STUDIA INFORMATICA 2013
Volume 34 Number 2A (111)

Dariusz R. AUGUSTYN
Politechnika Śląska, Instytut Informatyki

M2HSE – METODA ESTYMACJI SELEKTYWNOŚCI PEWNEJ KLASY ZAPYTAŃ ZAKRESOWYCH OPARTA NA WIELOWYMIAROWYM ROZKŁADZIE WARTOŚCI ATRYBUTÓW ORAZ ROZKŁADACH BRZEGOWYCH

Streszczenie. Selektywność jest parametrem wyznaczanym przez bazodanowy optymalizator zapytań w celu wczesnego oszacowania rozmiaru danych spełniających warunek zapytania. Jest to czynność niezbędna do znalezienia optymalnego planu wykonania zapytania. Selektywność jest na ogół oszacowywana na podstawie histogramów, które są nieparametrycznymi estymatorami rozkładów wartości atrybutów. Wyznaczanie selektywności dla zapytań z warunkiem selekcji opartym na kilku atrybutach wymaga wykorzystania wielowymiarowego histogramu estymującego łączny rozkład wartości atrybutów. Dokładność histogramów wielowymiarowych spada wraz ze wzrostem liczby wymiarów, co jest powszechnie znane pod nazwą problemu przekleństwa wymiarowości. Natomiast jednowymiarowe histogramy zbudowane dla pojedynczych atrybutów, które charakteryzują rozkład brzegowy, opisują ten jednowymiarowy rozkład dokładniej, ale oczywiście nie opisują zależności pomiędzy atrybutami. W niniejszym artykule zaproponowano metodę wyznaczania selektywności, opartą na histogramach opisujących zarówno rozkład łączny, jak i rozkłady brzegowe. Zaproponowana metoda (nazwana M2HSE) dotyczy pewnej klasy zapytań, w których zakresowy warunek selekcji oparty jest na wielu atrybutach. Dla takich zapytań przedstawiona metoda może pozwolić na wyznaczenie dokładniejszych przybliżeń wartości selektywności niż klasyczne metody, wykorzystujące histogramy opisujące tylko rozkład łączny albo tylko rozkłady brzegowe (gdzie zastosowane jest założenie o niezależności atrybutów).

Słowa kluczowe: estymacja selektywności zapytań, histogram, wielowymiarowy rozkład wartości atrybutów, rozkład brzegowy

M2HSE – THE SELECTIVITY ESTIMATION METHOD BASED ON MULTIDIMENSIONAL ATTRIBUTE VALUES DISTRIBUTION AND MARGINAL ONES FOR SOME KIND OF RANGE QUERIES

Summary. Selectivity is a parameter obtained by database query optimizer for early estimation of size of data that satisfying a query condition. This is needed for finding the optimal query execution plan. Commonly, selectivity is estimated using histograms that are non-parametric estimators of attribute values distribution. Obtaining a selectivity for a query with a selection condition bases on a few attributes requires a multimensional histogram estimating joint distribution. Accuracy of multidimensional histograms decreases for high dimensions. It is well-known as the curse of dimensionality problem. One-dimensional histograms describing marginal distributions are more accurate, but they do not describe dependency between attributes. In this paper we propose a method of selectivity estimation based on both types of histograms describing either a multidimensional joint distribution or marginal ones. The method (named M2HSE) may be used for some kind of queries with a range selection condition based on many attributes. For such kind of queries, this method may give more accurate selectivity estimations than classical methods based on multidimensional histogram only or marginal histograms only (where the AVI rule is assumed).

Keywords: query selectivity estimation, histograms, multidimensional distribution of attribute values, marginal distribution

1. Wprowadzenie

Faza realizacji zapytania przez serwer bazy danych jest poprzedzona fazą przygotowania. W ramach przygotowania następuje m.in. opracowanie optymalnego planu wykonania zapytania. Elementem opracowywania planu jest wstępne oszacowanie rozmiaru danych spełniających kryteria zapytania (tzn. spełniających tzw. warunek selekcji zapytania). Służy do tego obliczenie parametru zwanego selektywnością. Selektywność dla zapytań jednotablicowych to stosunek liczby wierszy spełniających kryteria zapytania do liczby wszystkich wierszy tablicy. Selektywność można również określić jako prawdopodobieństwo wylosowania wiersza spełniającego warunek selekcji ze zbioru wszystkich wierszy danej tablicy.

