STUDIA INFORMATICA 2012
Volume 33 Number 2A (105)

Dariusz R. AUGUSTYN, Daniel KOSTRZEWA
Politechnika Śląska, Instytut Informatyki

SZACOWANIE SELEKTYWNOŚCI ZAPYTAŃ OPARTE NA TRANSFORMACIE HOUGHA I METODZIE PCA

Streszczenie. Oszacowanie selektywności zapytania jest istotnym elementem procesu uzyskiwania optymalnego planu wykonania tego zapytania. Wyznaczenie selektywności wymaga użycia nieparametrycznego estymatora rozkładu wartości atrybutu, na ogół histogramu. Wykorzystanie wielowymiarowego histogramu jako reprezentacji łącznego rozkładu wielowymiarowego jest nieekonomiczne z powodu zajętości pamięciowej takiej reprezentacji. W artykule zaproponowano nową metodę, nazwaną HPCA, oszczędną pod względem zajętości, gdzie rozkład dwuwymiarowy w przybliżeniu może być reprezentowany w postaci zbioru histogramów jednowymiarowych. Metoda HPCA opiera się na transformacji Hougha i metodzie analizy składowych głównych. Dzięki HPCA można uzyskiwać dokładniejsze oszacowania selektywności zapytań niż te, otrzymane przy wykorzystaniu standardowych 2-wymiarowych histogramów.

Słowa kluczowe: optymalizacja zapytań, selektywność zapytań, histogramy, transformacja Hougha, redukcja wymiarowości, PCA

QUERY SELECTIVITY ESTIMATION BASED ON HOUH TRANSFORM AND PCA METHOD

Summary. Query selectivity estimation is an important element of obtaining optimal query execution plan. Selectivity estimation requires a nonparametric estimator of attribute values distribution – commonly a histogram. Using a multidimensional histogram as a representation of a joint multidimensional distribution of attributes values is not space-efficient. The paper introduces a new space-efficient method called HPCA, where a 2-dimesional distribution may be represented by a set of 1-dimensional histograms. HPCA is based on Hough transform and principal component analysis method. Using HPCA commonly gives more accurate selectivity estimation than standard methods based on a 2-dimensional histogram.

Keywords: query optimization, selectivity estimation, histograms, Hough transform, dimensionality reduction, PCA

1. Wprowadzanie – metody wyznaczania selektywności dla zapytań ze złożonym warunkiem wyszukiwania

Realizacja zapytania SQL przebiega w dwóch etapach, tzn. w ramach fazy przygotowania (ang. *prepare*) oraz fazy wykonania (ang. *execute*). W fazie przygotowania następuje m.in. wypracowanie tzw. planu wykonania zapytania. Zadanie to wykonuje moduł SZBD – optymalizator kosztowy. Uzyskanie planu wykonania jest tożsame z najlepszym, wg optymalizatora, sposobem realizacji zapytania. Ze względu na potencjalną mnogość liczby metod realizacji, każdy z potencjalnych sposobów jest wartościowany (wyznaczony jest szacunkowy koszt wykonania, mierzony np. estymowaną liczbą bloków koniecznych do pobrania z bazy danych). Wybierana jest metoda o najniższym koszcie.

Taka wczesna estymacja (oszacowanie kosztu realizacji zapytania przed jego „właściwym" wykonaniem) wymaga oszacowania parametru zwanego selektywnością. Dla zapytań jednotablicowych selektywność to ułamek, będący stosunkiem liczby wierszy spełniających kryterium zapytania do liczby wszystkich wierszy tablicy. Selektywność przyjmuje wartości z przedziału [0; 1] i można ją traktować jak prawdopodobieństwo wylosowania bez zwracania wiersza spełniającego kryterium wyszukiwania ze zbioru wszystkich wierszy tablicy. Selektywność dla zapytań jednotablicowych Q z zakresowym warunkiem selekcji, dotyczącym atrybutu X (z tablicy T) o ciągłej dziedzinie, można wyrazić następująco:

$$sel(Q(a \le X \le b)) = \int_a^b f\,(x)dx\,, \tag{1}$$

gdzie $f(x)$ to funkcja gęstości prawdopodobieństwa rozkładu wartości atrybutu X.

Stąd (wzór (1)) oszacowanie selektywności może odbywać się na podstawie estymatora funkcji gęstości prawdopodobieństwa. Z reguły w roli nieparametrycznego estymatora funkcji gęstości wykorzystany jest histogram (jednowymiarowy). W większości SZBD w tej roli zastosowany jest histogram *equi-with* (jednakowa szerokość podprzedziałów w dziedzinie wartości atrybutu) lub *equi-high* (możliwie jednakowa liczba wystąpień wartości atrybutu w podprzedziałach).

W przypadku bardziej złożonych zapytań jednotablicowych (tzw. „zapytanie N-wymiarowe"), tzn. takich, gdzie zakresowy warunek selekcji dotyczy kilku atrybutów, selektywność będzie wyrażać się wzorem:

$$sel(Q(a_1 \le X_1 \le b_1 \wedge \ldots \wedge a_N \le X_N \le b_N)) = \int_{a_1}^{b_1} \ldots \int_{a_N}^{b_N} f\,(x_1,\ldots,x_N)dx_1 \ldots dx_N\,, \tag{2}$$

gdzie $f(x_1,\ldots, x_N)$ jest N-argumentową funkcją gęstości prawdopodobieństwa łącznego rozkładu wartości atrybutów $X_1,\ldots, X_N$.

Zastosowanie wielowymiarowego histogramu w roli estymatora wielowymiarowej funkcji gęstości jest nieefektywne ze względu na dużą zajętość pamięci przez taką reprezentację rozkładu (szczególnie dla dużych wymiarowości, jeśli zakłada się również dużą rozdzielczość w każdym z wymiarów). Dlatego w większości komercyjnych SZBD nie ma obsługi łącznego rozkładu wartości atrybutów, a jedynie przechowuje się histogramy jednowymiarowe, opisujące rozkłady brzegowe. Selektywność złożonych zapytań liczona jest na podstawie uproszczonego i na ogół niespełnionego założenia o niezależności atrybutów (reguła AVI [1] – *attribute value idepedence*).

