STUDIA INFORMATICA 2013
Volume 34 Number 3 (113)

Dariusz BERNATOWICZ
Politechnika Koszalińska, Wydział Elektroniki i Informatyki
Anna BERNATOWICZ
Politechnika Koszalińska, Wydział Inżynierii Lądowej,
Środowiska i Geodezji

ZASTOSOWANIE KORELACJI W PROCESIE FRAGMENTACJI PIONOWEJ ROZPROSZONYCH BAZ DANYCH

Streszczenie: Celem poniższego artykułu jest przedstawienie podejścia dotyczącego zastosowania korelacji danych wejściowych opartych na statystyce zapytań i częstości ich wystąpienia w rozproszonych bazach danych. Podejście to stanowi alternatywną technikę redukcji liczby gałęzi w grafie podziału. Określa także kierunek i siłę zależności pomiędzy poszczególnymi elementami, która jest wykorzystywana przy ustalaniu kryterium podziału. Zawarto również krótką charakterystykę procesu fragmentacji pionowej opartej na statystyce zapytań oraz rozwinięcie algorytmu graficznego, umożliwiającego rozwiązanie problemu niespójności grafu.

Słowa kluczowe: fragmentacja pionowa, rozproszone bazy danych, graficzny algorytm podziału, macierz i graf korelacji

APPLICATION OF CORRELATION IN VERTICAL FRAGMENTATION PROCESS IN DISTRIBUTED DATABASES

Abstract: The main purpose of this paper is description of an approach of using data input correlation based on statistic of query and their frequency of occurrence in distributed databases. That approach is an alternative technique of reduction count of edges in graph. It also defines a direction and measure of dependence between particular elements. Described measure is used in determination of a partitioning criterium. This paper also presents a short characteristic a vertical fragmentation process based on statistic of query and development of a graphical partitioning algorithm which enable to solve disconnected graph problem.

Keywords: vertical fragmentation, distributed databases, graphical partitioning algorithm, matrix and graph of correlation

1. Wprowadzenie

Projektowanie rozproszonych baz danych dotyczy zwykle problemów fragmentacji, alokacji i replikacji. Jednakże, głównym celem rozproszenia jest poprawienie wydajności i zwiększenie niezawodności systemu. Pierwszy z dwóch aspektów jest szczególnie istotny, gdyż rozmieszczenie danych wynika często z natury rozproszonej organizacji i chęci uzyskania lokalnego wyszukiwania oraz przetwarzania danych. Ponieważ jednoczesne rozważanie powyższych zagadnień projektowania jest problemem złożonym (NP-trudnym), to w praktyce dąży się do ich rozdzielenia i szukania rozwiązania z wykorzystaniem heurystyk spełniających zadane kryteria.

Pierwszym etapem projektowania rozproszenia danych jest proces fragmentacji, definiowany jako podział pojedynczego zbioru atrybutów danej relacji na dwie lub więcej części w taki sposób, że połączenie tych części pozwala uzyskać oryginalną postać zbioru bez utraty jakiejkolwiek informacji [9]. Celem podziału jest uzyskanie minimalnego kosztu przetwarzania dla danego zbioru atrybutów, określonego przez funkcję celu oraz oszacowanie jakości uzyskanego podziału [1]. Ze względu na kryterium rozproszenia danych w procesie fragmentacji wyróżniamy podział: pionowy, poziomy i mieszany, będący syntezą dwóch poprzednich. Do najczęściej badanych zaliczamy podział pionowy, charakteryzujący się dużo większym stopniem skomplikowania niż podział poziomy.

Większość istniejących rozwiązań fragmentacji pionowej opiera się na modelu opisującym statystykę zapytań dostarczonych do systemu scentralizowanego, która określa empiryczne dane związane z typem i częstością ich wystąpienia. Podejście oparte na statystyce zapytań wykorzystuje głównie macierz pokrewieństwa jako dane wejściowe dla używanych algorytmów podziału. Do najczęściej stosowanych zaliczamy algorytmy oparte na teorii grafów, w których atrybuty i wyrazy macierzy pokrewieństwa określają wierzchołki oraz gałęzie grafu. Uproszczenie struktury grafu poprzez redukcję liczby gałęzi jest osiągane za pomocą technik modyfikacji macierzy pokrewieństwa lub alternatywnych podejść opartych na statystyce zapytań. Podejście przedstawione w artykule, oparte na zastosowaniu korelacji, która w naturalny sposób określa wzajemny związek między danymi wejściowymi, pozwala określić kierunek i miarę zależności pomiędzy atrybutami oraz może stanowić alternatywną technikę redukcji liczby gałęzi w grafie podziału.

