STUDIA INFORMATICA 2008
Volume 29 Number 2A (76)

Dariusz R. AUGUSTYN
Politechnika Śląska, Instytut Informatyki

ESTYMACJA SELEKTYWNOŚCI ZAPYTAŃ Z WYKORZYSTANIEM TRANSFORMATY KOSINUSOWEJ I FALKOWEJ

Streszczenie. Przeglądowy artykuł opisuje metody estymacji selektywności pewnej klasy zapytań ze złożonymi warunkami selekcji. Proste metody, wykorzystywane komercyjnie, zakładają pewne uproszczenie – niezależność wartości atrybutów tablic. Inne, dokładniejsze, bazują na estymacji wielowymiarowego rozkładu wartości atrybutów. Prezentowane, zaawansowane metody wykorzystują transformaty kosinusową i falkową dla efektywnego wyznaczania selektywności opierając się na stratnie skompresowanym widmie częstości wartości atrybutów.

Słowa kluczowe: selektywność zapytań w systemach zarządzania relacyjnymi bazami danych, estymacja wielowymiarowego rozkładu wartości atrybutu, rozkład macierzy na składowe główne, dyskretna transformata kosinusowa, dyskretna transformata falkowa

THE QUERY SELECTIVITY ESTIMATION USING COSINE AND WAVELET TRANSFORM

Summary. The paper is a literature survey of query selectivity methods for some kind of queries with composed selection conditions. Some simple commercially used methods base on the simplified assumption of table attribute value independence. The other methods presented in this paper are more accurate. They are based on the estimation of the multi-dimensional distribution of attribute values. Described methods use the discrete cosine transform and the discrete wavelet one for the effective query estimation based on the loss compressed spectrum of the attribute value frequency vector.

Keywords: query selectivity in relational database management systems, estimation of multi-dimensional distribution of attribute values, matrix singular value decomposition, discrete cosine transform, discrete wavelet transform

1. Metoda wyznaczania selektywności zapytań ze złożonym warunkiem selekcji, zakładająca niezależność atrybutów

Selektywność zapytania jest parametrem wyznaczanym w procesie optymalizacji zapytania. Wartość parametru jest liczbą z przedziału domkniętego [0, 1] i dla zapytań operujących na jednej tablicy, opisuje stosunek liczby wierszy spełniających kryterium zapytania, do wszystkich wierszy tablicy. Selektywność można też interpretować jako prawdopodobieństwo wylosowania pojedynczego wiersza, spełniającego kryteria zapytania ze zbioru wszystkich wierszy tablicy. Oszacowanie tej wartości, pozwalające na określenie przybliżonego rozmiaru wynikowego zbioru wierszy, jest wykorzystywane do wyboru najlepszego (optymalnego) sposobu realizacji zapytania. Selektywność zapytań, których warunek selekcji określony jest na jednym atrybucie, można wyznaczyć dzięki histogramowym metodom nieparametrycznej estymacji funkcji gęstości jednowymiarowego rozkładu prawdopodobieństwa wartości tego atrybutu (ogólny przegląd zawarty w [8]).

Wyznaczenie selektywności zapytania, w którym warunek selekcji określony jest na kilku atrybutach, wymaga estymacji wielowymiarowego rozkładu prawdopodobieństwa [1]. Jednak w większości komercyjnych zastosowań, przy szacowaniu selektywności takich zapytań, zakłada się niezależność atrybutów (reguła AVI ang. *attribute value independence*). Z twierdzenia, że prawdopodobieństwo iloczynu zdarzeń niezależnych jest iloczynem prawdopodobieństw tych zdarzeń, można oszacować selektywność zapytania Q o złożonym warunku za pomocą iloczynu selektywności prostych warunków następująco:

$$sel(Q(\Theta_1(R.X_1) \wedge ... \wedge \Theta_n(R.X_n))) = sel(Q_1(\Theta_1(R.X_1))).. sel(Q_n(\Theta_n(R.X_n))),$$

gdzie $R.X_i$ oznacza i-ty atrybut relacji/tablicy R, a Θ_i jest prostym warunkiem logicznym, określonym na tym atrybucie.

Formuła mnożenia selektywności prostych warunków pozwala na wykorzystanie jednowymiarowych metod estymacji selektywności. Metoda taka jest najczęściej stosowana przez optymalizatory komercyjnych serwerów baz danych. Jednak w praktyce założenie o niezależności atrybutów jest na ogół niespełnione i szacowanie bazujące na nim będzie obarczone dużym błędem, większym w odniesieniu do metod zaprezentowanych poniżej.

Poniżej prezentowane metody będą wykorzystywać wielowymiarową estymację funkcji gęstości rozkładu. Jednym z podstawowych problemów będzie minimalizacja rozmiaru danych potrzebnych do przechowywania parametrów, opisujących estymator rozkładu wielowymiarowego. Dane te przechowywane są w ramach tzw. słownika bazy danych (nazywanego inaczej metadanymi lub katalogiem systemowym b. d.).

