STUDIA INFORMATICA 2011
Volume 32 Number 2A (96)

Dariusz R. AUGUSTYN
Politechnika Śląska, Instytut Informatyki

METODA ANALIZY GŁÓWNYCH SKŁADOWYCH W SZACOWANIU SELEKTYWNOŚCI ZAPYTAŃ

Streszczenie. Selektywność zapytania jest parametrem pozwalającym określić spodziewany rozmiar wyniku zapytania. Oszacowanie selektywności wymagane jest do wyznaczania optymalnego sposobu realizacji zapytania. Zadaniem tym zajmuje się moduł optymalizatora SZBD. Obliczanie selektywności jest szczególnie utrudnione w zapytaniach z warunkami wieloatrybutowymi, gdzie potrzebny jest nieparametryczny estymator wielowymiarowego rozkładu wartości atrybutów. Zastosowanie wielowymiarowego histogramu w takiej roli może być zbyt kosztowne pod względem zajętości pamięci, szczególnie w przypadku wysokiej wymiarowości zagadnienia. W takiej sytuacji użyteczne może być podejście wykorzystujące metodę analizy składowych głównych, redukujące wymiarowość. Dodatkowo można zastosować metodę mnożenia selektywności, wyznaczonych niezależnie z jednowymiarowych rozkładów brzegowych, określonych w zredukowanej przestrzeni. Upraszcza to i przyspiesza przedstawioną w artykule metodę szacowania selektywności.

W artykule opisano również sposób implementacji zaproponowanego rozwiązania w SZBD Oracle, z wykorzystaniem modułu rozszerzającego działanie optymalizatora zapytań – Oracle Data Cartridge Interface Statistics.

Słowa kluczowe: selektywność zapytań, reprezentacja wielowymiarowego rozkładu, histogram, redukcja wymiarowości, analiza głównych składowych, rozszerzenie optymalizatora zapytań, Oracle ODCIStats

PRINCIPAL COMPONENT ANALYSIS IN QUERY SELECTIVITY ESTIMATION

Summary. Query selectivity allows to estimate the size of query results. It is required for obtaining the optimal method of query execution. This is a main goal of a query optimizer activities. Selectivity calculations for queries with a complex multi-attribute selection condition require a non-parametric estimator of multi-dimensional probability density function of distribution of table attribute values. Using a multi-dimensional histogram as a representation of multi-dimensional distribution is very space-consuming for high dimensions. The approach based on Principal Component

Analysis allows to reduce dimensionality and makes the representation space-efficient. Additionally the attribute value independence rule (with multiplicity of simple selectivities) may be used in a dimensions-reduced space so the method of the PCA-based selectivity estimation becomes simpler and more effective. The paper also presents the implementation of the proposed solution in DBMS Oracle as the extension of the query optimizer by using Oracle Data Cartridge Interface Statistics module.

Keywords: query selectivity, representation of multidimensional distribution, histogram, dimensionality reduction, Principal Component Analysis, extension of query optimizer, Oracle ODCIStats

1. Wstęp

Proces efektywnej realizacji zapytania wymaga wcześniejszego opracowania sposobu jego realizacji, zwanego planem wykonania (ang. *execution plan*). Zadanie to realizuje tzw. kosztowy optymalizator zapytań, który przy opracowywaniu uwzględnia spodziewany rozmiar danych, spełniających kryteria analizowanego zapytania. Selektywność zapytania jest parametrem określającym, jaki ułamek całości stanowi liczba wierszy zbioru wynikowego. Dla zapytań jednotablicowych selektywności jest to stosunek liczby wierszy spełniających warunek selekcji do liczby wszystkich wierszy tablicy.

W niniejszym artykule rozważana jest klasa zapytań, w której zakresowy warunek selekcji jest określony na atrybutach z ciągłą dziedziną wartości. Dla takich zapytań $Q_{1Dim}(a_{MIN} \leq T.A \leq a_{MAX})$ selektywność można określić jako:

$$sel(Q) = \int_{a_{MIN}}^{a_{MAX}} f_A(x)dx,$$

gdzie A – atrybut tablicy T, f_A – funkcja gęstości prawdopodobieństwa rozkładu wartości atrybutu A.

Estymacja selektywności, realizowana na wstępnym etapie przetwarzania zapytania, wymaga więc estymatora funkcji gęstości rozkładu wartości atrybutów. Najczęściej w roli nieparametrycznego estymatora funkcji gęstości wykorzystywany jest histogram.
Problem komplikuje się jeśli rozważymy zapytania wieloatrybutowe – $Q_{NDim}\left(\bigcap_{i=1}^{N}\left(a_{i,MIN} \leq T.A_i \leq a_{i,MAX}\right)\right)$. Wówczas selektywność można określić wzorem:

$$sel(Q_{NDim}) = \int_{a_{1,MIN}}^{a_{1,MAX}} \dots \int_{a_{N,MIN}}^{a_{N,MAX}} f_{A_1,\dots A_N}(x_1,\dots,x_N)dx_1 \dots dx_N,$$

gdzie $A_1 \dots A_N$ – atrybuty tablicy T, $f_{A_1,\dots,A_N}$ – funkcja gęstości prawdopodobieństwa łącznego rozkładu atrybutów.

Wyznaczanie selektywności na podstawie wielowymiarowych histogramów (estymujące $f_{A_1,\ldots,A_N}$) może być niekorzystne ze względu na dużą zajętość pamięci w przypadku wysokich wymiarowości (dużych wartości *N*).

W takich sytuacjach należy zastosować podejście redukujące efekt, który wynika z wymiarowości zagadnienia, co oczywiście odbywa się pewnym kosztem dokładności oszacowania selektywności.