Pozostała część artykułu wygląda następująco: w rozdziale 2 przedstawiono krótką charakterystykę procesu fragmentacji pionowej baz danych. W rozdziale 3 opisano koncepcję i właściwości podejścia, opartego na macierzy pokrewieństwa oraz algorytmie grafowym. W rozdziale 4 zaproponowano alternatywne podejście, oparte na macierzy korelacji, dalej w rozdziale 5 przedstawiono rozwinięcie algorytmu grafowego, umożliwiającego rozwiązanie problemu niespójności grafu, a w 6 zawarto wnioski i kierunki dalszej pracy.

2. Charakterystyka fragmentacji pionowej

Podział pionowy określany jest jako problem grupowania atrybutów relacji do pewnej liczby fragmentów zwanej schematem podziału. Schemat ten stanowi dane wejściowe dla procesu alokacji, jako następnego etapu projektowania rozproszenia danych. Powinien on dążyć do minimalizacji czasu wykonania aplikacji użytkownika, wykorzystującej otrzymane fragmenty. Duża liczba możliwych rozwiązań, określana liczbą Bella powoduje, że podział pionowy jest zbyt kosztowny przy wykorzystaniu tradycyjnych metod. Dla przykładu mając zbiór dziesięciu atrybutów otrzymujemy liczbę możliwych podziałów tego zbioru wynoszącą $B_{10} = 115\,975$, ale już dla piętnastu atrybutów wynosi ona $B_{15} = 1\,382\,958\,545$.

Tak szybki wzrost przestrzeni przeszukiwań powoduje konieczność zastosowania technik heurystycznych z określoną funkcją celu, a sam problem powinien być rozpatrywany w kontekście procesu optymalizacji. W potocznym znaczeniu technikami heurystycznymi nazywamy „niepełnowartościowe" algorytmy, umożliwiające w akceptowalnym czasie znalezienie przybliżonego (dostatecznie dobrego) rozwiązania badanego problemu. Efektywność tych algorytmów w procesie optymalizacji fragmentacji pionowej możemy określić w dwóch kategoriach, a mianowicie w jakości uzyskanego schematu podziału, wynikającego z funkcji celu oraz złożoności obliczeniowej danego algorytmu.

Ocena jakości uzyskanego schematu podziału, zaproponowana w pracy [1], określa minimalną wartość kosztu przetwarzania dla danego zbioru transakcji w zależności od częstości ich wystąpienia oraz fragmentacji elementów, do których transakcje uzyskują dostęp. Pozwala ona na porównanie i obliczenie „dobroci" uzyskanych schematów podziału dla różnych algorytmów, uwzględniając te same dane wejściowe. Umożliwia również zrównoważenie kosztu lokalnego i zdalnego dostępu do określonych atrybutów przez zadane transakcje.

Najlepsze rezultaty, zbliżone do rozwiązania optymalnego, uzyskują nowoczesne metaheurystyki [3], [5], ale ze względu na dużą złożoność np. zarządzanie populacją, cieszą się mniejszym zainteresowaniem niż algorytmy klasyczne. Uwzględniając kryterium złożoności obliczeniowej, algorytmy klasyczne i ich modyfikacje należą do najbardziej wydajnych. Do grupy tej możemy zaliczyć algorytmy takie, jak RBPA [8], GPA [10] i CBPA[4].

W wymienionych algorytmach jako dane wejściowe stosujemy obliczoną macierz pokrewieństwa, która może być wprowadzona bezpośrednio lub po dokonanej normalizacji.

3. Podejście oparte na macierzy pokrewieństwa

Większość algorytmów stosowanych dla podziału pionowego, opartych na statystyce zapytań jako danych wejściowych używa macierzy użycia atrybutów (ang. *Attribute Usage*

Matrix, AUM). Określa ona wzajemne powiązanie transakcji i atrybutów oraz częstość wystąpienia tych transakcji w badanym okresie. Komórki macierzy mogą przybierać następujące wartości:

$$AUM(T_i, A_j) = \begin{cases} 1 & \text{jeśli transakcja } T_i \text{ odwołuje się do atrybutu } A_j \\ 0 & \text{w przeciwnym przypadku} \end{cases} \quad (1)$$

Wyrazy macierzy określają odwołanie (uzyskanie dostępu) przez transakcje umieszczone w wierszach do poszczególnych atrybutów zawartych w kolumnach z pewną częstością. Przykładową postać macierzy AUM, rozważaną w [4], [8] i [10], przedstawiono w tabeli 1. Zawiera ona osiem transakcji odwołujących się do dziesięciu atrybutów z częstością określoną w kolumnie acc.