2. Zastosowanie twierdzenia o rozkładzie macierzy według wartości osobliwych w wyznaczaniu selektywności

Dla zapytań, w których warunek selekcji dotyczy dwu atrybutów, można rozważyć metodę wyznaczania selektywności, wykorzystującą estymację dwuwymiarowej funkcji rozkładu, bazującą na dekompozycji macierzy częstości na składowe główne (rozkład wg wartości osobliwych, rozkład SVD – ang. *Singular Value Decomposition*) [1].

Dla omówienia tej i kolejnych metod przyjmiemy większość oznaczeń z podrozdziału 2.2 z [8]. Zamiast wektora częstości, w metodzie bazującej na rozkładzie SVD, będziemy mieli do czynienia z macierzą częstości F (dokładniej macierz liczności; słowo częstość zostało użyte dla zachowania zgodności z przyjętą anglojęzyczną nomenklaturą). Macierz tę można potraktować jak dwuwymiarowy histogram liczności typu equi-width [8], czyli opisujący rozkład dwóch atrybutów relacji (rozkład par atrybutów). Numery elementów w wymiarach (numery wierszy albo kolumn w F) opowiadają pewnym podprzedziałom pojedynczego atrybutu. W wyznaczaniu selektywności konkretnego zapytania uwzględniane są te elementy macierzy F, których wiersz i kolumna odnoszą się do podprzedziałów, spełniających warunki logiczne zapytania zakresowego dla pierwszego i drugiego atrybutu. Selektywność jest liczona jako suma tych elementów, podzielona przez sumę wszystkich elementów F.

Macierz częstości można wyrazić w rozkładzie SVD jako $F = USV^T$, gdzie elementy macierzy U i V zawierają się w przedziale [-1, 1], a macierz S jest macierzą diagonalną, posiadającą nieujemne elementy na głównej przekątnej. Ciąg niezerowych elementów przekątnej S jest na ogół nierosnący, a często bardzo szybko malejący do zera.

Macierz F można odtworzyć w następujący sposób:

$$F = \sum_{k=1}^{N} d_k C_U(k) R_V(k),$$

gdzie:

- N – rozmiar kwadratowej macierzy S,
- d_k – element głównej przekątnej macierzy S ($s_{kk} = d_k$),
- $C_U(k)$ – k-ta kolumna macierzy U,
- $R_V(k)$ – k-ty wiersz macierzy V.

Idea metody estymacji rozkładu dwuwymiarowego przez SVD polega na przybliżonym odtwarzaniu macierzy F, z wykorzystaniem jedynie istotnych elementów d_k, tzn. elementów o wartościach większych od przyjętej wartości progowej.

Przykładowo, dla

$$F = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \end{bmatrix},$$ po dekompozycji można uzyskać:

$$U = \begin{bmatrix} -0.1525 & -0.8226 \\ -0.3499 & -0.4214 \\ -0.5474 & -0.201 \\ -0.7448 & 0.3812 \end{bmatrix},\ S = \begin{bmatrix} 14.2691 & 0 \\ 0 & 0.6268 \end{bmatrix},\ V = \begin{bmatrix} -0.6414 & -0.7672 \\ 0.7672 & -0.6414 \end{bmatrix}.$$

Stąd, zaniedbując współczynnik d_2 o małej wartości (przybliżając jego wartość do 0), można uzyskać estymator:

$$\hat{S} = \begin{bmatrix} 14.2691 & 0 \\ 0 & 0 \end{bmatrix}.$$

Wykorzystując $\hat{S}$, szukany estymator częstości $\hat{F}$ można znaleźć następująco:

$$\hat{F} = d_1 C_u(1) R_v(1).$$

Stąd:

$$\hat{F} = 14.2691 \begin{bmatrix} -0.1525 \\ -0.3499 \\ -0.5474 \\ -0.7448 \end{bmatrix} \begin{bmatrix} -0.6414 & -0.7672 \end{bmatrix},\ F \approx \hat{F} = \begin{bmatrix} 1.3956 & 1.6692 \\ 3.2026 & 3.8306 \\ 5.0097 & 5.9919 \\ 6.8167 & 8.1533 \end{bmatrix}.$$

Dla wykonania estymacji $\hat{F}$ w powyższym przykładzie wymagane jest zapamiętanie jednego elementu macierzy *S*, jednej kolumny macierzy *U* i jednego wiersza macierzy *V*.

Kryterium dotyczące liczby zapamiętywanych elementów z przekątnej *S* może być adaptacyjne (wyznaczana wartość progowa, poniżej której wartości będą przybliżane zerem) albo liczba ta może być określona arbitralnie (metoda SVD-*k*, gdzie *k* to z góry zadana ilość elementów do uwzględnienia). W ogólnym przypadku, zamiast zapamiętywania dokładnej macierzy częstości *F* o rozmiarze $M \times N$, po zastosowaniu metody SVD-*k* dla uzyskania estymatora częstości $\hat{F}$, przechowywane będą: *k* elementów macierzy *S*, *k* kolumn macierzy *U* łącznie o rozmiarze $k \times M$ oraz *k* wierszy macierzy *V* łącznie o rozmiarze $k \times N$. Zakładając, że *k* jest małe ($k = 2, 3, 4$), przy dużych *M* i *N*, użycie tego typu estymacji *F* jest korzystne pod względem minimalizacji zajętości pamięci.