Innym podejściem do opisanego problemu jest poszukiwanie propozycji stratnie skompresowanej reprezentacji rozkładu wielowymiarowego, w której można uzyskać kompromis w zakresie: zajętości reprezentacji – dokładności oszacowania selektywności. W takich podejściach nie ma wykorzystania reguły AVI, co skutkuje lepszymi wynikami estymacji selektywności.

Warto podkreślić, że chociaż komercyjne serwery na ogół nie implementują zaawansowanych metod wyznaczania selektywności, to jednak czasami umożliwiają integrację własnych, autorskich rozwiązań w tym zakresie. Przykładem takiego SZBD jest Oracle ze swoim modułem rozszerzającym ODCI Stat [8]. Przykładami rozszerzenia funkcjonalności optymalizatora zapytań SZBD Oracle mogą być rozwiązania wymienione w [6, 7, 11]. Przykłady te mają pokazać, że własne metody wyznaczania selektywności mogą być wykorzystane w praktycznych zastosowaniach, bazujących na komercyjnych SZBD.

Wśród zaproponowanych i zbadanych teoretycznie metod wyznaczania selektywności, działających na podstawie przybliżonej, pamięciowo-oszczędnej reprezentacji łącznego rozkładu wartości atrybutów, można wymienić metody wykorzystujące: wielowymiarowy estymator jądrowy [2], widmo dyskretnej transformaty cosinusowej [3], widmo transformaty falkowej [4], algorytm GenHist [2] czy sieci Bayesa [5, 7] i wiele innych.

Kolejnym podejściem może być koncepcja wykorzystania metody PCA (metoda analizy składowych głównych) do redukcji wymiarowości reprezentacji rozkładu. To podejście pozwala na zastosowanie wielowymiarowego histogramu, ale w przestrzeni zredukowanej, albo nawet jedynie kilku histogramów jednowymiarowych. Szczegóły takiego podejścia zostały opisane np. w [10]. Metoda jest skuteczna w przypadkach liniowej lub prawie liniowej zależności atrybutów.

W niniejszym artykule rozwinięto metodę zaproponowaną w [10]. W nowej metodzie na wstępnym etapie zachodzi „dekompozycja” zbioru (opisującego rozkład) o potencjalnie nieliniowej zależności na podzbiory z prawie liniową zależnością (użycie transformacji Hougha). Po tym etapie następuje zastosowanie metody PCA (redukcja wymiarowości reprezentacji) w ramach każdego z podzbiorów o prawie liniowych zależnościach atrybutów. W artykule

zaoponowano metodę, pozwalającą na wyznaczanie selektywności opierającej się na zbiorze kilku histogramów, opisujących rozkłady o zredukowanej wymiarowości. Całość metody określono nazwą HPCA (od nazw zastosowanych metod: Hougha i PCA).

2. Wstęp teoretyczny – opis wybranych metod z zakresu przetwarzania obrazów i rozpoznawania obiektów

2.1. Transformacja Hougha

Transformacja Hougha została opracowana jako metoda detekcji wzorców w obrazach binarnych [12]. Metodę tę bardzo szeroko wykorzystuje się w rozwiązywaniu wielu problemów, takich jak: wykrywanie krzywych [13], wykrywanie wcześniej zdefiniowanych kształtów [14], zaawansowana analiza obrazów dokumentów [15], nakładanie się na siebie obrazów medycznych [16]. Jednak najczęściej transformacja Hougha stosowana jest do znajdowania najdłuższych prostoliniowych odcinków, złożonych z punktów obrazu. Każda z prostych może być jednoznacznie przedstawiona w tzw. przestrzeni Hougha, która na ogół opisana jest współczynnikami ρ i θ, określającymi równanie prostej o postaci:

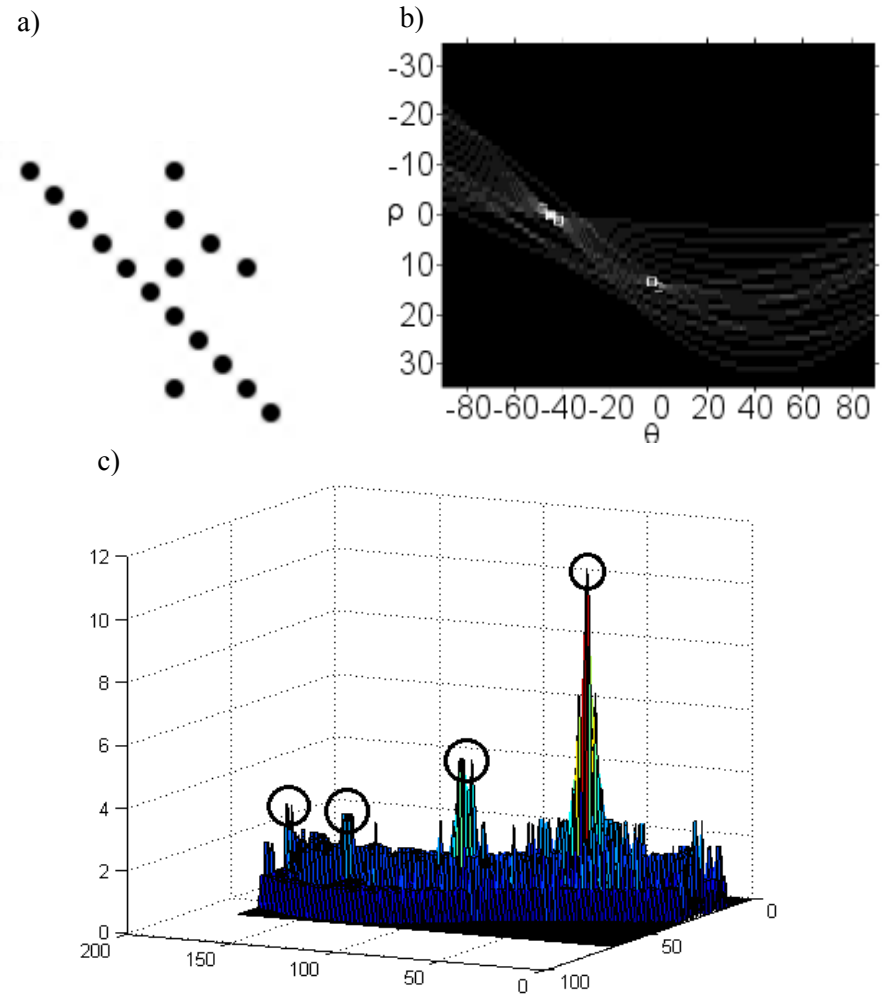
$$\rho = x \cdot cos(\theta) + y \cdot sin(\theta), \tag{3}$$

gdzie: ρ – odległość pomiędzy prostą a początkiem układu współrzędnych, θ – kąt pomiędzy ρ a osią OX.