Znane są podejścia, w których proponuje się oszczędną, pod względem zajętości, przybliżoną reprezentacją wielowymiarowego rozkładu wartości atrybutów, takich, jak np. wykorzystujące: wielowymiarowy estymator jądrowy [3], widmo dyskretnej transformaty cosinusowej [4], widmo transformaty falkowej [5], algorytm GenHist [3], STHOLES [6] czy sieci Bayesa [7].

Jednym z podejść do omawianego problemu jest koncepcja wykorzystująca analizę głównych składników (analiza komponentów głównych – ang. *Principal Component Analysis* – PCA [1]) do redukcji wymiarowości. W uproszczeniu, metoda PCA polega na takim obrocie układu współrzędnych w wielowymiarowej przestrzeni, aby zmaksymalizować wariancję dla pierwszej współrzędnej, potem dla drugiej itd. Metoda szereguje nowe współrzędne w kolejności malejących (nierosnących) wariancji, pozwalając na wybór współrzędnych znaczących (czyli pominięcie tych nieistotnych) – stąd redukcja wymiarowości nowej reprezentacji.

Artykuł będzie pokazywał zastosowanie PCA w metodzie szacowania selektywności zakresowych zapytań ze złożonym, wieloatrybutowym warunkiem selekcji.

2. Uzasadnienie wykorzystania metody PCA – prosty przykład użycia pamięcioooszczędnej reprezentacji wielowymiarowego rozkładu wartości atrybutów

W analizowanym poniżej przykładzie zakłada się, że zapytanie będzie dotyczyć tablicy *T*, zawierającej co najmniej atrybuty a_1, a_2, a_3, a_4 z ciągłą dziedziną wartości. Rozważane zapytania zakresowe, dotyczące 4-wymiarowej przestrzeni $A_1 \times \ldots \times A_4$, będą miały następującą postać:

$$(a_{1MIN} \leq a_1 \leq a_{1MAX} \wedge \ldots \wedge a_{4MIN} \leq a_4 \leq a_{4MAX}),$$

czyli będą dotyczyć pewnego, 4-wymiarowego hiperprostopadłościanu. Przykład posłuży do ilustracji pamięcioooszczędnej reprezentacji łącznego rozkładu prawdopodobieństwa wartości atrybutów $a_1,\ldots,a_4$, wykorzystanej do stosunkowo dokładnego szacowania selektywności zapytań *Q*.

2.1. Dane przykładowe w postaci próby losowej wierszy tablicy bazy danych i statystka opisująca rozkład w postaci klasycznego wielowymiarowego histogramu

Zwyczajowo, w terminologii baz danych, reprezentację rozkładu atrybutów wykorzystywaną w procesie optymalizacji zapytań nazwa się (niezbyt poprawnie) statystyką. Taka statystyka (na ogół w postaci histogramu) może być sporządzona na podstawie całego zbioru danych (np. ANALYZE TABLE … COMPUTE STATISTICS w SZBD Oracle [11]) albo (ze względu na zamiar skrócenia procesu) na podstawie części danych, czyli pewnej reprezentatywnej próby (np. ANALYZE TABLE … ESTIMATE STATISTICS SAMPLE … ROWS).

W klasycznym podejściu zastosowanie wielowymiarowego histogramu, jako reprezentacji łącznego rozkładu wartości atrybutów może okazać się pamięciowo niepraktyczne (tzw. problem eksplozji wymiarów).

W omawianym przykładzie, ze względu na czytelność, przyjęto bardzo niską rozdzielczość histogramu – *hres* = 4 na każdy z 4 wymiarów. Do zapamiętania całej statystyki trzeba będzie zaalokować strukturę danych o następującym rozmiarze (mierzonym zajętością pojedynczej liczby):

- $2 \times 4 = 8$ (definicja histogramu – dwie liczby na każdy wymiar: punkt początkowy b_i, długość podprzedziału d_i, gdzie i oznacza indeks wymiaru) plus
- $4 \times 4 \times 4 \times 4 = 256$ (liczba wszystkich komórek histogramu),

co łącznie daje zajętość równą aż 264.

Przyjmijmy, że statystyka jest wykonywana tylko na podstawie części danych z T, tzn. pewnego $n = 7$[1] elementowego zbioru wierszy T^* z wartościami atrybutów, jak w tabeli 1.

Tabela 1

Wartości atrybutów w podzbiorze wierszy tablicy T, stanowiących próbę losową T^*, wykorzystaną do tworzenia statystyk opisujących łączny rozkład $A_1 \times A_2 \times A_3 \times A_4$

a_1	a_2	a_3	a_4
4	2	4	15
8	4	1	15
16	8	9	34
21	14	50	89
34	15	31	75
51	28	34	117
63	38	20	135

Dla przykładowych danych z tabeli 1, definicja klasycznego 4-wymiarowego histogramu (z czterema podprzedziałami o równej długości w każdym wymiarze) jest następująca:

$$((b_i, d_i): i = 1,...,4) = ((4, 14.75), (2, 9), (1, 12.25), (15, 30))$$

[1] Tak mała wartość n (liczność próby) została przyjęta jedynie w celu zapewnienia czytelności przykładu. W rzeczywistości można się spodziewać, że n będzie o rzędy większe.