Tabela 1

Macierz użycia atrybutów dla przykładu 1

AUM	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	acc
T1	1	0	0	0	1	0	1	0	0	0	25
T2	0	1	1	0	0	0	0	1	1	0	50
T3	0	0	0	1	0	1	0	0	0	1	25
T4	0	1	0	0	0	0	1	1	0	0	35
T5	1	1	1	0	1	0	1	1	1	0	25
T6	1	0	0	0	1	0	0	0	0	0	25
T7	0	0	1	0	0	0	0	0	1	0	25
T8	0	0	1	1	0	1	0	0	1	1	15

Macierz pokrewieństwa atrybutów (ang. *Attribute Affinity Matrix*, AAM), zaproponowana w [6], bazuje na macierzy AUM i wektorze częstości acc. Określa ona wartości pokrewieństwa między atrybutami A_i i A_j relacji $R(A_1,A_2,\ldots,A_n)$ jako sumę częstości równoczesnego uzyskania dostępu do dowolnych dwóch atrybutów dla każdej transakcji. Wyrazy macierzy definiuje następujące wyrażenie:

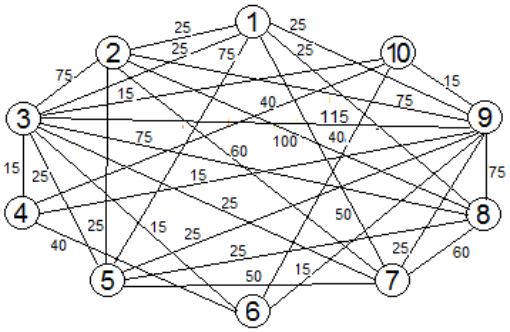
$$AAM_{ij} = \sum_{k|AUM(T_k,A_i)=1,\ AUM(T_k,A_j)=1} acc(T_k) \quad (2)$$

Przykładową macierz AAM, uzyskaną z powyższej macierzy AUM, przedstawiono w tabeli 2. Cechuje się ona symetryczną strukturą o największych wartościach występujących na przekątnej diagonalnej. Zarówno wiersze, jak i kolumny reprezentują atrybuty, a wyrazy macierzy należą do zbioru liczb całkowitych dodatnich ($N_+ \cup \{0\}$). Im większa wartość pokrewieństwa, tym częściej te same transakcje uzyskują dostęp do pary atrybutów, a co za tym idzie wzrasta możliwość wystąpienia danej pary atrybutów w tym samym fragmencie. Dla wartości bliskiej zeru para atrybutów nie ma transakcji wspólnych lub cechują się one niską częstością, a więc atrybuty powinny znaleźć się w różnych fragmentach.

Tabela 2

Macierz pokrewieństwa atrybutów (AAM) dla przykładu 1

AAM	1	2	3	4	5	6	7	8	9	10
1	75	25	25	0	75	0	50	25	25	0
2	25	100	75	0	25	0	60	100	75	0
3	25	75	115	15	25	15	25	75	115	15
4	0	0	15	40	0	40	0	0	15	40
5	75	25	25	0	75	0	50	25	25	0
6	0	0	15	40	0	40	0	0	15	40
7	50	60	25	0	50	0	85	60	25	0
8	25	110	75	0	25	0	60	110	75	0
9	25	75	115	15	25	15	25	75	115	15
10	0	0	15	40	0	40	0	0	15	40



Rys. 1. Graf pokrewieństwa uzyskany z AAM dla przykładu 1
Fig. 1. Affinity graph obtained from AAM for example 1

Na podstawie macierzy AAM zdefiniowano nieskierowany graf pokrewieństwa (rys. 1), w którym numery wierzchołków określają atrybuty, a krawędzie grafu zawierają wartość pokrewieństwa pomiędzy danymi wierzchołkami. Liczba krawędzi grafu wynosi różnicę pomiędzy liczbą krawędzi grafu pełnego, wyrażoną w postaci $n(n-1)/2$ a liczbą krawędzi o wartościach zerowych. Dla grafu z rys. 1 liczba krawędzi wynosi 45-15=30.