Zakładając pewną dyskretyzację reprezentacji obrazu, dla każdego niezerowego punktu w obrazie można wyznaczyć skończoną liczbę prostych, opisanych parami ρ oraz θ, przechodzących przez dany punkt [17]. Punktem niezerowym w obrazie binarnym nazywa się piksel o barwie różnej od barwy tła. Każda z wyznaczonych prostych jest odwzorowywana w przestrzeni Hougha jako punkt. Zatem każdy punkt (ρ, θ) przestrzeni zawiera liczbę prostych o takich współczynnikach (rys. 1b). Powstałe w ten sposób maksima powierzchni odpowiadają zbiorom punktów leżących na jednej prostej (rys. 1c), natomiast współrzędne tych maksimów jednoznacznie określają położenie prostych w pierwotnym obrazie binarnym.

Powierzchnia w przestrzeni Hougha często charakteryzuje się dużą liczbą lokalnych maksimów, o zbliżonych wartościach, położonych bardzo blisko siebie. Jest to zwykle efekt niepożądany, który można zniwelować przez zastosowanie powszechnie znanych metod przetwarzania obrazów, np. filtra medianowego.

Filtr medianowy stosowany jest do usuwania punktowych zakłóceń, typu „pieprz i sól”, przy minimalnej utracie jakości obrazu [18]. Przekształcenie filtrem medianowym obrazu, będącego reprezentacją przestrzeni Hougha, powoduje usunięcie pojedynczych lokalnych maksimów przy jednoczesnym uwydatnieniu grup maksimów położonych blisko siebie.



Rys. 1. Wynik transformacji Hougha: a) obraz pierwotny, b) przestrzeń Hougha, c) przestrzeń Hougha z zaznaczonymi wartościami maksymalnymi

Fig. 1. The result of Hough transform: a) original image, b) Hough space, c) Hough space with maximum values

2.2. Analiza głównych składowych

Analiza głównych składowych (PCA, ang. *Principal Component Analysis*) jest jedną z najpopularniejszych metod analizy czynnikowej. Zbiór danych wejściowych stanowi K obserwacji o D zmiennych; inaczej ujmując, jest to K punktów w przestrzeni D-wymiarowej. Celem metody jest obrót układu współrzędnych, w którym maksymalizuje się wariancje kolejnych współrzędnych. Konstruuje się w ten sposób nową przestrzeń, w której największa zmienność danych występuje dla współrzędnych początkowych. Korzystając z tej cechy, można znacznie zredukować rozmiar danych, rezygnując ze współrzędnych o najmniejszej zmienności [19] i opisując każdą z K obserwacji przy pomocy d zmiennych, gdzie $d < D$.

W metodzie PCA są wykonywane następujące kroki:

1. Wyznaczenie średnich dla wierszy.
2. Wypełnienie macierzy kowariancji.

3. Obliczenie wartości własnych macierzy kowariancji.
4. Wyznaczenie wektorów własnych.
5. Rzutowanie danych wejściowych na wybrane wektory własne.

3. Opis zaproponowanej metody estymacji selektywnosci zapytań

Zaproponowana metoda wyznaczania selektywności zapytań zakresowych obejmuje dwa etapy:

- etap wstępny – tzw. faza tworzenia statystyk – tworzenie przybliżonego opisu rozkładu dwuwymiarowego za pomocą zbioru kilku histogramów jednowymiarowych,
- etap wyznaczania wartości selektywności dla konkretnego dwuwymiarowego zapytania zakresowego.

W dalszych rozważaniach przyjęto, że parametr selektywności od razu będzie zawierał estymowaną liczbę wierszy spełniających kryteria (a nie tylko frakcje wierszy spełniających kryteria, tak jak jest to w podstawowej definicji, przedstawionej we wstępie artykułu).

3.1. Algorytm tworzenia opisu rozkładu

Niech $T(X_1, X_2)$ oznacza schemat tablicy bazy danych, gdzie X_k dla $k = 1, 2$ to atrybuty tablicy T. Zaproponowany algorytm, przeznaczony do tworzenia opisu rozkładu za pomocą zbioru histogramów jednowymiarowych, można przedstawić następująco (listing 1):

01 Pobranie N-elementowej próby losowej $\{(x_{1j}, x_{2j})$ dla j=1..$N\}$ – próba 2-wymiarowej zmiennej, wyznaczona na podstawie zawartości wybranej tablicy bazy danych.

02 Uzyskanie zbioru wartości w przestrzeniu akumulacyjnej *theta* × *rho* przez zastosowanie transformaty Hougha do wykrywania prostych. Krok realizowany za pomocą funkcji ***hough*** [20], pochodzącej z Image Processing Toolbox systemu Matlab.

03 Analiza danych w przestrzeni akumulacyjnej – badanie funkcji *a*(*theta, rho*)- sprawdzenie istotności warunku wystąpienia prostych (porównanie wartości maksimów przestrzeni akumulacyjnej z wartością średnią, np. wartość w maksimum większa co najmniej K = 10 razy od wartości średniej). Jeśli nie zachodzi ww. warunek istnienia prostych, to koniec algorytmu (wtedy metoda nie nadaje się dla rozkładu danych pobranych w ramach kroku 01).

04 Filtracja transformaty – zastosowanie dwuwymiarowego filtru dolnoprzepustowego w celu eliminacji nadmiaru lokalnych ekstermów funkcji *a*(*theta, rho*). Krok realizowany za pomocą funkcji ***medfilt2*** [22].

05 Selekcja kliku największych lokalnych maksimów funkcji *a*(*theta, rho*), ze scalaniem maksimów, bliskich w sensie odległości w dziedzinie argumentu funkcji *a*. Krok realizowany przy pomocy funkcji ***houghpeaks*** [21].

06 Redukcja liczby maksimów. Jeśli liczba maksimów jest większa od wartości zadanego parametru *mt* (arbitralnie przyjęta wartość progowa, np. 5) - zastosowanie grupowania maksimów za pomocą podziałowego algorytmu ***KMeans*** [24]. Odpowiednia funkcja grupująca została napisana w języku Matlab na potrzeby niniejszego algorytmu.