Przedstawiona metoda, chociaż pozwala na minimalizację zajętości pamięci i jest efektywna pod względem złożoności obliczeń (istnieje tzw. szybki algorytm obliczania SVD), ze względu na swoje ograniczone zastosowanie do dwóch wymiarów, nie jest implementowana w komercyjnych rozwiązaniach, ale stanowi przykład kierunku badań nad

estymatorami więcej niż jednowymiarowymi. Podobny mechanizm odrzucania nieistotnych elementów o małej wartości został wykorzystany w metodach, przedstawianych w dalszej części artykułu.

3. Estymacja selektywności z wykorzystaniem własności dyskretnej transformaty kosinusowej

Dyskretna transformata kosinusowa (DCT – ang. *discrete cosine transform*) wykorzystawana jest w dziedzinie przetwarzania sygnałów i obrazów. Jednym z jej zastosowań jest stratna kompresja danych (np. kompresja obrazów JPEG).

DCT może być wykorzystywana w procesie stratnej kompresji widma wielowymiarowego histogramu częstości [2]. W wyniku dziania DCT na wielowymiarowy histogram częstości F uzyskiwane jest widmo – wielowymiarowa transformata G. Dla większości rozkładów F (dane skorelowane) elementy istotne widma G skupione są w niewielkim, początkowym obszarze (dla indeksów o niskich wartościach). W ramach aproksymacji wyznaczane jest przybliżone widmo $\hat{G}$, przez zredukowanie do zera części widma G o małych wartościach modułu. Dla celów wyznaczania selektywności pamiętana jest tylko niezerowa część $\hat{G}$, stąd potencjalnie mała zajętość pamięci wymaganej do przechowywania $\hat{G}$.

Własności DCT pozwalają na szybkie wyznaczenie selektywności bezpośrednio na podstawie transformaty $\hat{G}$, czyli bez potrzeby wyznaczania odwrotnej transformaty kosinusowej (IDCT – ang. *inverse cosine transform*).

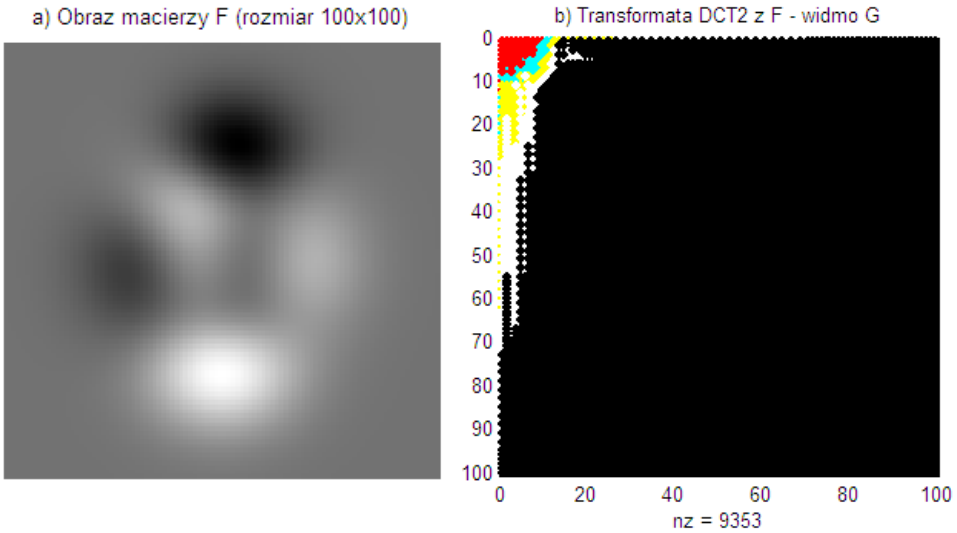
Na potrzeby wielowymiarowej estymacji rozkładu atrybutów w pracy [2] rozszerzono algorytm DCT na przypadek wielowymiarowy. W niniejszym opracowaniu metodę wykorzystującą DCT zilustrowano przez przypadek estymacji rozkładu dwuwymiarowego. Zakładając dwuwymiarową macierz częstości F o rozmiarze $M \times N$, poprzez zastosowanie dwuwymiarowej DCT, można uzyskać poszczególne współczynniki macierzy widma G następująco:

$$g(u,v) = \sqrt{\frac{2}{M}}k(u)\sum_{m=0}^{M-1}\left\{\sqrt{\frac{2}{N}}k(v)\sum_{n=0}^{N-1} f(m,n)\cos\left[\frac{(2n+1)v\pi}{2N}\right]\right\}\cos\left[\frac{(2m+1)u\pi}{2M}\right],$$

$$\text{gdzie } k(x) = \begin{cases} \frac{1}{\sqrt{2}} \text{ dla } x = 0 \\ 1 \quad \text{dla } x \neq 0 \end{cases}.$$