Opierając się na grafie pokrewieństwa realizowany jest klasyczny algorytm podziału atrybutów na fragmenty GPA, w którym najpierw dokonujemy linearyzacji grafu w postaci drzewa rozpiętości, a następnie wyszukujemy cykle identyfikujące fragmenty. Mimo że algorytm ten cechuje się najmniejszą złożonością obliczeniową $O(n^2)$, to uzyskana jakość podziału, określana funkcją celu, jest stosunkowo niska, zwłaszcza dla dużej liczby atrybutów [10]. Niska skalowalność algorytmu spowodowała powstanie alternatywnych podejść opartych na modyfikacji grafu. Dotyczą one głównie redukcji krawędzi grafu, uzyskanej poprzez normalizację macierzy pokrewieństwa oraz wprowadzenie dodatkowych parametrów ograniczających. Przykładem takiego algorytmu jest CBPA [4], w którym dokonujemy normalizacji macierzy do macierzy połączeń (ang. *connection matrix*), gdzie poszczególne wyrazy określają siłę połączenia między parą atrybutów. Następnie budujemy graf połączeń ograniczając liczbę krawędzi grafu poprzez przyjęcie arbitralnie minimalnego progu siły połączenia par atry-

butów, zwanego progiem akceptacji. Brak jednoznacznych reguł zdefiniowania minimalnej wartości progu akceptacji oraz określenie dodatkowych parametrów stanowi duży problem w doborze ich wielkości i wpływa na jakość podziału.

Zaproponowane poniżej podejście umożliwia redukcję krawędzi grafu bez parametru ograniczającego poprzez zastosowanie korelacji jako miary zależności między atrybutami. Dodatkowo, występowanie niespójności jest wykorzystywane do podziału grafu na mniejsze w fazie wstępnej.

4. Macierz i graf korelacji atrybutów

Macierz pokrewieństwa określa zależność między parami atrybutów jako sumę częstości równoczesnego uzyskania dostępu do tych atrybutów przez poszczególne transakcje. Ponieważ wyrazy macierzy należą do liczb całkowitych dodatnich, to nawet wystąpienie jednej transakcji odwołującej się do analizowanej pary atrybutów powoduje powstanie krawędzi w grafie. Mimo że niska wartość pokrewieństwa w stosunku do znacznie większych wartości innych krawędzi nie ma wpływu na schemat podziału, to zwiększa złożoność grafu. W celu uniknięcia nieistotnych krawędzi i redukcji złożoności grafu wykorzystano właściwości współczynnika korelacji Pearsona.

Współczynnik korelacji Pearsona ρ jest unormowaną miarą kierunku i siły zależności między dwoma zmiennymi [11]. Kierunek zależności może być dodatni lub ujemny, a siła przybiera wartości liczbowe z przedziału $-1 \le \rho \le 1$. Wykorzystując tę zależność generujemy symetryczną macierz korelacji atrybutów nazywaną dalej ACRM (ang. *Attribute Correlation Matrix*), określającą wzajemny związek pomiędzy atrybutami A_i i A_j macierzy AUM. Wyrazy macierzy mogą przybierać następujące wartości:

$$ACRM_{ij} = \begin{cases} 0 < \rho_{A_iA_j} \le 1, & \textit{jeśli zależność dodatnia} \\ -1 \le \rho_{A_iA_j} < 0, & \textit{jeśli zależność ujemna} \\ 0 & \textit{jeśli brak zależności} \end{cases} \tag{3}$$

Charakterystyczną cechą macierzy ACRM, a zarazem różnicą w stosunku do macierzy AAM jest możliwość wystąpienia zależności ujemnej między parą atrybutów. Oznacza to, że większość transakcji odwołuje się tylko do jednego atrybutu, a wspólne transakcje dla obydwu atrybutów charakteryzują się niską częstością występowania. W przypadku braku zależności żadna z transakcji nie odwołuje się do obydwu atrybutów jednocześnie. W obydwu powyższych przypadkach określona para atrybutów nie powinna znajdować się w tym samym fragmencie, a wyrazy macierzy ACRM, odnoszące się do tych atrybutów mogą zostać pominięte i nie będą uwzględniane w dalszej analizie. Takie założenie pozwala na redukcję

liczby rozpatrywanych wyrazów macierzy bez konieczności określenia empirycznego progu akceptacji, jak ma to miejsce w algorytmie CBPA. Macierz korelacji dla przykładu 1, obliczoną z macierzy AUM, przedstawiono w tabeli 3.