07 Dla każdego elementu ze zbioru maksimów – tzn. każdej pary $theta_i$-rho_i opisującej wykrytą prostą p_i:

08 Wyselekcjonowanie grupy g_i punktów sąsiadujących z prostą p_i (tzn. takich, których odległość od prostej p_i jest najmniejsza w stosunku do odległości do pozostałych prostych p_j).

09 Dla punktów należących do grupy g_i wyznaczanie nowej reprezentacji przestrzeni dwuwymiarowej metodą analizy komponentów głównych. Krok realizowany za pomocą funkcji [c_i, s_i, l_i] = *princomp*(…)[23], pochodzącej z Statistics Toolbox systemu Matlab.

10 Sprawdzenie warunku istotności komponentów – weryfikacja warunku, że wartość własna $l_i(1)$ jest istotnie większa od $l_i(2)$, np. $l_i(1)/(l_i(1) + l_i(2)) > 0.95$ – co pozwala na redukcję wymiarowości (redukcja z dwu do jednego wymiaru). Jeśli nie zachodzi ww. warunek, to koniec algorytmu.

11 Budowa i zapamiętanie jednowymiarowego histogramu h_i o stałej szerokości podprzedziałów w oparciu o wartości $s_i(1)$.

12 Zapamiętanie macierzy c_i (2x2) oraz wektora m_i średnich z {(x_{1j}, x_{2j}) dla j € g_i} (2 liczby) na potrzeby określenia przekształcenia $X_1 \times X_2 \rightarrow S_i(1)$.

3.2. Algorytm wyznaczania wartości selektywności zapytania

Rozważmy następujące „dwuwymiarowe" zapytanie zakresowe: $q_x(a_1 \le X_1 \le b_1 \wedge a_2 \le X_2 \le b_2)$, gdzie a_k, b_k dla $k = 1, 2$ oznaczają granice przedziałów w każdym z wymiarów.

Załóżmy, że łączny rozkład wartości X_1 i X_2 będzie się dało zdekomponować z użyciem M prostych, czyli może on być opisany M histogramami jednowymiarowymi.

Zaprezentowany dalej algorytm ma na celu estymację selektywności zapytań klasy q_x. Zaproponowaną metodę wyznaczania wartości selektywności na podstawie M-elementowego zbióru histogramów jednowymiarowych można opisać następująco (listing 2):

01 Dla każdego histogramu h_i (i = 1..M):

02 Transformacja granic zapytań z przestrzeni dwuwymiarowej do jednowymiarowej:
$[sa_1\ sa_2] = ([a_1\ a_2] - m_i)/\ c_i^{-1}$,
$[sb_1\ sb_2] = ([b_1\ b_2] - m_i)/\ c_i^{-1}$.
Tym samym przekształcenie zapytania w przestrzeni $X_1 \times X_2$:
$q_x(a_1 <= X_1 <= b_1$ and $a_2 <= X_2 <= b_2)$
do prawie równoważnego zapytania w jednowymiarowej przestrzeni S:
$q_s(\ sa_1 <= S_1 <= sb_1)$.

03 Wyznaczenie selektywności sel_i na podstawie histogramu h_i oraz granic przedziału zapytania [sa_1 sb_1]. Wykorzystanie uproszczającego założenia, że wewnątrz podprzedziałów histogramu rozkład jest w przybliżeniu równomierny.

04 Wyznaczenie wagi w_i dla selektywności sel_i. Przy założeniu rozkładu normalnego zmiennej S_2 (i znanej wartości wariancji S_2) następuje wyznaczenie prawdopodobieństwa występowania danych w wymiarze S_2, w przedziale określonym przez sa_2 i sb_2. Wartość tego prawdopodobieństwa jest szukaną wagą w_i.

05 Agregacja selektywności składowych - selektywność wynikowa = suma wszystkich w_i * sel_i dla $i = 1..M$.

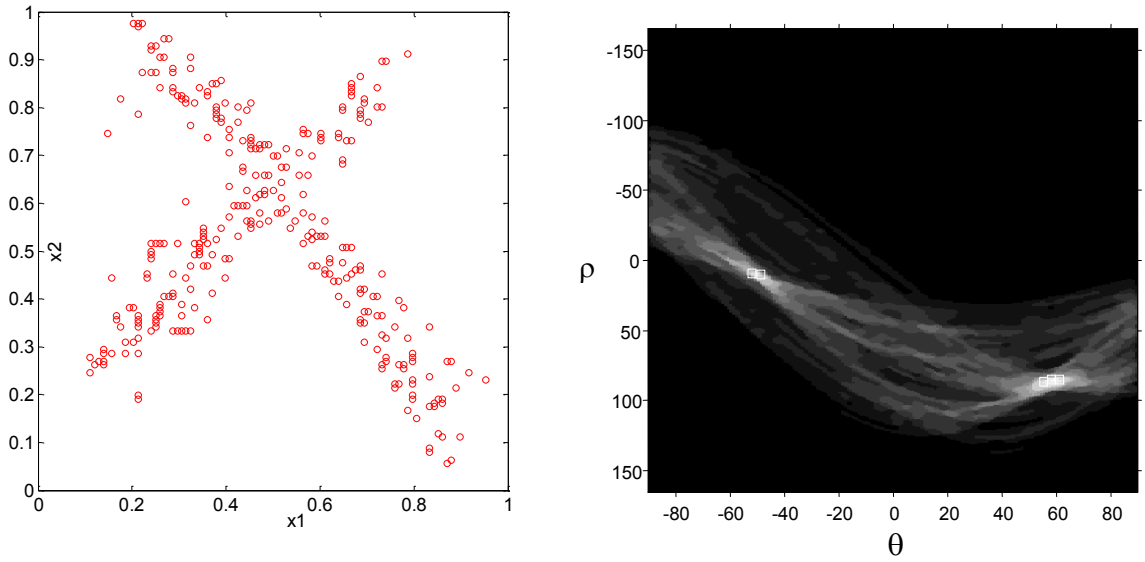
4. Przykład użycia metody

Przedstawiony dalej przykład stanowi ilustrację sposobu budowy reprezentacji rozkładu łącznego (dwuwymiarowego) za pomocą zaproponowanej metody (zbiór histogramów jednowymiarowych) oraz klasycznej reprezentacji rozkładu łącznego (za pomocą histogramu dwuwymiarowego). W obu przypadkach zakłada się, że rozmiar reprezentacji – zajętość pamięci – jest jednakowy i określony wartością parametru *STAT_SIZE*. W rozdziale pokazano, jak wyznaczana jest selektywność dla zapytania zakresowego na podstawie histogramów jednowymiarowych i histogramu dwuwymiarowego. Pokazano zaletę nowego podejścia – mniejszy względny błąd estymacji selektywności w przypadku użycia zaproponowanej metody z histogramami jednowymiarowymi.