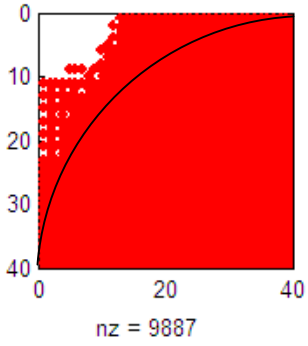
W przykładzie na rys. 1a zaprezentowano przybliżoną funkcję gęstości prawdopodobieństwa, uzyskaną przez interpolację i skalowanie wartości macierzy częstości F (jaśniejsze

kolory odpowiadają większym wartościom funkcji). Na rys. 1b pokazano wartości macierzy G – widmo uzyskane przez zastosowanie DCT. Poszczególne odcienie szarości odpowiadają zakresom wielkości modułów współczynników widma ($|g(u,v)|$ w podziale na podprzedziały wartości, których granice wyznaczają elementy zbioru $\{0.001, 0.01, 0.1, 1, 10\}$). Kolorem czarnym zaznaczono wartości bliskie zeru (tzn. mniejsze niż 0.001). Rysunek 1b pokazuje, że niemal cała moc przykładowego dwuwymiarowego sygnału F jest umiejscowiona w części widma G o niskich wartościach indeksów u i v (moduł 9353 współczynników z 10000 ogólnej liczby jest równy zero).



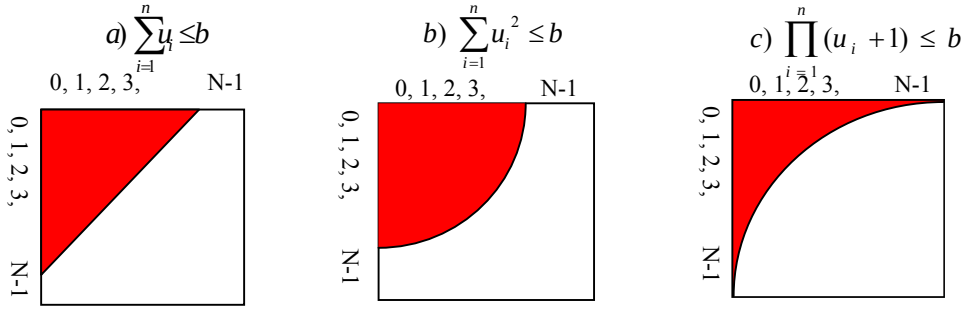
Rys. 1. Obrazy: a) wygładzonej macierzy częstości F, b) widma G
Fig. 1. The picture of: a) the smoothed frequency matrix F, b) the spectrum matrix G

Po przyjęciu pewnej wartości progowej, np. 0.1, przybliżając do zera niektóre współczynniki, np. te, dla których $|g(u,v)| < 0.1$, można uzyskać przybliżoną macierz widma $\hat{G}$ jak na rys. 2. Kolorem białym zaznaczono istotne współczynniki widma. Przy tak założonym progu w całej macierzy $\hat{G}$, 9887 współczynników z 10000 ma moduł wartości równy zeru.



Rys. 2. Macierz $\hat{G}$ – aproksymacja widma dla $|g| \geq 0.1$ (obszar biały); powiększony fragment widma dla $u,v \leq 40$
Fig. 2. The matrix $\hat{G}$ – the spectrum approximation for $|g| \geq 0.1$ (the white area); the zoomed spectrum partition for $u,v \leq 40$

W związku z często występującym, charakterystycznym rozkładem wielkości współczynników widma (na ogół skupionych w obszarze niskich wartości współczynników u i v), w pracy [2] rozważa się różne metody próbkowania widma, tzn. uwzględniania tylko istotnej jego części, przez wybór geometrycznego kształtu strefy – zakresu współczynników u i v (ang. *geometric sampling zone*). Kształty te rozważane są w pracy [2] dla przypadku wielowymiarowego. Dla przypadku dwuwymiarowego, niektóre kształty stref próbkowania pokazano na rys. 3, gdzie u_i opisuje i-ty wymiar, a parametr b jest przyjętą wartością progową, wpływającą na rozmiar strefy.



Rys. 3. Przykładowe typy stref próbkowania widma: a) trójkątna, b) sferyczna, c) odwrotnie proporcjonalna
Fig. 3. Examples of sampling zones types of the spectrum: a) triangle, b) spherical, c) reciprocal

W pracy [2] rozważa się kilka typów stref próbkowania: prostokątne, trójkątne, sferyczne, odwrotnie proporcjonalne. Analiza ilościowa (potwierdzająca wynik pokazany na rys. 2) wykazała, że najbardziej efektywne jest próbkowanie na podstawie strefy odwrotnie proporcjonalnej, które dla danych skorelowanych pozwala na uzyskanie bardzo dobrego przybliżenia widma $\hat{G}$, przy możliwie małej liczbie przechowywanych próbek (współczynników).