Tabela 3

Macierz korelacji atrybutów (ACRM) dla przykładu 1

ACRM	1	2	3	4	5	6	7	8	9	10
1	1	–0,07	–0,26	–0,45	1	–0,45	0,47	–0,07	–0,26	–0,45
2	–0,07	1	0,26	–0,45	–0,07	–0,45	0,47	1	0,26	–0,45
3	–0,26	0,26	1	0	–0,26	0	–0,26	0,26	1	0
4	–0,45	–0,45	0	1	–0,45	1	–0,45	–0,45	0	1
5	1	–0,07	–0,26	–0,45	1	–0,45	0,47	–0,07	–0,26	–0,45
6	–0,45	–0,45	0	1	–0,45	1	–0,45	–0,45	0	1
7	0,47	0,47	–0,26	–0,45	0,47	–0,45	1	0,47	–0,26	–0,45
8	–0,07	1	0,26	–0,45	–0,07	–0,45	0,47	1	0,26	–0,45
9	–0,26	0,26	1	0	–0,26	0	–0,26	0,26	1	0
10	–0,45	–0,45	0	1	–0,45	1	–0,45	–0,45	0	1

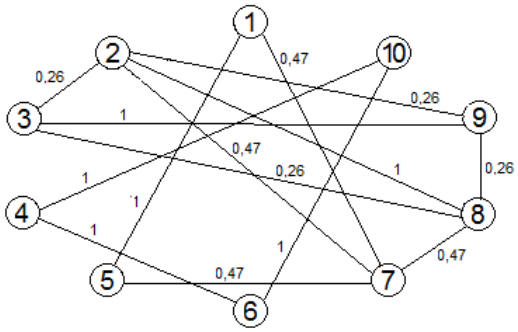
Na podstawie macierzy ACRM tworzony jest nieskierowany graf korelacji, podobny w koncepcji do grafu pokrewieństwa, przedstawionego na rys. 1. Wierzchołki określają poszczególne argumenty, a krawędzie – wyrazy macierzy o dodatniej zależności. Wszystkie rozpatrywane wyrazy macierzy ACRM zawarto w tabeli 4, a odpowiadający im graf przedstawiono na rys. 2. Przyjęcie wierzchołków tylko o dodatnim współczynniku korelacji pozwoliło zmniejszyć liczbę krawędzi grafu korelacji w stosunku do grafu pokrewieństwa z 30 do 14.

Tabela 4

Dodatnie wyrazy macierzy korelacji atrybutów

ACRM	1	2	3	4	5	6	7	8	9	10
1	1				1		0,47			
2		1	0,26				0,47	1	0,26	
3		0,26	1					0,26	1	
4				1		1				1
5	1				1		0,47			
6				1		1				1
7	0,47	0,47			0,47		1	0,47		
8		1	0,26				0,47	1	0,26	
9		0,26	1					0,26	1	
10				1		1				1

Nietrudno zauważyć, że liczba krawędzi w grafie (oprócz zastosowanego podejścia) zależy bezpośrednio od charakterystyki transakcji zawartej w wejściowej macierzy AUM. Tabela 5 przedstawia wyniki eksperymentu, określającego średnią liczbę krawędzi grafu w zależności od stopnia wypełnienia macierzy AUM dla różnych podejść, opartych na macierzy pokrewieństwa AAM, korelacji ACRM i połączeń ACM.



Rys. 2. Graf korelacji uzyskany z ACRM dla przykładu 1
Fig. 2. Correlation graph obtained from ACRM for example 1

Stopień wypełnienia macierzy SWM określamy jako stosunek liczby wyrazów dodatnich do wszystkich wyrazów macierzy. Przedział wartości SWM przyjęto na podstawie analizy przykładów z literatury, dla których wartość ta zawierała się w przedziale 25-40%. Poziom minimalnego progu akceptacji dla macierzy połączeń ACM przyjęto dla dwóch wartości, a mianowicie 0,2 i 0,4. Średnią liczbę krawędzi grafu LKG uzyskano z próby n=10000 losowań różnych macierzy AUM o zadanym rozmiarze i stopniu wypełnienia. Dodatkowo, zamieszczono średnią liczbę niespójnych grafów LNG otrzymanych dla próby n, w stosunku do stopnia wypełnienia macierzy AUM. Wartości średnie wyliczono na podstawie wielokrotnego powtórzenia pomiarów.

Na podstawie uzyskanych wyników przedstawionych w tabeli 5 możemy zauważyć, że podejście oparte na korelacji wykazuje dużą niezależność między spadkiem liczby krawędzi w grafie a procentowym spadkiem wypełnienia macierzy AUM. Dopiero przy dolnej granicy wypełnieniu (na poziomie 25%) liczba gałęzi w grafie dla poszczególnych podejść zaczyna się wyrównywać, co wynika ogólnie ze specyfiki transakcji, a mianowicie małej liczby odwołań do atrybutów. Przy wypełnieniu 45-50% dla macierzy AAM i ACM z przyjętym progiem akceptacji na poziomie 0,2 liczba krawędzi jest bliska liczbie krawędzi grafu pełnego, wynoszącej 45. Nawet zwiększenie wartości progu akceptacji do poziomu 0,4 i jednoczesne zmniejszenie liczby krawędzi do wartości 25-30 nadal stanowi około 60% grafu pełnego.