4.1. Budowa histogramów jednowymiarowych

Do ewaluacji zaproponowanej metody zastosowano zbiór syntetycznych danych (pobranych z tablicy bazy danych, zgodnie z krokiem 01 listingu 1) o rozkładzie zaprezentowanym na rys. 2a. Wartości X_1 i X_2 należą do przedziału [0; 1], co nie umniejsza ogólności rozwiązania. W przykładzie liczność próby wynosi $N = 308$.

a) b)



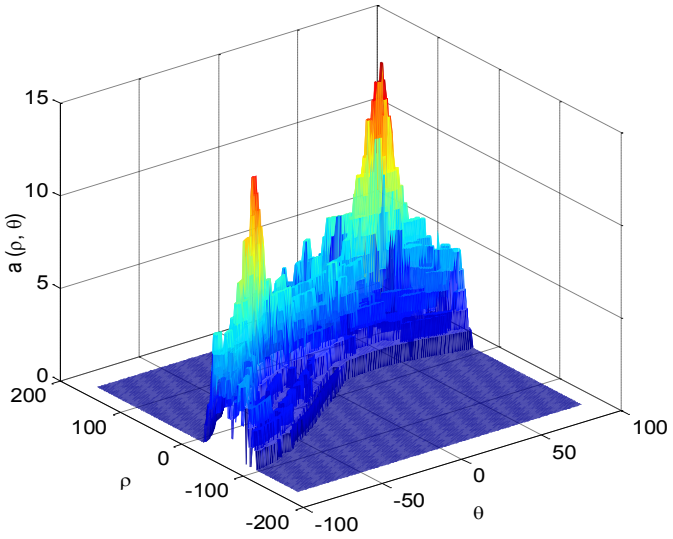
Rys. 2. Ilustracja dwuwymiarowego rozkładu próby losowej w przestrzeni $X_1 \times X_2$ (a), przestrzeń akumulacyjna $\theta \times \rho$ (b) – wynik transformacji Hougha dla danych z rys. 2a (białe kwadraty oznaczają 5 największych maksimów lokalnych funkcji $a\ (\theta \times \rho)$)

Fig. 2. Illustration of bivariate distribution of sample in $X_1 \times X_2$ space (a), $\theta \times \rho$ accumulation space (b)– the result of applying Hough transform for data from fig. 2a (white squares denotes places of 5 greatest local maxima of function $a\ (\theta \times \rho)$)

Dla pobranych danych wykonano przekształcenie Hougha (krok 02 listingu 1), uzyskując wartości w przestrzeni akumulacyjnej pokazanej na rys. 2b. Jasne punkty oznaczają duże wartości transformaty.

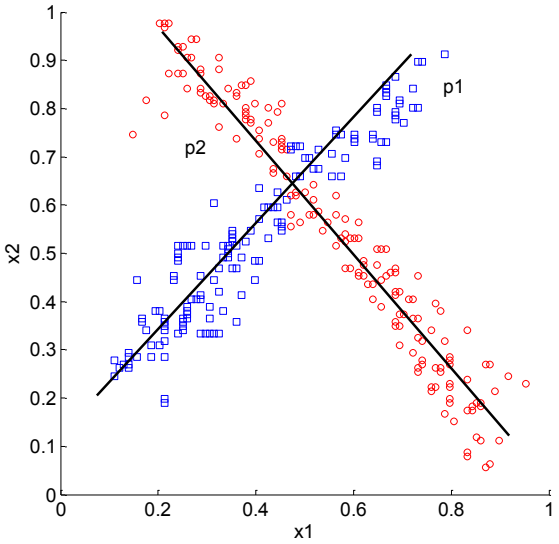
Likwidacja lokalnych maksimów położonych blisko siebie została zrealizowana przez zastosowanie filtru medianowego (krok 04 listingu 1). Wygładzoną transformację Hougha (wartości funkcji a ($\theta \times \rho$)) przedstawiono na rys. 3.

W rozpatrywanym przykładzie przyjęto, że rozkład danych będzie przybliżany maksymalnie mt = 5 histogramami jednowymiarowymi, tzn. dane „będą dopasowywane" maksymalnie do 5 różnych prostych. Jednak już po wygładzeniu i zastosowaniu kryterium „maksimum większe co najmniej K = 10 razy od wartości średniej" wyselekcjonowano tylko dwa maksima (dwa istotnie duże szczyty funkcji a pokazane na rys. 3). Stąd krok 06 listingu 1, polegający na ewentualnym grupowaniu zbioru maksimów, jako nadmiarowy, nie był wykonywany.



Rys. 3. Wartości wygładzonej funkcji akumulacyjnej z dwoma istotnymi szczytami
Fig. 3. Values of the smoothed accumulation function with two significant peaks

Znalezione maksima funkcji a odpowiadają dwóm prostym p_1 oraz p_2 pokazanym na rys. 4. Na podstawie kryterium najmniejszej odległości od prostych p_i (krok 08 listingu 1) podzielono cały zbiór danych z rys. 2a na dwie grupy: g_1 i g_2 (o licznościach odpowiednio 166 i 142). Podział ten zilustrowano na rys. 4.