Jednak podstawową zaletą omawianego podejścia, jak już zostało wspomniane, jest wykorzystanie możliwości wyznaczenia selektywności bezpośrednio na podstawie $\hat{G}$ („całkowanie widma"), tzn. bez potrzeby wykonywania IDCT. Dla przypadku dwuwymiarowego selektywność może być wyznaczona następująco:

$$sel(Q(a < R.X < b \wedge c < R.Y < d)) = \int_c^d \int_a^b f(x,y)dxdy$$

$$= \int_c^d \sqrt{\frac{2}{M}} \sum_{u=0}^{M-1} k(u) \left\{ \int_a^b \sqrt{\frac{2}{N}} \sum_{v=0}^{N-1} k(v) g(u,v) \cos xv\pi \, dx \right\} \cos yu\pi \, dy$$

$$\approx \sqrt{\frac{2}{M}} \sqrt{\frac{2}{N}} \sum_{(u,v)\in Z} k(u)k(v)g(u,v) \int_c^d \cos(u\pi y)dy \int_a^b \cos(v\pi x)dx,$$

gdzie Z oznacza przyjętą strefę próbkowania, tzn. $(u,v) \notin Z \Rightarrow g(u,v) \equiv 0$.

Kolejną zaletą użycia DCT w omawianym zastosowaniu jest możliwość wykorzystania liniowej własności transformaty, przy realizacji aktualizacji współczynników widma, tzn.

własności: $\mathrm{DCT}(\alpha X + \beta Y) = \alpha\,\mathrm{DCT}(X) + \beta\,\mathrm{DCT}(Y)$, gdzie X, Y to sygnały, a α, β wartości skalarne. Dzięki temu aktualizacja widma po zmianie sygnału F (histogramu częstości F) może się odbywać z wykorzystaniem przyrostów wartości F, tzn. $\Delta G = DCT(\Delta F)$ (nie ma potrzeby wyznaczania transformaty G na podstawie całego F, w przypadku jego częściowej modyfikacji).

4. Wykorzystanie dyskretnej transformaty falkowej do wyznaczania selektywności zapytań

Podobnie jak DCT, dyskretna transformata falkowa (DWT – ang. *discrete wavelet transform*) wykorzystywana jest w dziedzinie przetwarzania sygnałów i obrazów (np. kompresja obrazów JPEG 2000).

DWT pozwala, przez wielorozdzielczą reprezentację wektora skumulowanych częstości, na efektywne obliczanie selektywności [3, 4]. W niniejszym podrozdziale, dla przejrzystości opisu, omówione zostanie zastosowanie DWT z użyciem rozwinięcia Haara w wyznaczaniu selektywności dla przypadku jednowymiarowego. Oczywiście, istnieje uogólnienie zaprezentowanej techniki na przypadek wielowymiarowy i z wykorzystaniem innych rodzajów funkcji bazowych [3, 4].

Dla wektora częstości F (estymującego, z dokładnością do współczynnika proporcjonalności, funkcję gęstości prawdopodobieństwa) można zdefiniować wektor skumulowanych częstości F^+ (estymującego, z dokładnością do współczynnika proporcjonalności, dystrybuantę), tzn.:

$$F = [\, f_j : j = 0..\, N-1\,] \;\Rightarrow\; F^+ = [\, c_l : c_l = \sum_{j \le l} f_j, l = 0..\, N-1].$$

Przykładowo, dla F = [2 0 1 1 1 0 1 2] wektor F+= [2 2 3 4 5 5 6 8].

Tabela 1 przedstawia sposób tworzenia falkowego widma wektora F^+. Dla najwyższej rozdzielczości wartości średnie odpowiadają oryginalnemu sygnałowi, czyli wektorowi F^+, natomiast detale nie są określone. Dla pozostałych rozdzielczości wartości $a_{i,k}$ uzyskiwane są przez obliczenie średniej z $a_{i+1,2k}$ oraz $a_{i+1,2k+1}$ – wartości z poziomu rozdzielczości $i + 1$. Wartości detali $d_{i,k}$ liczone są jako różnica pomiędzy wartościami średnimi z sąsiednich poziomów, tzn. $a_{i+1,2k} - a_{i,k}$.