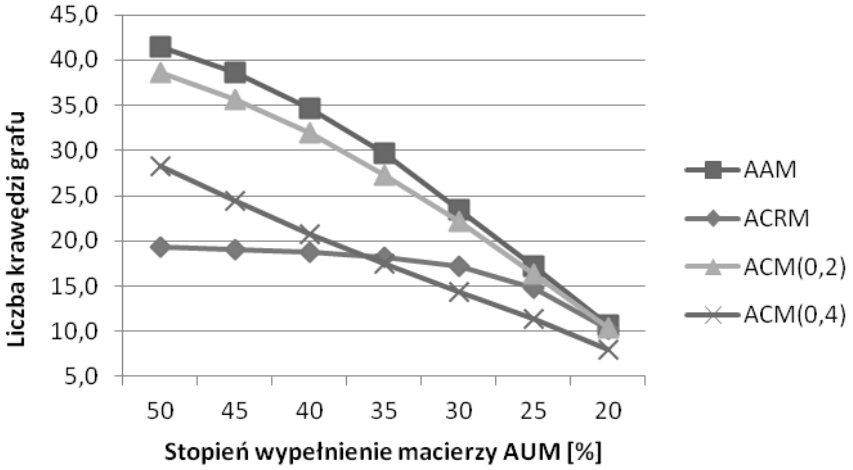
Tabela 5

Średnia liczba krawędzi grafu w zależności od stopnia wypełnienia macierzy AUM dla przykładu 1

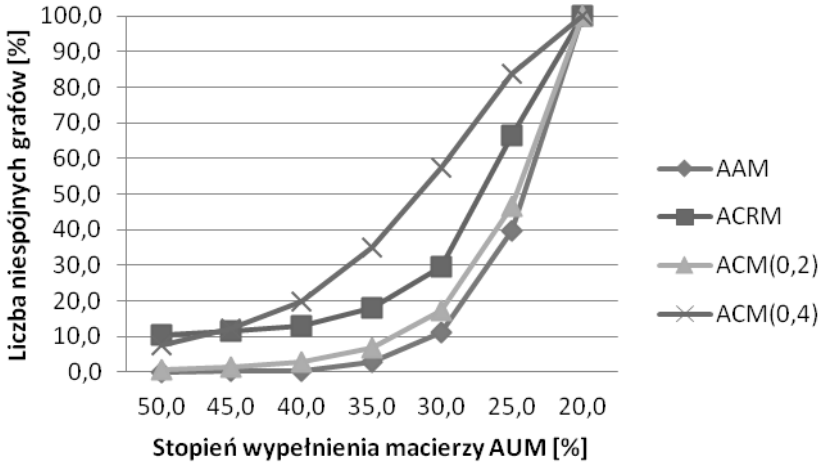
	AAM		ACRM		ACM (0,2)		ACM (0,4)	
SWM [%]	LKG	LNG [%]	LKG	LNG [%]	LKG	LNG [%]	LKG	LNG [%]
50	41,4	0,1	19,3	10,3	38,7	0,7	28,2	7,5
45	38,6	0,2	19,1	11,7	35,6	1,4	24,4	12,3
40	34,7	0,4	18,7	12,8	31,9	2,8	20,8	20,0
35	29,7	2,8	18,2	18,2	27,3	7,0	17,5	35,1
30	23,5	11,3	17,2	29,6	22,1	17,2	14,4	57,5
25	17,2	39,7	14,8	66,4	16,3	46,4	11,4	83,8
20	10,6	100,0	10,3	100,0	10,4	100,0	8,0	100,0

Zastosowanie współczynnika korelacji Pearsona, określającego zależności między atrybutami i uwzględnienie tylko jego dodatnich wartości pozwoliło zmniejszyć liczbę krawędzi w grafie do 20, co daje około 50% redukcję w stosunku do macierzy AAM i ACM. Dla stopnia wypełnienia 30-40% redukcja krawędzi nie jest już tak znaczna, ale nadal waha się w przedziale 20-40%. Zastosowanie korelacji jako miernika oceny zależności atrybutów pozwala również uniknąć problemu z empirycznym doborem parametrów, jak ma to miejsce w przypadku macierzy ACM.

Redukcja liczby krawędzi może prowadzić do niespójności grafu. Zależność wystąpienia niespójności w grafie od stopnia wypełnienia macierzy AUM przedstawiono na rys. 4.