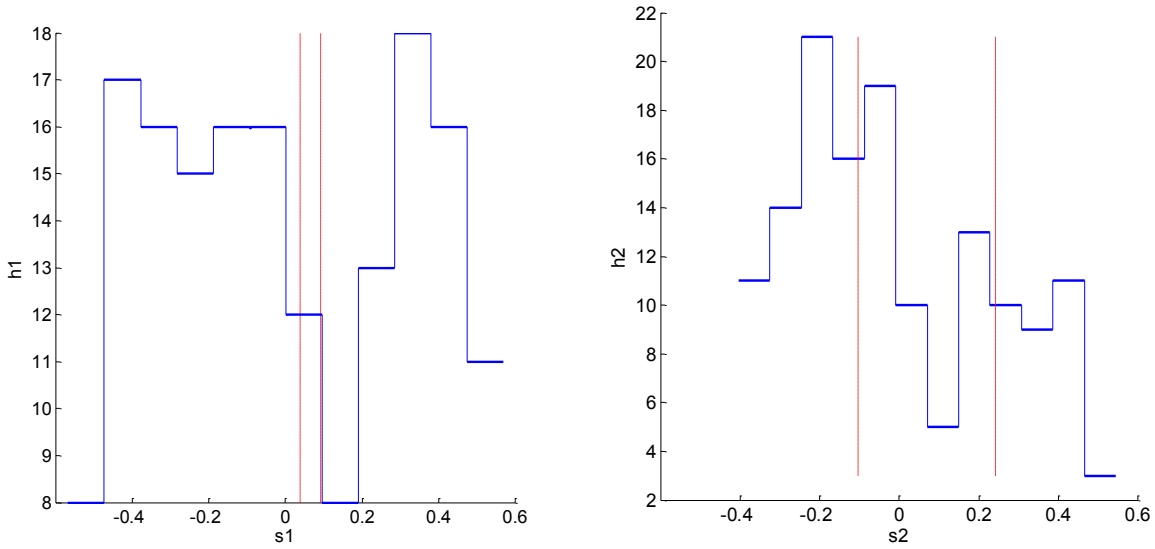


Rys. 4. Istotne wykryte proste oraz przydział punktów do rozłącznych grup, odpowiadających wykrytym prostym
Fig. 4. Significant detected lines and corresponding data points groups

Dla każdej grupy g_i (i = 1.. M = 2) zastosowano metodę PCA (krok 09 listingu 1). Dla grupy g_1 wektor l_1 wynosi [0,1053 0,0017], dla g_2 wektor l_2 wynosi [0,0635 0,0019], czyli $l_1(1)/\text{sum}(l_1) = 0{,}9844 > 0{,}95$ i $l_2(1)/\text{sum}(l_2) = 0{,}9715 > 0{,}95$, co pozwala na przyjęcie założenia, że w obu grupach można zastosować redukcję danych do jednego wymiaru (krok 10 listingu 1). Po realizacji PCA dla obu grup zapamiętane zostają macierze współczynników $c_1 = \begin{bmatrix} -0{,}6293 & 0{,}7771 \\ 0{,}7771 & 0{,}6293 \end{bmatrix}$, $c_2 = \begin{bmatrix} 0{,}6985 & -0{,}7156 \\ 0{,}7156 & 0{,}6985 \end{bmatrix}$ oraz wartości średnich w grupach m_1 = [0,5655 0,5386] i m_2 = [0,3951 0,5344].

Na potrzeby przykładu przyjęto założenie o rozmiarze reprezentacji rozkładu – zajętość pamięci wyrażona ilością liczb reprezentujących rozkład wynosi $STAT\_SIZE = 36 + 2 * 3 = 42$.

Dla każdej grupy g_1 i każdej grupy g_2 stworzono reprezentacje rozkładu w zredukowanej przestrzeni, tzn. w postaci histogramów jednowymiarowych h_1, h_2. W definicji każdego z histogramów h_i, o stałej szerokości podprzedziału, określa się punkt startowy s_i, długość podprzedziału w_i i liczbę podprzedziałów n_i. Założono, że liczba podprzedziałów $n_1 = n_2 = 12$. Zajętość pamięci, wynikająca z definicji każdego z histogramów h_i (s_i, w_i, n_i), wynosi 3. Stąd $STAT\_SIZE$ = 2 * (liczba podprzedziałów n_i + rozmiar definicji histogramu h_i + rozmiar macierzy współczynników c_i + rozmiar wektora średnich m_i) = 2 * (12 + 3 + 4 + 2) = 42.
Korzystając z wartości s_i otrzymanej z transformacji $X_1 \times X_2$ -> $S_i(1)$ dla punktów z grupy g_i oraz przyjmując rozdzielczość histogramu n_i = 12 (krok 11 listingu1), uzyskano jednowymiarowe histogramy h_1 (-0,5672, 0,0946, 12) i h_2 (-0,4047, 0,0791, 12), opisujące rozkłady w grupach g_1 i g_2. Histogramy zostały zaprezentowane na rys 5.

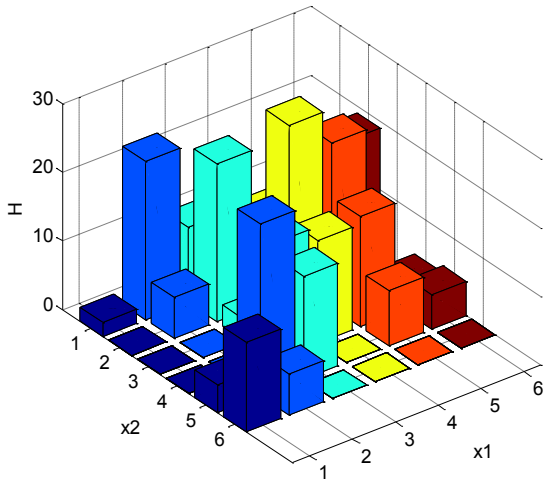


a) b)

Rys. 5. Histogramy jednowymiarowe o rozdzielczości 12: a) histogram h_1 dla grupy g_1, b) histogram h_2 dla grupy g_2

Fig. 5. 1-dimensional histograms with resolution equals 12: a) histogram h_1 for g_1 group, b) histogram h_2 for g_2 group

4.2. Budowa klasycznego dwuwymiarowego histogramu



Rys. 6. Histogram dwuwymiarowy H o rozdzielczości 6 × 6

Fig. 6. 2-dimesional histogram H with resolution equals 6 × 6

Na potrzeby porównania zaproponowanej reprezentacji z dostępnymi rozwiązaniami klasycznymi zbudowano histogram dwuwymiarowy H, opisujący rozkład danych przykładowych, pokazanych na rys 2a. Przyjęto, że zajętość reprezentacji rozkładu H musi być taka

sama jak zajętość reprezentacji rozkładu w postaci dwóch histogramów (opisanych w poprzednim podrozdziale) i wynosi *STAT_SIZE* = 42. Biorąc pod uwagę, że definicja w każdym z dwóch wymiarów powoduje zajętość 3 liczb, na dane opisujące liczebność w każdym (prostokątnym) obszarze zostaje 42 – 2 * 3 = 36. Stąd liczba podprzedziałów w każdym z wymiarów $n_1 = n_2 = \sqrt{36} = 6$.