2.2. Wykorzystanie metody PCA do uzyskania reprezentacji danych o zredukowanej wymiarowości

Zastosowanie metody PCA dla zbioru danych T^* pozwala na wyznaczenie następujących elementów:

- macierz S – zbiór wierszy odpowiedników dla wierszy T^* w nowym układzie współrzędnych, tzn. przestrzeni komponentów głównych,
- macierz współczynników komponentów głównych C – określa liniowe, wzajemnie jednoznaczne odwzorowanie pomiędzy S i T^*; każda kolumna macierzy C jest wektorem własnym macierzy kowariancji wartości T^* i zawiera współczynniki dla wybranego komponentu; kolejność kolumn w macierzy C wynika z malejącej (nierosnącej) wariancji komponentu,
- wektor L – wektor wartości własnych macierzy kowariancji wartości T^* lub inaczej wektor wariancji kolumn macierzy S; wartości własne są uporządkowane malejąco (nierosnąco); uporządkowanie odpowiadających komponentów głównych jest takie, jak to, które wynika z uporządkowania kolumn w macierzy C; wektor L może być wykorzystany do ilościowej oceny istotności komponentów głównych, a tym samym pozwala na selekcję istotnych (znaczących) kolumn macierzy C.

Przyjmijmy, że ET^* jest poziomym wektorem średnich wartości w kolumnach T^*. Elementy ET^*_j można wyznaczyć następująco:

$$ET^*_j = \frac{1}{n}\sum_{i=1}^{n} T^*_{ij} . \tag{1}$$

Oznaczmy przez ETT^* macierz, której wiersze to wektory ET^*, takie że elementy ETT^* można wyrazić następująco:

$$\underset{i=1\ldots n}{\forall} ETT^*_{ij} = ET^*_j . \tag{2}$$

Zależność:

$$T^* = ETT^* + S\,C^T \tag{3}$$

pozwala wyznaczyć T^* na podstawie S.

Zależność:

$$S = (T^* - ETT^*)(C^T)^{-1} \tag{4}$$

pozwala wyznaczyć S na podstawie T^*.

Wartości w S i C, wyznaczone metodą PCA dla wartości macierzy T^*, przedstawiono w tabeli 2.

Dla danych zawartych w T^*, wyznaczony wektor wartości własnych macierzy kowariancji L wynosi [3113.3, 225.8, 10.065, 0.104]. Widać, że dwa pierwsze elementy tego wektora są znacząco większe od pozostałych. W typowych zastosowaniach metody PCA wybiera się kilka pierwszych elementów wektora L tak, żeby ich suma stanowiła pewien przyjęty procent

(np. 95%) sumy wszystkich elementów. W rozpatrywanym przykładzie podzbiór złożony z dwóch pierwszych elementów L spełnia z nadmiarem to umowne kryterium (suma obu stanowi 99,7% całkowitej sumy).

Tabela 2

Wartości z przestrzeni komponentów głównych – S, współczynniki komponentów głównych – C

s_1	s_2	s_3	s_4
-62.6297	-2.9053	2.1165	0.5912
-61.3290	-7.7176	0.3170	-0.3634
-39.0996	-4.6226	-0.810	-0.3122
21.0680	28.9995	3.2368	-0.1248
9.8262	6.2544	-5.8852	0.1366
56.4000	-0.3866	-1.6686	0.0862
75.7641	-19.6218	2.6935	-0.0137

S

0.3780	-0.4293	-0.8195	0.0340
0.2258	-0.2438	0.1936	-0.9231
0.2210	0.8691	-0.3638	-0.2518
0.8702	0.0290	0.3982	0.2887

C

Takie podejście pozwala na zredukowanie wymiarowości problemu z 4 do 2 (uwzględnienie tylko dwóch znaczących komponentów głównych). Macierz $\hat{S}$ zawiera dane w nowej, tylko 2-wymiarowej przestrzeni komponentów. $\hat{S}$ zbudowana jest jedynie z kolumn s_1 i s_2 macierzy S. Użycie $\hat{S}$ (zamiast S) pozwoli na redukcję zajętości pamięci przez strukturę danych przechowującą statystykę, bez istotnej utraty dokładności odwzorowywanego T^*.

Przyjmijmy oznaczenie $\hat{S}_{zero-padded}$, odpowiednik $\hat{S}$, macierz o wymiarach 4×4, gdzie dwie pierwsze kolumny to s_1 i s_2, a dwie pozostałe są wypełnione zerami. Wówczas wartości danych $\hat{T}^*$ (przybliżone wartości T^*), wyznaczone z $\hat{S}$ (przybliżenia S), będzie można uzyskać następująco:

$$\hat{T}^* = ETT^* + \hat{S}_{zero-padded} C^T . \tag{5}$$

Błąd względny wyniku odwzorowania wykonanego na podstawie $\hat{S}$ (wzór 5) w stosunku do tego, wykonanego na podstawie S (wzór 3) zdefiniowany jest następująco:

$$RErrT = \frac{\left|\hat{T}^* - T^*\right|}{T^*} 100\% . \tag{6}$$

Wartości $\hat{T}^*$ i $RErrT$ dla danych z omawianego przykładu pokazane zostały w tabeli 3.

Tabela 3

Przybliżone wartości danych – $\hat{T}^*$, wartości błędu względnego [%] – $RErrT$

a_1	a_2	a_3	a_4
5.7144	2.1360	4.9189	13.9866
8.2722	3.6032	1.0238	14.9787
15.3468	7.8687	8.6267	34.4127
23.6569	13.2581	51.1462	87.7472
29.1723	16.2655	28.8933	77.3039
49.6296	28.4026	33.4146	117.6395
65.2079	37.4659	20.9765	133.9315

$\hat{T}^*$

30.0014	6.3680	18.6810	7.2458
3.2904	11.0126	2.3284	0.1423
4.2565	1.6692	4.3275	1.1992
11.2310	5.5955	2.2410	1.4277
16.5491	7.7800	7.2915	2.9803
2.7613	1.4176	1.7518	0.5436
3.3859	1.4256	4.6552	0.7978

$RErrT$

Wartości $\hat{T}^*$ nieznacznie odbiegają od wartości oryginalnych T^*. Wartość średnia błędu względnego wynosi około 5,8%, co subiektywnie można uznać za zdecydowanie dopuszczalne, zważywszy, że docelowo chodzi o uzyskiwanie (na podstawie $\hat{S}$) przybliżonych wartości selektywności.