Rys. 3. Liczba krawędzi grafu dla AAM, ACRM i ACM w zależności od stopnia wypełnienia macierzy AUM
Fig. 3. Count of graph edges for AAM, ACRM and ACM depending on the degree of fulfillment AUM



Rys. 4. Liczba niespójnych grafów w zależności od stopnia wypełnienia macierzy AUM dla n = 10 000
Fig. 4. Count of disconnected graphs depending on the degree of fulfillment AUM for n = 10 000

We wszystkich przypadkach zmniejszenie stopnia wypełnienia powoduje wzrost prawdopodobieństwa niespójności otrzymanych grafów. Ponad połowę niespójnych grafów uzyskano w przedziale 25-30% wypełnienia macierzy AUM, a 100% niespójności otrzymano dla 20% wypełnienia.

Ponieważ wystąpienie niespójności w grafie uniemożliwia zastosowanie grafowego algorytmu podziału w procesie fragmentacji, dlatego w dalszej części zaproponowano modyfikację algorytmu GPA.

5. Dwufazowy algorytm podziału oparty na korelacji

W celu rozwiązania problemu niespójności grafu korelacyjnego zaproponowano dwufazowy algorytm podziału, zwany dalej CRBPA (ang. *Correlation-Based Partitioning Algorithm*). Pierwsza faza, zwana wstępną wykorzystuje strukturę zbiorów rozłącznych (ang. *disjoin sets*) [2], pozwalającą na podział grafu korelacji na wzajemnie niezależne podgrafy. Etap ten realizowany jest iteracyjnie rozpatrując po kolei wszystkie krawędzie grafu. Rozpoczynając od jednoargumentowych zbiorów w każdej iteracji pobieramy jedną krawędź i wykonujemy operację złączenia zbiorów (unionSets), w których występują wierzchołki (argumenty) rozpatrywanej gałęzi. W ten sposób stopniowo zwiększamy zawartość zbiorów, zmniejszając jednocześnie ich liczbę. W przypadku wystąpienia niespójności, na koniec etapu otrzymujemy przynajmniej dwa zbiory, w których największy zawiera k atrybutów dla $k < n$. Złożoność obliczeniowa pierwszego etapu wynosi $O(m)$, gdzie m oznacza liczbę krawędzi grafu korelacji.

W fazie drugiej z uzyskanych wcześniej zbiorów generujemy niezależne podgrafy i dokonujemy ostatecznego podziału poprzez użycie klasycznego algorytmu GPA dla każdego podgrafu oddzielnie. Ponieważ algorytm GPA ma złożoność obliczeniową $O(n^2)$, gdzie n określa liczbę atrybutów, to wykonując go równolegle dla otrzymanych podgrafów zmniejszamy złożoność obliczeniową do $O(k^2)$, gdzie k określa liczbę atrybutów największego zbioru. Zatem uzyskana złożoność obliczeniowa całego algorytmu w przypadku wystąpienia niespójności wynosi $O(m+k^2)$ lub $O(m+n^2)$ w sytuacji odwrotnej.

Rozpatrując przykład 1 składający się z $n = 10$ atrybutów w etapie wstępnym uzyskujemy dwa podgrafy, złożone z następujących wierzchołków (4,6,10) i (1,2,3,5,7,8,9). Ponieważ liczba krawędzi uzyskana z macierzy ACRM wynosi $m = 14$, a liczba elementów największego zbioru wynosi $k = 7$, więc złożoność obliczeniowa algorytmu CRBPA jest mniejsza od GPA, a zatem $O(m+k^2) < O(n^2)$.

6. Wnioski

W artykule przedstawiono problem fragmentacji pionowej w procesie projektowania rozproszonych baz danych oraz zaproponowano podejście dotyczące zastosowania korelacji danych wejściowych opartych na statystyce zapytań i częstości ich występowania. W opisanym podejściu przedstawiono technikę redukcji liczby krawędzi w grafie korelacji, charakteryzującą się niskim poziomem wrażliwości na stopień wypełnienia macierzy danych wejściowych AUM.

Istniejąca możliwość wystąpienia niespójności w grafie korelacji zależy głównie od charakterystyki danych wejściowych i uniemożliwia wykorzystanie w procesie podziału klasycznego algorytmu GPA. Rozwiązanie problemu niespójności grafu uzyskano przez zmodyfikowanie algorytmu GPA o dodatkowy etap wstępny, oparty na strukturze zbiorów rozłącznych. Natomiast zmniejszenie złożoności algorytmu uzyskano przez równoległe przetwarzanie otrzymanych podgrafów z etapu wstępnego. Pozwoliło to na uzyskanie złożoności obliczeniowej algorytmu na poziomie $O(m+k^2) < O(n^2)$, gdzie $k < n$, w przypadku niespójnych grafów.