Warto zwrócić uwagę, że rozdzielczość histogramu H w każdym z wymiarów wynosi 6, natomiast rozdzielczość każdego z dwu histogramów h_i (opisanych w poprzednim podrozdziale) wynosi 12. Właśnie w tej, dwukrotnie lepszej rozdzielczości upatrywana jest przewaga zaproponowanego rozwiązania dla rozpatrywanego przykładu, tzn. polepszenie szacowania selektywności.

Na rys. 5 pokazano histogram H, gdzie każdy z dwu wymiarów został zdefiniowany przez (s_i, w_i, n_i), tzn. (0,1111, 0,1404, 6) i (0,0556, 0,9762, 6). Zamiast wartości granicznych dla podprzedziałów x1 i x2 na rysunku pokazano jedynie numery podprzedziałów.

4.3. Wyznaczenie selektywności dla przykładowego zapytania

Wykorzystanie zaproponowanej reprezentacji rozkładu zostanie zilustrowane przez pokazanie sposobu wyznaczania selektywności dla następującego zapytania:

$$q_x(0{,}3234 \le X_1 \le 0{,}6320 \ \wedge \ 0{,}4593 \le X_2 \le 0{,}6418).$$

Zapytanie to zostaje przetransformowane do dwu zapytań (krok 02 listingu 2): $q_{S1}(0{,}0384 \le S_1 \le 0{,}0908)$ oraz $q_{S2}(-0{,}1039 \le S_2 \le 0{,}2423)$. Granice zakresów obu zapytań zostały pokazane na rys. 5a i 5b w postaci przerywanych linii.

Selektywności składowe sel_i, wyznaczone (krok 03 listingu 2) na podstawie histogramów h_i (rys. 5), wynoszą odpowiednio $sel_1 = sel\ (q_{s1}) = 6{,}6424$ i $sel_2 = sel\ (q_{s2}) = 51{,}9465$.

Ostatecznie selektywność sumaryczna wynosi $sel_{\mathrm{HPCA}}\ (q_x) = sel_1 + sel_2 = 58{,}5889$ (krok 04 listingu 2).

Natomiast selektywność zapytania q_x, wyznaczona na podstanie histogramu dwuwymiarowego H (rys. 6), wynosi $sel_{\mathrm{H2D}}\ (q_x) = 47{,}4452$.

Rzeczywista selektywność zapytania, wyznaczona bezpośrednio na podstawie danych z rys. 2a, wynosi $sel\ (q_x) = 61$.

Błąd względny estymacji selektywności $RelErrSel_{\mathrm{HPCA}}$ z użyciem zaproponowanej metody HPCA wynosi:

$$RelErrSel_{HPCA} = \frac{\left|sel_{\mathrm{HPCA}}(q_x) - sel(q_x)\right|}{sel(q_x)} 100\% \approx 4\%. \tag{4}$$

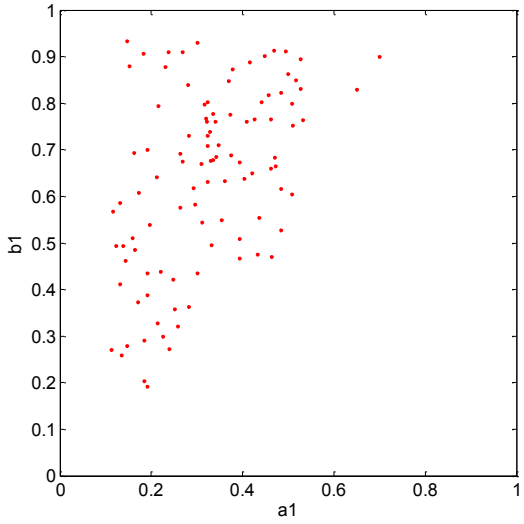
Natomiast błąd względny estymacji selektywności $RelErrSel_{\mathrm{H2D}}$ z bezpośrednim użyciem histogramu dwuwymiarowego H wynosi:

$$RelErrSel_{H2D} = \frac{\left|sel_{\mathrm{H2D}}(q_x) - sel(q_x)\right|}{sel(q_x)} 100\% \approx 21\%. \tag{5}$$

4.4. Analiza błędu selektywności zaproponowanej metody, dla danych przykładowych, w odniesieniu do metody opartej na klasycznym histogramie dwuwymiarowym

Pojedynczy przykład zapytania przedstawionego w poprzednim podrozdziale pokazał pewną przewagę zaproponowanej metody (dla danych przykładowych).

W celu uogólnienia tego wniosku na większą liczbę zapytań rozważono testowy zbiór zakresowych N_q = 100 zapytań z losowymi granicami zakresów a_i i b_i (dla i = 1..2). Wartości a_i pochodziły z generatora *rand*(), generującego liczby pseudolosowe o rozkładzie jednostkowym na odcinku [0; 1]. Natomiast wartości b_i były wyznaczane następująco: $b_i = a_i + (1 - a_i)$ * *rand*(). Rozkład granic zakresów (tj. rozkład par a1, b1) zapytania dla zmiennej X_1 został pokazany na rys. 7.



Rys. 7. Rozkład granic zakresów zapytania (a_1, b_1) dla atrybutu X_1
Fig. 7. Distribution of query bounds (a_1, b_1) for X_1 attribute

Błąd względny metody HPCA – $MeanRelErrSel_{\mathrm{HPCA}}$, uśredniony po zbiorze N_q zapytań, wyniósł ok. 7%.

Błąd względny metody wyznaczania selektywności opartej na histogramie dwuwymiarowym – $MeanRelErrSel_{\mathrm{H2D}}$, uśredniony po zbiorze N_q zapytań, wyniósł ok. 18%.

Wynik dla zaproponowanej metody HPCA jest więc lepszy od wyniku uzyskanego metodą standardową – H2D.