Przestrzeń $A_1 \times A_2 \times A_3 \times A_4$ będzie nazywana dalej przestrzenią podstawową. Przestrzeń o zredukowanej liczbie wymiarów, tzn. $S_1 \times S_2$, będzie dalej nazywana przestrzenią zredukowaną.

2.3. Transformacja warunku selekcji zapytania na potrzeby metody wyznaczania selektywności opartej na metodzie PCA

Przy wyznaczaniu selektywności (w omawianym przykładzie) zakresowy warunek selekcji dla przestrzeni podstawowej, oparty na 4-wymiarowym hiperprostopadłościanie, można zastąpić warunkiem selekcji dla przestrzeni 2-wymiarowej, opartym na prostokącie.

Dla przykładu rozważmy następujące zapytanie:

$$Q(2 \le a_1 \le 22 \wedge 3 \le a_2 \le 15 \wedge\ 0 \le a_3 \le 60\ \wedge\ 14 \le a_4 \le 90). \tag{7}$$

Granicom hiperprostopadłościanu, określonego warunkiem selekcji, odpowiadają wektory [2 3 0 14] i [22 15 60 90]. Korzystając ze wzoru (4) można obliczyć, że wektorowi z przestrzeni podstawowej [2 3 0 14] odpowiada następujący wektor z przestrzeni zredukowanej:

$$([2\,3\,0\,14] - ET^*)(C^T)^{-1} = [\text{-}64.9142 \quad \text{-}5.7960 \quad 5.0062 \quad 0.3187]. \tag{8}$$

Wektorowi z przestrzeni podstawowej [22 15 60 90] odpowiada następujący wektor z przestrzeni zredukowanej:

$$([22\ 15\ 60\ 90] - ET^*)(C^T)^{-1} = [24.7522 \quad 37.0467 \quad \text{-}0.6291 \quad \text{-}3.2433]. \tag{9}$$

W wektorach uzyskanych we wzorach 8 i 9, ze względu na redukcję wymiaru przestrzeni, oczywiście istotne są tylko dwa początkowe elementy. Stąd ostatecznie można stwierdzić, że warunkowi selekcji zapytania Q będzie odpowiadać następujący warunek selekcji zapytania $\hat{Q}$ z przestrzeni zredukowanej:

$$\hat{Q}([-64.9142 \le s_1 \le 24.7522\ \wedge\ -5.7960 \le s_2 \le 37.0467\ ). \tag{10}$$

2.4. Wykorzystanie histogramu jako reprezentacji rozkładu wartości w zredukowanej przestrzeni

Jako nową reprezentację rozkładu (oszczędną pod względem zajętości pamięci) można przyjąć 2-wymiarowy histogram, zbudowany na podstawie wartości ze zredukowanej przestrzeni. Biorąc pod uwagę dane z S (tabela 1), a dokładniej minima i maksima w każdej kolumnie s_1 i s_2 oraz zakładając dotychczasowo przyjętą rozdzielczość w każdym wymiarze $h_{res} = 4$, odpowiedni, dwuwymiarowy histogram miałby następującą definicję:

$((b_1, d_1), (b_2, d_2)) = ((\text{-}62.6297, 34.5985), (\text{-}19.6218, 12.1553))$.

Zajętość takiego histogramu wynosi:

- 2×2 = 4 (do przechowania definicji histogramu) plus
- 4×4 = 16 (do przechowywania wartości w komórkach histogramu),

co w sumie daje 20 liczb do zapamiętania (wynik w sposób oczywisty jest znacznie lepszy od wyniku, dotyczącego histogramu budowanego z danych w przestrzeni podstawowej, opisanego w podrozdziale 2.1).

Przedstawiony sposób reprezentacji rozkładu, chociaż efektywny w zakresie oszczędności pojemności wymaganej do przechowania dwuwymiarowego histogramu, może zostać jeszcze poprawiony, co zostanie pokazane w kolejnym podrozdziale.

2.5. Zastosowanie reguły AVI w przestrzeni zredukowanej – przyjęcie założenia dotyczącego niezależności zmiennych z przestrzeni zredukowanej

Optymalizatory komercyjnych serwerów baz danych, estymując selektywność zapytań ze złożonymi warunkami selekcji, zbudowanymi na kilku atrybutach, stosują regułę AVI (ang. *attribute values independence*). Reguła ta opiera się na uproszczonym (a najczęściej wręcz niespełnionym) założeniu, że atrybuty są względem siebie niezależne, tzn. zakłada się, że dla warunku złożonego selektywność jest równa iloczynowi selektywności warunków prostych (na zasadzie wyznaczania prawdopodobieństwa iloczynu zdarzeń niezależnych). Na ogół serwery baz danych umożliwiają jedynie tworzenie i aktualizację statystyk dla pojedynczych atrybutów w postaci jednowymiarowych histogramów. Histogramy te wykorzystywane są do wyznaczania selektywności predykatów prostych (określonych na jednym atrybucie).

Dokładne wyznaczenie selektywności warunku złożonego wymaga zastosowania nieparametrycznego estymatora funkcji gęstości prawdopodobieństwa wielowymiarowego rozkładu wartości atrybutów, jak np. wielowymiarowego histogramu.