Tabela 1

Analiza wielorozdzielcza wektora skumulowanych częstości F^+

Skala (rozdzielczość) i	Wartości średnie $a_{i,k}$								Współczynniki szczegółowe (detale) $d_{i,k}$			
3	2	2	3	4	5	5	6	8	-	-	-	-
2	2		3.5		5		7		0	-0.5	0	-1
1	2.75				6				-0.75		-1	
0	4.375								-1.625			

Szukaną transformatę, tzn. widmo falkowe G dla przykładowego F^+ można wyrazić wzorem: DWT($[c_0, c_1, .., c_7]$)=DWT($[a_{3,0}, a_{3,1}, .., a_{3,7}]$) = $[a_{0,0}, d_{0,0}, d_{1,0}, d_{1,1}, d_{2,0}, d_{2,1}, d_{2,2}, d_{2,3}]$ = $[g_0, g_1, .., g_7]$, czyli DWT([2 2 3 4 5 5 6 8]) = [4.375 -1.625 -0.75 -1 0 -0.5 0 -1] (elementy wchodzące w skład widma zostały zaznaczone w tabeli 1 kolorem szarym).

Oryginalne wartości sygnału (tzn. wektora F^+ w omawianym przykładzie) mogą być łatwo wyznaczone z widma, wyrażonego przez drzewo dekompozycji (ang. *error tree*), którego przykład został pokazany na rys. 4. Przejścia drogą od korzenia do liścia pozwalają na wyznaczenie oryginalnego sygnału. W węzłach drzewa odpowiednio umieszczone są wartości współczynników widma. Wartości pojedynczych składowych oryginalnego sygnału c_k ($a_{3,k}$) z poziomu liści drzewa mogą być dynamicznie wyznaczane na podstawie przejścia (nie są faktycznie przechowywane w ramach drzewa).

Przyjmując oznaczenia:

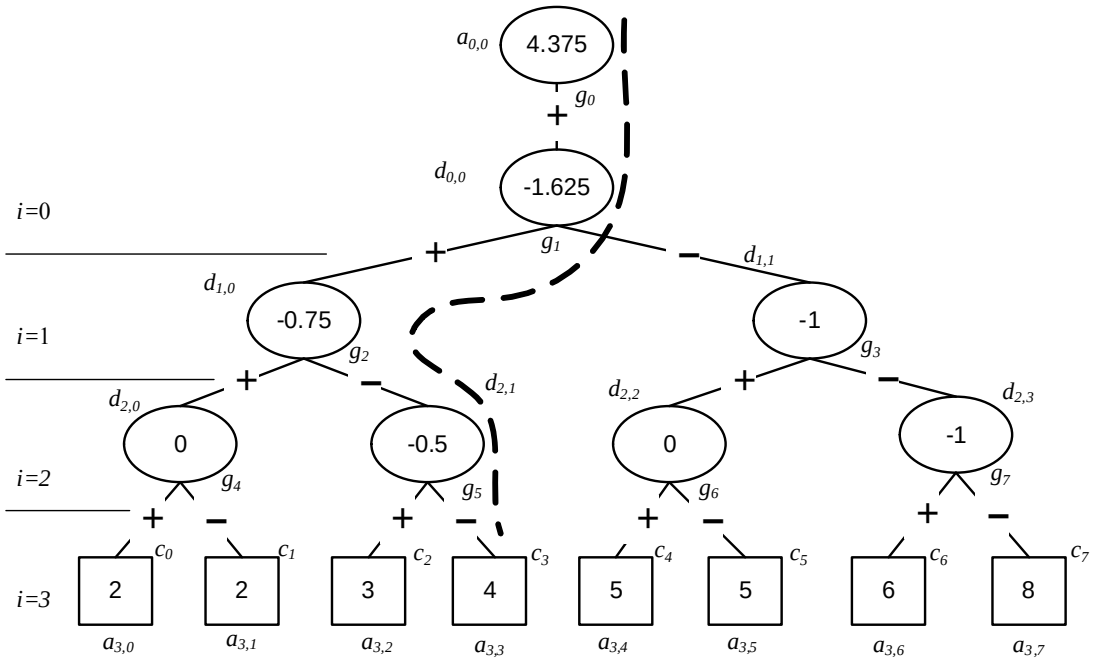
- c_k – k-ta składowa sygnału oryginalnego F^+,
- g_j – węzeł drzewa dekompozycji falowej (j-ta składowa widma G),
- $path_k$ – ścieżka (droga) w drzewie dekompozycji, o początku w korzeniu, pozwalająca na wyznaczenie c_k (lista indeksów pewnego podciągu G); np. $path_3$ = [0, 1, 2, 5],
- *leftleaves*(c_i)/ *rightleaves*(c_i) – funkcja zwracająca te węzły, które występując w roli korzenia tworzą poddrzewo, zawierające element c_i w lewej/prawej części tego poddrzewa; np.: *leftleaves*(c_3) = {g_1}, *leftleaves* (c_1) = {g_1, g_2}, *rightleaves*(c_7) = {g_1, g_3, g_7}

wartość składową można wyznaczyć następująco:

$$c_k = \sum_{j \in path_k} \delta_{k,j} g_j ,$$

gdzie $\delta_{k,j} = \begin{cases} 1 \text{ gdy } g_j \in leftleaves(c_k) \vee j = 0, \text{czyli } g_j \text{ jest korzeniem,} \\ -1 \text{ gdy } g_j \in rightleaves(c_k). \end{cases}$

Przykładowo, wyznaczanie c_3 odbywa się następująco: $c_3 = a_{3,3} = g_0 + g_1 - g_2 - g_5 = a_{0,0} + d_{0,0} - d_{1,0} - d_{2,1} = 4$ (odpowiednia droga zaznaczona została na rys. 4 linią przerywaną).