5. Podsumowanie

Artykuł dotyczy tematyki metod szacowania selektywności zapytań w ramach procesu optymalizacji zapytań SQL. Rozważono w nim klasę zakresowych zapytań jednotablicowych. Estymacja selektywności złożonych zapytań wymaga utrzymywania estymatora łącznego rozkładu wartości atrybutów. Przechowywanie reprezentacji rozkładu w postaci wielowymiarowego histogramu może być nieefektywne z powodu zbyt dużej zajętości. Z tego powodu zaproponowano nową metodę, nazwaną HPCA, opartą na metodach i technikach znanych z dziedziny przetwarzania obrazów (wyszukiwanie prostych transformacją Hougha, filtracja dolnoprzepustowa, grupowanie danych metodą K-średnich, redukcja wymiarowości metodą PCA).

Zaproponowana metoda będzie skuteczna, jeśli dane (na podstawie których tworzony jest opis rozkładu łącznego) będą zależne (niekoniecznie liniowo). W metodzie określono kryteria jej stosowalności (wystarczająco duże maksima wartości transformacji Hougha – krok 03 listingu 01, odpowiednio mała istotność komponentów odrzucanych w metodzie PCA – krok 10 listingu 01).

HPCA pozwala na znalezienie takiej reprezentacji łącznego rozkładu (w postaci zbioru histogramów jednowymiarowych), że – przy z góry określonym rozmiarze zajętości reprezentacji – wyznaczane wartości selektywności będą dokładniejsze niż te, obliczone klasycznie, tzn. na podstawie histogramu dwuwymiarowego (pod rozdz. 4.4).

Obecnie metoda HPCA dotyczy rozkładów dwuwymiarowych (ze względu na własności użytej transformacji Houhga), czyli znajduje zastosowanie dla sytuacji, w której zakresowy warunek selekcji jest określony na dwóch atrybutach. Dalsze prace będą koncentrować się m.in. na próbie rozszerzenia metody na rozkłady o większej wymiarowości.

BIBLIOGRAFIA

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. Gunopulos D., Kollios G., Tsotras V. J.: Approximating Multi-Dimensional Aggregate Range Queries Over Real Attributes. ACM SIGMOD 2000, Dallas 2000.
3. 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.

4. 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.
5. 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.
6. 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, s. 585÷593.
7. Augustyn D. R.: Zastosowanie sieci Bayesa w szacowaniu selektywności zapytań w optymalizatorze zapytań serwera bazy danych Oracle. Studia Informatica, Vol. 32, No. 1A (94), Wydawnictwo Politechniki Śląskiej, Gliwice 2011, s. 25÷42.
8. Oracle 10g. Using extensible optimizer (2010), http://download.oracle.com/docs/cd/B14117 01/appdev.101/b10800/dciextopt.htm.
9. Oracle® Database SQL Reference. Analyze (2011), http://download.oracle.com/docs/cd/B19306_01/server.102/b14200/statements_4005.htm.
10. Augustyn D. R.: Metoda analizy głównych składowych w szacowaniu selektywności zapytań. Studia Informatica, Vol. 32, No. 2A(96), Wydawnictwo Politechniki Śląskiej, Gliwice 2011, s. 21÷36.
11. Döller M.: The MPEG-7 Multimedia DataBase System (MPEG-7 MMDB). Dissertation: 117-125 University Klagenfurt, Austria 2004.
12. Hough P. V. C.: Method and means for recognizing complex patterns. United States Patent Office, U.S. Patent 3,069,654, 1962.
13. Xu L., Oja E.: Randomized Hough Transform (RHT): Basic Mechanisms, Algorithms, and Computational Complexities. CVGIP: Image Understanding, Vol. 57, No. 2, 1993, s. 131÷154.
14. Ballard D. H.: Generalizing the Hough Transform to Detect Arbitrary Shapes. IEEE Pattern Recognition, Vol. 13, No. 2, Great Britain 1981, s. 111÷122.
15. Hinds S. C., Fisher J. L., D'Amato D. P.: A Document Skew Detection Method Using Run-Length Encoding and the Hough Transform. Proceedings of 10th International Conference on Pattern Recognition (ICPR), Atlantic City, NJ, USA 1990, s. 464÷468.
16. Chmielewski L.: Nakładanie obrazów metodą transformaty Hougha. Prace XIII Krajowej Konferencji Biocybernetyka i Inżynieria Biomedyczna KBIB 2003, t. 2, Gdańsk 2003, s. 830÷835.
17. Illingworth J., Kittler J.: A Survey of the Hough Transform. Computer Vision, Graphics, and Image Processing, Vol. 44, Elsevier 1988, s. 87÷116.
18. Tukey J.: Exploratory Data Analysis. Addison-Wesley Menlo Park, CA, USA 1977.

19. Jolliffe I. T.: Principal Component Analysis, Second Edition. Springer-Verlag, New York – Berlin – Heidelberg 2002.
20. Hough transform – MATLAB (2012), http://www.mathworks.com/help/toolbox/images/ref/hough.html.
21. Identify peaks in Hough transform – MATLAB (2012), http://www.mathworks.com/help/toolbox/images/ref/houghpeaks.html.
22. 2-D median filtering – MATLAB (2012), http://www.mathworks.com/help/toolbox/images/ref/medfilt2.html.
23. Principal component analysis (PCA) on data – MATLAB (2012), http://www.mathworks.com/help/toolbox/stats/princomp.html.
24. k-means clustering (2012), http://en.wikipedia.org/wiki/K-means_clustering.

Wpłynęło do Redakcji 31 stycznia 2012 r.

Abstract

Query selectivity estimation is an important element of obtaining optimal query execution plan. Selectivity estimation requires a nonparametric estimator of attribute values distribution – commonly a histogram. A multidimensional histogram may be used as a representation of a joint multidimensional distribution of attributes values. This may be not space-efficient for high dimensionality.

The paper introduces a new method of attributes values distribution representation and selectivity calculation. The method is called HPCA. In HPCA a 2-dimensional distribution may be represented by a set of 1-dimensional histograms. Applying Hough transform allows to detect a set of linear dependencies. This allows to divide data to a set of separate subsets with linear or almost linear dependency. Then PCA method is applied for every subsets, so dimensionality of subsets are reduced. A 1-dimensional histogram is created for every subset.

For given size of representation, usage HPCA distribution representation (a set of 1-dimensional histogram) gives better selectivity estimation than using a standard 2-dimensional histograms.

Adresy

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

Daniel KOSTRZEWA: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska, daniel.kostrzewa@polsl.pl.