Histogram opisujący rozkład wartości w przestrzeni zredukowanej $S_1 \times S_2$ stanowi efektywną zajętościowo, ale przybliżoną reprezentację rozkładu wartości w przestrzeni $A_1 \times A_2 \times A_3 \times A_4$.

Macierz kowariancji zmiennych S_1 i S_2 jest macierzą z wartościami nieujemnymi jedynie na głównej przekątnej (poza główną przekątną wszystkie elementy są równe zero). Przykładowo, dla wartości S z omawianego przypadku (tabela 2) kowariancja wynosi:

$$\operatorname{cov}(S_1, S_2) = \begin{bmatrix} 3112.3 & 0 \\ 0 & 225.8 \end{bmatrix}. \tag{11}$$

Kowariancja zmiennych niezależnych jest równa zero. Twierdzenie odwrotne nie jest prawdziwe.

W istniejących optymalizatorach wyznaczenie selektywności dotyczące np. warunku selekcji, zbudowanego na atrybutach a_1 i a_2 najczęściej sprowadzałoby się do wyznaczenia iloczynu selektywności dla predykatu z a_1 i selektywności dla predykatu z a_2, mimo iż

zmienne A_1 i A_2 z pewnością nie są niezależne, jak pokazuje to poniższa macierz kowariancji (niezawierająca zer):

$$\mathrm{cov}(A_1, A_2) = \begin{bmatrix} 493.1429 & 287.7381 \\ 287.7381 & 172.6190 \end{bmatrix}. \tag{12}$$

Skoro regułę AVI praktycznie stosuje się dla zależnych atrybutów (i daje to akceptowalne wyniki!), tym bardziej można dopuścić jej zastosowanie dla atrybutów nieskorelowanych (co nie dowodzi ich niezależności). Takie podejście pozwala jeszcze bardziej uprościć sposób reprezentacji rozkładu. Zamiast wielowymiarowego histogramu, dotyczącego rozkładu wartości w przestrzeni zredukowanej (w przykładzie – histogramu dwuwymiarowego), można z powodzeniem używać histogramów jednowymiarowych dla każdej zmiennej z przestrzeni zredukowanej (w przykładzie – dwa histogramy jednowymiarowe). Wówczas estymacja selektywności będzie odbywać się na podstawie reguły AVI, ale zastosowanej dla zmiennych w przestrzeni zredukowanej.

Dla omawianego przykładu histogramy jednowymiarowe, z $h_{res} = 4$ podprzedziałami, są następująco zbudowane:

- $hist_{S1}$ – histogram dla zmiennej S_1: $(b, d) = (-62.6297, 34.5985)$,
- $hist_{S2}$ – histogram dla zmiennej S_2: $(b, d) = (-19.6218, 12.1553)$.

Ostatecznie zajętość proponowanej statystyki będzie wynosić:

- 2 + 4 = 6 (definicja i wartości w histogramie dla zmiennej S_1) plus
- 2 + 4 = 6 (definicja i wartości w histogramie dla zmiennej S_2),

czyli 12.

Dodatkowo, na potrzeby wyznaczania transformacji wartości z przestrzeni podstawowej do przestrzeni zredukowanej, wymagane jest przechowywanie wektora wartości średnich ET^* (czteroelementowego, w omawianym przykładzie).

Ostatecznie uzyskano wynik równy 16, zdecydowanie lepszy niż 264 (dla klasycznego histogramu, opisującego bezpośrednio rozkład wartości w przestrzeni podstawowej).

Uogólniając, zajętość zaproponowanej statystyki jest sumą następujących składników:

- N – rozmiar wektora ET^* (w przykładzie $N = 4$; N to liczba atrybutów tablicy T, której dotyczy złożony warunek selekcji zapytania Q),
- $2 * K$ – definicje każdego z K histogramów jednowymiarowych (w przykładzie $K = 2$; $K \leq N$ lub nawet $K \ll N$; K to liczba znaczących komponentów głównych, wyznaczona na podstawie wektora L),
- $h_{res} * K$ – liczba komórek we wszystkich K jednowymiarowych histogramach (w przykładzie $h_{res} = 4$; h_{res} to liczba podprzedziałów w histogramie).

Uogólniając, zajętość statystyki opartej na klasycznym, wielowymiarowym histogramie utworzonym dla wartości z przestrzeni podstawowej jest sumą następujących składników:

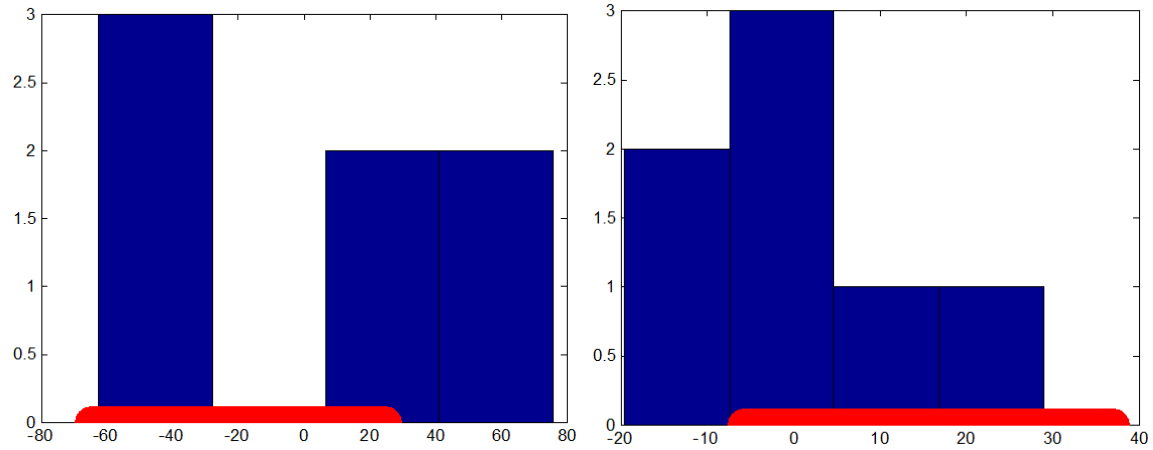
- $2 * N$ – definicja każdego z N wymiarów histogramu,

- h_{res}^N – liczba wszystkich komórek N-wymiarowego histogramu.

