyjnych, a w mniejszym stopniu cytowanie raportów, preprintów, itp.

B. 5.3. Rozkład średniego wieku bibliografii

Innym parametrem danych, na który ma wpływ redukcja jest średni wiek bibliografii pracy /jest to suma wieku cytowań dzielona przez liczbę cytowań w bibliografii publikacji/. Histogramy średniego wieku przed i po redukcji przedstawiono w TABELACH X i XII.

TABELA X. Histogram średniego wieku cytowań przed redukcją

średni wiek bibl.	0÷1	1÷2	2÷3	3÷4	4÷5	5÷6	6÷7	7÷8	8÷9
liczba prac	0	7	34	57	60	78	59	49	30
liczba przewidyw.	1,71	9,42	25,9	47,48	65,29	71,82	65,83	51,7	35,6
$\chi^2_{i,11}$	1,53		2,53	1,9	0,43	0,53	0,71	0,14	0,88
$\chi^2_{i,7}$	0,43			1,9	0,43	0,53	0,71	0,14	0,88

średni wiek bibl.	9÷10	10÷11	11÷12	12÷13	13÷14	14÷15	15÷16	18÷19
liczba prac	15	19	7	2	3	2	2	1
liczba przewidyw.	21,73	11,95	5,97	4,6				
$\chi^2_{i,11}$	2,08	0,09	0,18	6,34				
$\chi^2_{i,7}$	0,72							

Dla poziomu istotności 0,05 nie ma podstaw do odrzucenia hipotezy, iż histogram częstości ma obcięty rozkład Poissona z parametrem $\lambda = 5,5$ /EX = 5,475/, bo $\chi^2_{(0,05;11)} = 19,675 > \sum \chi^2_{i,11} = 17,34$, a ponadto /wyrównanie licznosci/ $\chi^2_{(0,10;7)} = 12,017 > \sum \chi^2_{i,7} = 5,74$. Przyczyna tego faktu nie jest jednak jasna.

TABELA XI. Histogram średniego wieku cytowań po redukcji .

średni wiek bibliografii	0÷1	1÷2	2÷3	3÷4	4÷5	5÷6	6÷7	7÷8	8÷9	9÷10
liczba prac	7	46	91	103	88	47	14	11	1	1

Wzięto pod uwagę 409 prac, bo 10 w wyniku redukcji pozostało bez bibliografii. Wyraźnie widoczne jest znaczne "odmłodzenie" średniego wieku bibliografii /EX = 3,18/. W tym aspekcie redukcja ma charakter pozytywny - wyodrębnia prace niedawne, a więc o hipotetycznie większym podobieństwie. Z drugiej strony, autorzy prac często określają źródło pierwotnej inspiracji /prace starsze/ co może mieć znaczenie dla historyka rozwoju nauki.

Okazuje się ponadto, że w wyniku redukcji w 18 przypadkach nastąpiło zwiększenie średniego wieku bibliografii, a w 117 pozostał on bez zmian.

C. Podsumowanie

Na podstawie wykonanych obliczeń można stwierdzić, że:

a/ Redukcja bibliografii przy odpowiednio dobranym progu starości oraz ograniczeniu się do cytowań w wybranych czasopismach ma szereg zalet:

- skraca długość bibliografii /ekonomiczne gromadzenie danych/,
- zmniejsza radykalnie wielkość zbioru cech - prac cytowanych /u nas ok. 4-o krotnie/,
- poprawia selektywność s.w.i.,
- skraca średni wiek bibliografii,
- "poprawia" kształt rozkładów - możliwość szacowania parametrów eksploatacyjnych s.w.i.,

Wady redukcji:

- część prac pozostaje bez bibliografii,
- znikają, istotne w innych zastosowaniach /socjologia i historia nauki/, więzi między autorami,
- następuje wydłużenie /nieznaczne/ średniego wieku cytowania,

- b/ Parametry bibliograficzne /zgodność z prawami Zipfa, Bradforda, Lotki/ wybranego zbioru eksperymentalnego nie odbiegają w rażący sposób od tych stwierdzonych prawidłowości - baza danych nie została dobrana tendencyjnie,
- c/ Możliwym okazuje się automatyczne wyodrębnienie prac o charakterze przeglądowym tylko na podstawie analizy długości bibliografii.

Thematical Grouping of the Data Bases in the Library Information Systems

The presented work deals with some problems of information storage and retrieval systems. Especially, the problem of document clustering on monothematical groups in a bibliographical data base is developed in detailed way. The proposed grouping algorithms use the degree of similarity between descriptions of documents. The bibliography of a document has been treated as its description. The Author proposes four ways of evaluating the coefficients of objects similarity; all of them are based on the analysis of science citation structure / using k-citation relation /. The structure of science citation is presented as a oriented citation graph. This approach enables grouping of similar documents without the necessity of their manual indexing. This fact is very valuable from the point of view of practical realization of such a information retrieval system / possibility of elimination of descriptor classifications that are very popular in the indexing /. In the experiments grouping are constructed by means of graph-theoretical algorithms using the concept of k-connectivity relation. Data base $B = \{b_i\}$ is presented in the form of a graph $G = (V(G), E(G), \mu)$. The objects of the data base are vertices $V(G)$, edges $E(G)$ represent object similarity relation and $\mu: E(G) \rightarrow \mathbb{R}_+$ is object similarity / dissimilarity / function. The graph G is finite, simple, weighted on edges by the function μ , and nonoriented. After detailed analysis of properties of graph-theoretical methods of grouping, usage of k-components, k-bonds, k-blocks as a cluster definition has been proposed. Comparison of graph-theoretical and graph-heuristical quasi-optimal grouping methods / for example: Zahn's, Gotlieb-Kumar's, Bonner's, Volkina-Zagorujko's methods / gives some interesting motions. Empirical properties of ten selected graph algorithms for grouping have been tested on an exemplary data base contains articles from fifteen journals / dealing with information systems / gathered in the years 1970-1980. The bibliographies from overview works of well known specialists in the science have been taken as a pattern grouping. The obtained grouping have been compared with the experts' grouping by means of a modified Rand's matching coefficient. The experiments have proved generall consistency between these groupings. Besides, the obtained groupings are more precise and stable than the grouping of data base made according to key-words. The significant result of the work is empirical analysis of dependence of description reduction on the effectiveness of grouping algorithms. Some bibliometric properties of the considered data base have been evaluated and discussed.

All the experiments have confirmed the usefullness of properly reduced bibliography of articles for the indexing and grouping purposes. Theoreti-

cal and practical meaning of the analysis of grouping problems in the field of a graph theory is quite evident because of the possibility of a priori approximation of the parameters of resulting grouping and of small computational complexity of the graph-algorithms.

Mgr inż. Włodzimierz Wnuk

Raport złożono w Redakcji
Instytutu w marcu 1981 r.

Niniejszy Raport otrzymują:

1. Biblioteka Międzyinstytutowa	1 egz.
2. Biblioteka Główna	1 egz.
3. Z-ca Dyrektora d/s Rozwoju Kadry Naukowej	1 egz.
4. Promotor	1 egz.
5. Recenzenci	2 egz.
6. Autor	2 egz.
7. Wydział Informatyki Uniwersytetu Jagiellońskiego	1 egz.

Razem: 9 egz.