W ramach dalszych prac można przebadać zastosowanie korelacji w ocenie wpływu zależności opisującej liczność transakcji na schemat grupowania atrybutów oraz rozpatrzyć wpływ poziomu istotności na redukcję krawędzi w grafie.

BIBLIOGRAFIA

1. Chakravathy S., Muthuraj J., Varadarajan R., Navathe S.: An Objective Function for Vertically Partitioning Relations in Distributed Databases and its Analysis, Distributed and Parallel Databases, Vol. 2, No. 1, San Diego 1993, pp. 183-207.
2. Cormen T.H., Leiserson C.E., Rivest R.L., Stein C.: Introduction in Algorithms, Third Edition, The MIT Press, USA 2009.
3. Du J., Alhajj R., Barker K.: Genetic algorithms based approach to database vertical partitioning, Journal of Intelligent Information Systems, Vol. 26, Issue 2, 2006, pp. 167-183.
4. Du J., Barker K., Alhajj R.: Attraction – Global Affinity Measure for Database Vertical Partitioning, In proc. of ICWI, 2003, pp. 538-548.
5. Goli M., Raolnkoohi R., Taghi S.M.: A new vertical fragmentation algorithm based on ant collective behavior in distributed database systems. Knowledge and Information Systems, Vol. 30, 2012, pp. 435-455.
6. Hoffer A., Severance D.: The use of cluster analysis in Physical Database design, In. Proc. First Int. Conf. on very large Database, New York 1975.

7. Muthuray J., Chakravarthy S., Varadarajan R., Navathe S.: A Formal Approach to the Vertical Partitioning Problem in Distributed Database Design, n Technical Report. CIS Dept, University of Florida, 1993.
8. Navathe S.B., Ceri S., Wiederhold G., Dov J.: Vertical Partitioning Algorithms for Database Design, ACM Trans. On Database Systems, Vol. 9, No. 4, 1984, pp. 680-710.
9. Ozsu M.T., Valduriez P.: Principles of Distributed Database Systems, Second Edition, Prentice Hall, 1999.
10. Ra M., Navathe S.B.: Vertical partitioning and Database Design: A Graphical Algorithm, In Proceedings of the ACM SIGMOD International Conference on Management of Data, Portland 1989, pp. 440-450.
11. Sobczyk M.: Statystyka. Aspekty praktyczne i teoretyczne, Wydawnictwo UMCS, Lublin 2006.

Wpłynęło do Redakcji 15 marca 2013 r.

Abstract

The first stage of the design of the distribution of databasis is a fragmentation process. It defines a single attribute set partition into two or more independent parts without losing any information. The main goal of that partition is to obtain the minimum of cost set processing, which is specified by an objective function and estimation of the partition's quality obtained. Because of data distribution criterium in fragmentation process there is distinguished a vertical, horizontal and mixed fragmentation. The last one is a synthesis of the previous two. Among the most there is vertical one used fragmentation whose features are much more complicated than the horizontal one's.

Most of the existing solutions of vertical fragmentation are based on the model describing statistic of query, delivered to the stand-alone system. An approach based on statistic of query mainly uses an affinity matrix as data input for partitioning algorithms. The most commonly useful algorithms are based on graph theory, their attributes specify vertex and the cells of the matrix describe graph's edges. Simplification of a graph's structure by reduction of numbers of edges is obtained by modification of techniques of affinity matrix or by using alternative approaches based on statistic of query.

The main purpose of this paper is to describe an approach of using data input correlation based on statistic of query and their frequency of occurrence in distributed databases. That approach is an alternative technique of reduction count of edges in graph. It also defines

a direction and measure of dependence between particular elements. The described measure is used in determination of a partitioning criterium. A solution to a disconnected graph problem is obtained by modifying GPA algorithm by adding initial phase based on the structure of disjoin sets. Reduction of the complexity of the algorithm is achieved by parallel processing subgraphs which are obtained from the initial phase.

This paper also presents a short review of a vertical fragmentation process, affinity matrix and graph and also includes approach based on statistic query.

Adresy

Dariusz BERNATOWICZ: Politechnika Koszalińska, Wydział Elektroniki i Informatyki, ul. Śniadeckich 2, 75-453 Koszalin, Polska, dariusz.bernatowicz@tu.koszalin.pl
Anna BERNATOWICZ: Politechnika Koszalińska, Wydział Inżynierii Lądowej, Środowiska i Geodezji, ul. Śniadeckich 2, 75-453 Koszalin, Polska, anna.bernatowicz@tu.koszalin.pl