Ostatecznie, zajętość pamięci dla zaproponowanej „statystyki" jest rzędu O($K\ h_{res}$), natomiast dla „statystyki klasycznej" zajętość jest rzędu O(h_{res}^N), co eliminuje tę ostatnią z zastosowań dla dużych N. Przybliżona „statystyka", o małej zajętości, chociaż oparta na histogramach jednowymiarowych, dobrze opisuje rozkład wielowymiarowy na potrzeby wyznaczania selektywności zapytań ze złożonym, zakresowym warunkiem selekcji.

2.6. Przykład i ocena estymacji selektywności zapytania opartej na metodzie PCA

Sposób wyznaczania selektywności zostanie zilustrowany z użyciem przykładowego zapytania Q, zdefiniowanego wzorem 7, w odniesieniu do danych zawartych w T*. Wiersze z T* (sztuk 3), spełniające warunek zapytania Q, w tabeli 1 zostały zaznaczone kolorem szarym. Łatwo obliczyć, że selektywność dokładna sel_exact(Q) wynosi.



a) b)

Rys. 1. Jednowymiarowe histogramy dla wartości z przestrzeni zredukowanej (kolor niebieski)[2] oraz przedziały warunków selekcji (kolor czerwony): a) histogram $hist_{S1}$ i warunek selekcji $-64.9142 \leq s_1 \leq 24.7522$, b) histogram $hist_{S2}$ i warunek selekcji $-5.7960 \leq s_2 \leq 37.0467$

Fig. 1. 1-dimensional histograms for values of reduced space (blue color) and selection condition intervals (red color): a) histogram $hist_{S1}$ and selection condition $-64.9142 \leq s_1 \leq 24.7522$, b) histogram $hist_{S2}$ and selection condition $-5.7960 \leq s_2 \leq 37.0467$

Selektywność przybliżona sel_approx(Q) zostanie obliczona po przetransformowaniu warunku zapytania Q do przestrzeni zredukowanej, tzn. po zbudowaniu nowego warunku dla zapytania $\hat{Q}$. Selektywność ta jest równa sel_approx($\hat{Q}$) i będzie iloczynem selektywności:

$$\text{sel}(-64.9142 \leq s_1 \leq 24.7522) * \text{sel}(-5.7960 \leq s_2 \leq 37.0467), \quad (13)$$

wyznaczonych niezależnie, na podstawie każdego z histogramów $hist_{S1}$ i $hist_{S2}$.

[2] Artykuł dostępny jest z rysunkami w kolorze na portalu zeszytów Studia Informatica: http://znsi.aei.polsl.pl/.

Rysunek 1 przedstawia histogramy $hist_{S1}$ i $hist_{S2}$ oraz przedziały wartości dla s_1 i s_2, wynikające z warunku selekcji zapytania $\hat{Q}$.

Histogramy $hist_{S1}$ i $hist_{S2}$ zawierają bezwzględną liczbę wystąpień w czterech podprzedziałach (sumaryczna liczba wystąpień wynosi n = 7). Tradycyjnie zakłada się, że rozkład wartości wewnątrz podprzedziału jest równomierny.

$\text{sel}(-64.9142 \leq s_1 \leq 24.7522)$ wyznaczona zostanie na podstawie 1. i 3. podprzedziału w $hist_{S1}$ (w 2. podprzedziale liczba wystąpień równa jest 0 i dlatego zostanie on pominięty). Pierwszy podprzedział [-62.6297 -28.0313) pokryty jest w całości przedziałem warunku selekcji i do wyrażenia na selektywność wnosi składnik $\frac{3}{7} \approx 0.4286$. Trzeci podprzedział [6.5672 41.1656) pokryty jest częściowo. Jego uwzględnienie wnosi do wyrażenia składnik $\frac{(24.7522 - 6.5672)}{34.5985}\frac{2}{7} \approx 0.1502$, stąd $\text{sel}(-64.9142 \leq s_1 \leq 24.7522)$ wynosi 0.4286 + 0.1502 = 0.5788. Podobnie można wyznaczyć selektywność $\text{sel}(-5.7960 \leq s_2 \leq 37.0467)$ na podstawie $hist_{S2}$, która wynosi 0.5322.

Ostatecznie, na podstawie wzoru 13 (zastosowanie reguły AVI), otrzymujemy:

$$\text{sel\_approx}(\hat{Q}) = 0.5788 * 0.5322 = 0.3080. \tag{14}$$

Błąd względny oszacowania selektywności wynosi:

$$RErrSel = \frac{\left|\text{sel\_approx}(\hat{Q}) - \text{sel\_exact}(Q)\right|}{\text{sel\_exact}(Q)} 100\% \approx 28\%, \tag{15}$$

co (w subiektywnej ocenie autora) i tak jest dobrym rezultatem, zważywszy na bardzo niską rozdzielczość histogramów (h_{res} = 4).