Rys. 4. Drzewo dekompozycji falkowej przykładowego wektora skumulowanych częstości
Fig. 4. The error tree for the wavelet decomposition of the example cumulative frequency vector

Selektywność prostych, jednotablicowych zapytań zakresowych np. $sel(Q(v_i \le R.X \le v_j))$, gdzie v_i i v_j są elementami V – wektora unikalnych wartości X (definicja V zgodna z oznaczeniami z [8]), może być wyliczana na podstawie widma falkowego G jako $(c_{i-1} - c_j)/c_{N-1}$, gdzie wartości c_{i-1} i c_j wyznaczane są na podstawie przejść przez drzewo dekompozycji, a c_{N-1} jest równe liczbie wierszy/krotek tablicy/relacji R.

Ze względu na potrzebę minimalizacji zajętości pamięci przyjmuje się pewną aproksymację $\hat{G}$ widma falkowego G, przez pominięcie w wektorze G elementów mało istotnych (co sprowadza się do przybliżenia zerem elementów G o małej wartości). W przypadku przyjęcia pewnego warunku aproksymacji $|g_j| \le 0.5 \Rightarrow \hat{g}_j \equiv 0$, przybliżenie przykładowego widma G będzie następujące $\hat{G}$ = [4.375 -1.625 -0.75 -1 0 0 0 -1], a zrekonstruowany na podstawie $\hat{G}$ przybliżony wektor skumulowanych częstości będzie wyrażony następująco $\hat{F}^+$ = [2 2 3.5 3.5 5 5 6 8]. Jednak taki algorytm pomijania nie uwzględnia faktu różnego zakresu wpływu poszczególnych elementów widma na rekonstruowany wektor sygnału oryginalnego.

Algorytm eliminacji mało istotnych elementów G powinien, dla danego, potencjalnie usuwanego elementu, uwzględniać ilość składowych oryginalnego wektora F, na które dany

element ma wpływ (co wynika z wysokości położenia tego współczynnika w drzewie dekompozycji). Łatwo zauważyć, że elementy wyżej położone w drzewie obejmują swym zakresem wpływu większą liczbę składowych *F*.

Na potrzeby ilościowej oceny skutków ewentualnego pominięcia elementów w drzewie dekompozycji (zerowania wartości) wprowadza się wskaźnik L_2Error – błąd kwadratowy rekonstrukcji, określony jako:

$$L_2Error = \sqrt{\frac{1}{N}(F^+ - \hat{F}^+)(F^+ - \hat{F}^+)^T} = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1}(c_i - \hat{c}_i)^2} \ .$$

Dla przykładowego, ośmioelementowego wektora *G*, przy pominięciu elementu g_1 błąd wyniesie $|g_1|$, odpowiednio przy g_3 wyniesie $\frac{|g_3|}{\sqrt{2}}$, a przy g_6 wyniesie $\frac{|g_6|}{2}$. Uogólniając, zmiana wartości w węźle *i*-tego poziomu drzewa dekompozycji wpływa na błąd L_2Error z wagą $\frac{1}{\sqrt{2^i}}$. Wektor *G* po normalizacji uwzględniającej wagi $\frac{1}{\sqrt{2^i}}$, dla składowych będących węzłami *i*-tego poziomu, oznaczony jest symbolem G_{norm}. Dla przykładowego widma *G*, wektor znormalizowany ma następującą postać:

$$G_{norm} = [4.375 \ \ -1.625 \ \ -\frac{0.75}{\sqrt{2}} \ \ -\frac{1}{\sqrt{2}} \ \ 0 \ \ -0.25 \ \ 0 \ \ -0.5].$$

Najprostszą metodą uzyskania $\hat{G}$ (zmniejszonego w odniesieniu do aproksymowanego widma *G*), zawierającego zadaną liczbę *b* « N elementów istotnych (najistotniejszych), jest wybór *b* elementów o największej wartości modułu z wektora G_{norm}. Jest to metoda optymalna pod względem kryterium minimalizacji wskaźnika L_2Error.

Ostatecznie, selektywność oszacowywana jest na podstawie składowych przybliżonego wektora skumulowanych częstości $\hat{F}^+$, wyznaczanych dynamicznie na podstawie wektora $\hat{G}$, czyli *b*-elementowej aproksymacji widma.