Dla zakresowego, jednotablicowego zapytania Q z prostym warunkiem selekcji $a < x < b$, określonym na atrybucie x tablicy T, tj.:

```
Q: select  *  from  T where  a < T.x < b,
```

selektywność wyraża się wzorem:

$$\mathrm{Sel}(Q(a < x < b)) = F(b) - F(a), \tag{1}$$

gdzie: a, b to stałe, F to dystrybuanta rozkładu wartości atrybutu x. Dla atrybutu x o ciągłej dziedzinie wartości selektywność wyraża się następująco:

$$\mathrm{Sel}(Q(a \le x \le b)) = \int_a^b f\,(x)dx\,, \tag{2}$$

gdzie $f(x)$ to funkcja gęstości prawdopodobieństwa rozkładu wartości atrybutu x. Dla zakresowego, jednotablicowego zapytania Q ze złożonym warunkiem selekcji określonym na x_1, x_2,…, x_d, tj. kilku ciągłych atrybutach tablicy, selektywność wyraża się wzorem:

$$\mathrm{Sel}(Q(a_1 \le x_1 \le b_1 \,\&\, ... \,\&\, a_d \le x_d \le b_d)) = \int_{a_1}^{b_1} ... \int_{a_d}^{b_d} f(x_1,...,x_d)dx_1...dx_d\,, \tag{3}$$

gdzie $f(x_1,\ldots, x_d)$ to d-argumentowa funkcja gęstości prawdopodobieństwa wielowymiarowego, łącznego rozkładu $X_1 \times \ldots \times X_d$.

Z formuły (3) wynika, że szacowanie selektywności wymaga użycia estymatora funkcji gęstości. W tej roli najczęściej stosuje się histogram [1]. W komercyjnych systemach zarządzania bazami danych mogą być tworzone histogramy jednowymiarowe, określające rozkłady pojedynczych atrybutów. Z punktu widzenia rozkładu łącznego kilku atrybutów te jednowymiarowe histogramy można traktować jako rozkłady brzegowe rozkładu łącznego.

W wielu SZBD selektywność zapytania zakresowego ze złożonym warunkiem (tzn. takiego jak we wzorze (3)) jest liczona jako iloczyn selektywności warunków prostych (tzn. określonych na pojedynczych atrybutach), tj.:

$$\mathrm{Sel}(Q(a_1 \le x_1 \le b_1 \,\&\, ... \,\&\, a_d \le x_d \le b_d)) = \prod_{j=1}^{d} \mathrm{Sel}(Q(a_j < x_j < b_j))\,. \tag{4}$$

Takie podejście (tj. zastosowanie prawa o prawdopodobieństwie zdarzeń niezależnych) jest równoważne założeniu o niezależności atrybutów. Znane jest ono pod nazwą reguły AVI (ang. *attribute value independence assumption*) [2]. Jednak założenie to często jest niespełnione (wartości atrybutów są często skorelowane) i zastosowanie metod wyznaczania selektywności na podstawie reguły AVI prowadzi do niedokładnych oszacowań.

Oczywistym sposobem na usunięcie tego problemu jest zastosowanie histogramu wielowymiarowego reprezentującego rozkład łączny atrybutów. Jednak tutaj pojawia się kolejny problem – wraz ze wzrastającą liczbą wymiarów i niewielką, z góry określoną, liczbą kubełków histogramu (liczbą definiującą zajętość reprezentacji histogramu w pamięci) aproksymacja rozkładu łącznego staje się niedokładna. Problem oszczędnej pamięciowo reprezentacji łącznego rozkładu wielowymiarowego jest szeroko dyskutowany. Przykładami podejść do tego problemu mogą być rozwiązania oparte na: wielowymiarowym estymatorze jądrowym [3], widmie transformaty cosinusowej [4, 5, 6], widmie transformaty falkowej [7], sieci Bayesa [8] czy transformaty Hougha [9] i innych. Przykładami konkretnych implementacji mogą być rozwiązania takie jak [10, 11].

Uwzględniając powyższe, w niniejszym opracowaniu zaproponowano „mieszaną” metodę wyznaczania selektywności, opartą zarówno na rozkładzie łącznym (uwzględniającym zależ-

ności pomiędzy atrybutami), jak i na rozkładach brzegowych (czyli histogramach jednowymiarowych o dokładnej reprezentacji rozkładu dla pojedynczego atrybutu). Opracowanie stanowi formalną prezentację kroków procedury estymacji selektywności. Omawiana metoda dotyczy tylko pewnej klasy złożonych zapytań zakresowych, tzn. takich, w których większość warunków składowych (warunków prostych, określonych na pojedynczych atrybutach) jest „szeroka" (tzn. pokrywa znaczącą część dziedziny tego atrybutu).

Chociaż metodę omawia się w kontekście zapytań jednotablicowych, to może być użyteczna w przypadku zapytań wielotablicowych, gdzie wyznaczenie selektywności predykatów dotyczących selekcji na łączonych tablicach może determinować sposób i kolejność złączeń (podobny przykład znajduje się w [11]).

W artykule zaproponowano następującą, skróconą nazwę metody – M2HSE, od ang. *Multidimensional and Marginal Histograms-based Selectivity Estimation Method*.

2. Motywacja

Niech x_1 i x_2 są atrybutami relacji R (inaczej kolumnami tablicy R), o ciągłej dziedzinie wartości, odpowiednio D_1, D_2. Przyjmijmy, że:

$f_{x1x2}(x_1, x_2)$ – dwuwymiarowa funkcja gęstości rozkładu $X_1 \times X_2$,

$f_{x1}(x_1) = \int_{D_2} f_{x1x2}(x_1, x_2)dx_2$ – jednowymiarowy rozkład brzegowy wg X_1,

$f_{x2}(x_2) = \int_{D_1} f_{x1x2}(x_1, x_2)dx_1$ – jednowymiarowy rozkład brzegowy wg X_2.