3. Zastosowanie praktyczne – możliwość rozszerzenia funkcjonalności optymalizatora zapytań SZBD Oracle

Zaprezentowany sposób wyznaczania selektywności możliwy jest do implementacji jako rozszerzenie standardowej funkcjonalności optymalizatora zapytań w SZBD Oracle, dzięki możliwości użycia modułu ODCIStats (ang. ***O**racle **D**ata **C**artridge **I**nterface **Stat**istics*) [10].

Rozszerzenie:

- umożliwi wywołanie funkcji (nazywanej dalej *create_user-defined_statistics*) do tworzenia i utrwalania w bazie danych własnej statystyki, opartej na koncepcji PCA,
- zapewni, że optymalizator zapytań może skorzystać ze zdefiniowanej przez użytkownika funkcji (nazywanej dalej *user-defined_selectivity_function*), przeznaczonej do wyznaczania selektywności, w fazie przygotowania zapytania (jeśli optymalizator natrafi na odpowiedni predykat w warunku selekcji).

Zakłada się, że programista budujący dziedzinowy system informatyczny utworzy również w języku PL/SQL:

- *user-domain_type* – własny typ użytkownika – typ domenowy, złożony, specyficzny dla dziedziny; w tym wypadku (przykład z poprzedniego podrozdziału) to struktura z polami odpowiadającymi atrybutom: a_1,.., a_4,
- *user-domain_table* – tablicę bazy danych z danymi dziedzinowymi, która m.in. ma atrybut typu domenowego,
- *user-domain_function* – funkcję domenową, charakterystyczną dla dziedziny, operującą na typie domenowym.

Problem sprowadza się do stworzenia takiego rozwiązania, które pozwoli optymalizatorowi na wyznaczanie selektywności dla predykatów związanych z funkcją *user-domain_function* (operującą na argumentach typu złożonego).

Rozszerzenie optymalizatora powinno zapewniać następujące działanie SZBD Oracle:

- (w fazie uruchomienia lub konserwacji systemu informatycznego)
  w sytuacji wywołania standardowej komendy ANALYZE TABLE …, tworzącej statystyki dla kolumn *user-domain_table*, dla atrybutu typu *user-domain_type*, SZBD automatycznie uruchomi funkcję *create_user-defined statistics*;
- (w fazie eksploatacji systemu informatycznego)
  w sytuacji, w której optymalizator podczas ewaluacji zapytania SQL natrafi na warunek selekcji, oparty na funkcji *user-domain_function*, selektywność dla tego warunku zostanie obliczona na podstawie wyniku *user-defined_selectivity_function*.

Wcześniej, programista systemowy, budujący rozszerzenie SZBD powinien wykonać następujące czynności:

- implementacja w języku Java klasy:
  - ze statyczną metodą (nazywanej dalej *PCA-based-stat_create*), która odpowiedzialna jest za utworzenie statystyki (histogramy w przestrzeni zredukowanej) oraz zapisanie jej w tablicy systemowej,
  - ze statyczną metodą (nazywanej dalej *PCA-based_selectivity*), wyznaczającą selektywność na podstawie wcześniej utworzonej statystyki.
- Stworzenie funkcji pośredniczących w języku PL/SQL *create_user-defined_statistics* i *user-defined_selectivity_function*, odwołujących się do javowych metod statycznych: *PCA-based_stat_create* i *PCA-based_selectivity* (użycie komendy CREATE FUNCTION … AS JAVA LANGUAGE NAME …).
- Definicja i implementacja typu obiektu w języku PL/SQL (nazywanego dalej *stat_type*), który ma implementować wybrane, statyczne funkcje interfejsu ODCIStats (funkcja *stat_type.ODCIStatsCollect* ma wywoływać *create_user-defined_statistics*; funkcja *stat_type.ODCIStatsSelectivity* ma wywoływać *user-defined_selectivity_function*).

- Skojarzenie typu domenowego ze *stat_type* (komenda ASSOCIATE STATISTICS WITH TYPES *user-domain_type* USING *stat_type*) po to, aby umożliwić tworzenie własnej, specyficznej statystyki w ramach wywołań komend ANALYZE TABLE...
- Skojarzenie funkcji dziedzinowej ze *stat_type* (komenda ASSOCIATE STATISTICS WITH FUNCTIONS *user-domain_function* USING *stat_type*) po to, aby umożliwić optymalizatorowi zapytań użycie właściwej funkcji wyznaczania selektywności, w przypadku wystąpienia predykatu z *user-domain_function*; takie skojarzenie zapewni, że jeśli wystąpi warunek selekcji z funkcją *user-domain_function*, to wywołane zostanie *stat_type.ODCIStatsSelectivity*.

Zakres artykułu nie pozwala na szczegółowe omówienie zagadnienia realizacji technicznej rozszerzenia, ale przykłady zastosowania modułu ODCIStats do implementacji metod wyznaczania selektywności, oparte na koncepcji sieci Bayesa czy dyskretnej transformacie cosinusowej zostały przedstawione m.in. w [8, 9]. W tych wspomnianych rozwiązaniach również chodzi o pokazanie:

- tworzenia przybliżonej, efektywnej pod względem zajętości reprezentacji wielowymiarowego rozkładu wartości atrybutów,
- sposobu implementacji specyficznej metody wyznaczania selektywności zapytań ze złożonym warunkiem zapytania.