5. Podsumowanie

W niniejszym przeglądowym artykule zaprezentowano podstawowe cechy metod estymacji selektywności z zastosowaniem transformaty kosinusowej i falkowej. Opisane metody wykorzystują własności DCT i DWT w sposób zbliżony do użycia tychże transformat w dziedzinie przetwarzania sygnałów i obrazów. W obu przypadkach metoda wyznaczania selektywności bazuje na stratnie skompresowanym widmie albo częstości, albo skumulowanej częstości wystąpień wartości atrybutów. Zaletą zaprezentowanych metod jest możli-

wość oszacowywania selektywności wprost na podstawie aproksymowanego widma, bez potrzeby wyznaczania odwrotnej transformaty całego widma.

W artykule opisano sposób oszacowywania selektywności pewnej klasy zapytań – zapytań jednotablicowych, z warunkiem selekcji będącym iloczynem zakresowych warunków logicznych dla kilku atrybutów tablicy. Prawidłowe oszacowanie selektywności takich zapytań wymaga wyznaczenia i wykorzystania estymatora wielowymiarowej funkcji gęstości. Jednak w większości komercyjnych zastosowań, tzn. w algorytmach zaimplementowanych w ramach modułu optymalizatora zapytań serwera b.d., bazuje się na założeniu o niezależności atrybutów (AVI). Na ogół dla rzeczywistych zbiorów danych założenie AVI jest najczęściej niespełnione (dane są skorelowane) i metoda bazująca na nim (mnożenie selektywności, wyznaczonych dla prostych warunków składowych) jest w oczywisty sposób obarczona błędem. Stąd, w przyszłych implementacjach serwerów b. d. prawdopodobnie należy spodziewać się częstszego wykorzystania metod opartych na estymacji wielowymiarowej funkcji gęstości atrybutów, np. tych, bazujących na DCT lub DWT.

Obecnie, podobnie jak w dziedzinie przetwarzania obrazów, większą uwagę poświęca się badaniom metod opartych na własnościach DWT niż DCT. Niezależnie od sposobów tutaj zaprezentowanych, od lat rozwijane są też inne metody estymacji rozkładów wielowymiarowych na potrzeby oszacowywania selektywności, np. MHIST [1, 5, 6], GENHIST [5, 6] czy STHOLES [7].

LITERATURA

1. Possala V., Ioannidis Y. E.: Selectivity Estimation without the Attribute Value Independence Assumption. Proc. of the 23rd Int. Conf. on Very Large Databases, The VLDB Journal, Athens 1997.
2. Lee. J., Deok-Hwan K., Chin-Wan Ch.: Multi-dimensional Selectivity Estimation Using Compressed Histogram Estimation Information. Proc. of ACM SIGMOD Int. Conf. on Management of Data. ACM, Philadelphia 1999.
3. Matias Y., Vitter J. S., Wang M.: Wavelet-Based Histograms for Selectivity Estimation. Proc. of ACM SIGMOD Int. Conf. on Management of Data. ACM, Washington 1998.
4. Garofalakis M.: Wavelet-Based Approximation Techniques in Database Systems. Exploratory DSP. Signal Processing Magazine, IEEE vol. 23 no. 6, 2006.
5. Gunopulos D., Kollios G., Tsotras V. J.: Approximating Multi-Dimensional Aggregate Range Queries Over Real Attributes. ACM SIGMOD 2000, Dallas 2000.

6. Gunopulos D., Kolios G., Tsotras V. J., Domeniconi C.: Selectivity estimators for multidimensional range queries over real attributes. The international Journal on Very Large Data Bases. The VLDB Journal, vol. 14 no. 2, Springer Berlin / Heidelberg 2005.
7. Bruno N., Chanchuri S., Gravano L.: STHoles: A Multidimesional Workload-Aware Histogram. Proc. of the 2001 ACM SIGMOD International Conference on Management of Data, Santa Barbara 2001.
8. Augustyn D. R.: Szacowanie selektywności zapytań z wykorzystaniem metod estymacji jednowymiarowych rozkładów prawdopodobieństwa. Studia Informatica, vol. 29 no. 2 (76), Gliwice 2008.

Recenzent: Dr inż. Maciej J. Bargielski

Wpłynęło do Redakcji 6 grudnia 2007 r.

Abstract

The paper contains a survey of query selectivity methods that use properties of the discrete cosine transform (DCT) and the discrete wavelet one (DWT). The article mainly concentrates on queries with composed selection conditions defined on many table attributes.

The paper explains the attribute values independence assumption (the AVI rule), which is base of most algorithms implemented in commercial database query optimizer modules.

Often, for a real-life data, this assumption is false, so the other methods that use an estimation of multi-dimensional attribute values distributions can give a better accuracy. Presented DCT and DWT methods use a spectrum representation of the multi-dimensional distribution of attribute values. The DCT method uses the loss compressed spectrum of the attribute value frequency vector for the selectivity estimation. The DWT one uses loss compressed spectrum of the cumulated frequency vector. The possibility of the estimating selectivity which can be directly calculated using the approximate spectrum (without calculating inverse transform for the whole spectrum) is the only one of many advantages of shown methods.

Adres

Dariusz Rafał AUGUSTYN: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska, draugustyn@polsl.pl.