4. Podsumowanie

W artykule przedstawiono metodę wyznaczania selektywności zapytań opierając się na koncepcji wynikającej z metody analizy głównych składowych (PCA). Dla zapytań z warunkiem selekcji opartym na kilku atrybutach, obliczanie selektywności wymaga użycia reprezentacji wielowymiarowego rozkładu wartości tych atrybutów. Dla dużej liczby atrybutów w warunku selekcji, wielowymiarowy histogram nie sprawdza się w roli estymatora rozkładu, z powodu zbyt dużej zajętości pamięci. Jednym z podejść do rozwiązania tego problemu jest redukcja wymiarowości reprezentacji rozkładu z wykorzystaniem PCA.

Artykuł na konkretnym przykładzie pokazuje sposób obliczania selektywności z użyciem statystyk, otrzymanych przez zastosowanie metody PCA.

Użycie metody PCA, jako znanego „narzędzia redukcji wymiarowości danych” w zastosowaniach związanych z SZBD, było postulowane przez innych autorów. Pierwsze sugestie, pośrednio związane z tym zagadnieniem, pojawiły się w [2], przy rozważaniach dotyczących przybliżonej, ale pamięciooszczędnej reprezentacji rozkładu dwuwymiarowego z użyciem metody rozkładu macierzy, według wartości osobliwych (ang. *singular value decomposition* – SVD).

Nowym elementem omawianej pracy jest wykorzystanie reguły AVI w przestrzeni zredukowanej. Pozwala to na wykorzystanie histogramów jednowymiarowych w przestrzeni zredukowanej do przybliżonej reprezentacji rozkładów wielowymiarowych z przestrzeni podstawowej. Takie podejście przede wszystkim zmniejszy zajętość pamięci struktury danych, przechowującej odpowiednią statystykę, ale przy okazji, upraszczając opis rozkładu do kilku histogramów jednowymiarowych, pozwoli na przyspieszenie samego procesu wyznaczania selektywności według zaproponowanej metody.

W artykule rozważano histogramy o stałej długości podprzedziału (ang. *equi-width*), ale nic nie stoi na przeszkodzie, żeby omawiana metoda mogła wykorzystać histogramy kwantylowe (powszechnie używane w komercyjnych SZBD, np. SZBD Oracle), czyli takie o stałej wysokości, ale nierównomiernie rozłożonych granicach podprzedziałów (ang. *equi-height*, *equi-depth*).

Artykuł omawia kroki prowadzące do implementacji omawianego rozwiązania (z użyciem modułu Oracle ODCIStats [10]), co umożliwi praktyczne wykorzystanie w SZBD Oracle metody estymacji selektywności opartej na PCA.

Dalsze prace będą dotyczyć ilościowej oceny omawianego rozwiązania.

BIBLIOGRAFIA

1. Krzanowski W. J.: Principles of Multivariate Analysis: A User's Perspective. Oxford University Press, 2000.
2. 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.
3. Gunopulos D., Kollios G., Tsotras V. J.: Approximating Multi-Dimensional Aggregate Range Queries Over Real Attributes. ACM SIGMOD 2000, Dallas 2000.
4. 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.
5. Chakrabarti K., Garofalakis M., Rastogi R., Shim K.: Approximate Query Processing Using Wavelets. VLDB Journal. vol. 10, no. 2-3, Springer-Verlag, New York 2001.
6. Bruno N., Chaudhuri S., Gravano L.: STHoles: a multidimensional workload-aware histogram. Proc. of ACM SIGMOD Int. Conf. on Management of Data 30(2). ACM, New York 2001.
7. Getoor L., Taskar B., Koller D.: Selectivity estimation using probabilistic modes. Proc. of ACM SIGMOD Int. Conf. on Management of Data. ACM, New York 2001.

8. Augustyn. D. R.: Applying advanced methods of query selectivity estimation in Oracle DBMS. Advances in Soft Computing. Man-Machine Interactions. Springer-Verlag, Berlin Heidelberg 2009.
9. Augustyn. D. R., Warchał Ł.: Zastosowanie sieci Bayesa w szacowaniu selektywności zapytań w optymalizatorze zapytań serwera bazy danych Oracle. Studia Informatica, Vol. 32, No. 1A (94), Gliwice 2011.
10. Oracle 10g. Using extensible optimizer (2010). http://download.oracle.com/docs/cd/B14117 01/appdev.101/b10800/dciextopt.htm.
11. Oracle® Database SQL Reference. Analyze (2011). http://download.oracle.com/docs/cd/B19306_01/server.102/b14200/statements_4005.htm.

Wpłynęło do Redakcji 31 stycznia 2011 r.

Abstract

Query selectivity allows to estimate the size of query result. For single-table queries the selectivity value can be defined as a division the number of rows satisfying a query selection condition by the total number of table rows.

The selectivity value is required by DBMS query optimizer for choosing the best query execution method for given query.

A calculation of selectivity value is based on some estimators of a probability density function of distribution of table attribute values. Commonly a histogram is used as non-parametric estimator of density function of attribute values distribution.

But using a multi-dimensional histogram for selectivity estimation for queries with a multi-attribute complex selection condition is too much space-consuming for high dimensions.

The method of Principal Component Analysis (PCA) may be used for dimension reduction of a representation of attribute value distribution. The paper shows the PCA-based selectivity method which operates on dimension-reduced distribution representations (PCA-based statistics). Additionally the AVI rule in the dimesion-reduced space in the dimesion-reduced space (where variables are uncorrelated, not necessary independent) was applied in the proposed method. The AVI rule (attribute value independence assumption) states that a resultant selectivity for a complex condition may be obtained by multiplying selectivities of simple conditions. This makes PCA-based statistics more simple.

The paper also describes steps to implement the proposed solution as an extension of query optimizer of DBMS Oracle using Oracle Data Cartridge Interface Statistics module.

Adres

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