Załóżmy, że zostały utworzone jednowymiarowe histogramy H_{x1} i H_{x2} typu *equi-width* (histogramy o stałej szerokości przedziału) dla x_1 i x_2. Histogramy te są nieparametrycznymi estymatorami funkcji gęstości $f_{x1}(x_1)$ i $f_{x2}(x_2)$. Mechanizmy tworzenia histogramów jednowymiarowych należą do standardowych funkcjonalności SZBD.

Przez H_{x1x2} oznaczmy histogram dwuwymiarowy estymujący funkcję gęstości $f_{x1x2}(x_1, x_2)$. Przyjmijmy, że M to rozdzielczość histogramów jednowymiarowych, tj. liczba kubełków-podprzedziałów histogramów H_{x1} i H_{x2}. Dla uproszczenia zakłada się jednakową liczbę podziałów dziedzin D_1 i D_2, czyli liczbę podprzedziałów $m_{x1}^{(1)} = m_{x2}^{(1)} = M$ dla H_{x1} i H_{x2}. Długości podprzedziałów histogramów można wyznaczyć odpowiednio:

$$h_{x1}^{(1)} = \frac{\overline{\overline{D_1}}}{m_{x1}^{(1)}} = \frac{\overline{\overline{D_1}}}{M}, h_{x2}^{(1)} = \frac{\overline{\overline{D_2}}}{m_{x2}^{(1)}} = \frac{\overline{\overline{D_2}}}{M}, \tag{5}$$

gdzie $\overline{\overline{D_1}}, \overline{\overline{D_2}}$ to długości przedziałów dziedzin atrybutów x_1 i x_2, tzn.

$$\overline{\overline{D_i}} = \max\{x_i\} - \min\{x_i\}. \tag{6}$$

Niech M będzie również rozdzielczością histogramu dwuwymiarowego H_{x1x2}, tj. liczbą kubełków-prostokątów o wymiarach $h_{x1}^{(2)} \times h_{x2}^{(2)}$.

Ponownie załóżmy jednakową liczbę podziałów dziedzin D_1 i D_2 w każdym z dwóch wymiarów histogramu H_{x1x2}, odpowiednio równą:

$$m_{x1}^{(2)} = m_{x2}^{(2)}. \tag{7}$$

Ponieważ

$$m_{x1}^{(2)} \cdot m_{x2}^{(2)} = M, \tag{8}$$

więc uwzględniając (7), można określić:

$$m_{x1}^{(2)} = m_{x2}^{(2)} = \sqrt{M}. \tag{9}$$

Biorąc pod uwagę (9), można wyznaczyć odpowiednie długości krawędzi kubełków-prostokątów histogramu H_{x1x2}:

$$h_{x1}^{(2)} = \frac{\overline{\overline{D_1}}}{m_{x1}^{(2)}} = \frac{\overline{\overline{D_1}}}{\sqrt{M}}, h_{x2}^{(2)} = \frac{\overline{\overline{D_2}}}{m_{x2}^{(1)}} = \frac{\overline{\overline{D_2}}}{\sqrt{M}}. \tag{10}$$

W związku z założeniem, że rozdzielczości wszystkich histogramów (mierzone liczbą kubełków) są jednakowe, można łatwo zauważyć zmniejszenie dokładności aproksymacji dwuwymiarowej funkcji gęstości (w stosunku do dokładności aproksymacji funkcji gęstości rozkładu brzegowego) – liniowe rozmiary kubełków uległy zwiększeniu:

$$\frac{h_{x1}^{(2)}}{h_{x1}^{(1)}} = \frac{h_{x2}^{(2)}}{h_{x2}^{(1)}} = \sqrt{M}. \tag{11}$$

Jest to przejaw znanego zjawiska funkcjonującego pod nazwą „przekleństwa wymiarowości" (ang. *curse of dimensionality*).

Dla przypadku d-wymiarowego histogramu ($d > 2$) jest to w oczywisty sposób jeszcze bardziej widoczne – stosunek długości krawędzi kubełka-hiperkostki (o rozmiarach $h_{x1}^{(d)} \times ... \times h_{xd}^{(d)}$) do długości podprzedziału odpowiedniego histogramu jednowymiarowego wynosi:

$$\frac{h_{x1}^{(d)}}{h_{x1}^{(1)}} = \frac{h_{x2}^{(d)}}{h_{x2}^{(1)}} = ... = \frac{h_{xd}^{(d)}}{h_{xd}^{(1)}} = \frac{M}{\sqrt[d]{M}} = M^{\frac{d-1}{d}} \tag{12}$$

i jest jeszcze większy niż w (11).

Teoretycznie wykorzystanie łącznego rozkładu pozwala na lepsze oszacowanie selektywności zapytań złożonych niż oszacowanie oparte jedynie na znajomości rozkładów brzegowych i użyciu zasady AVI [2] (czyli przyjęcia hipotezy o niezależności zmiennych). Jeżeli jednak rozkłady brzegowe mają dokładniejszą reprezentację (wzór (12)), wówczas nasuwa się

pomysł wykorzystania obu rodzajów opisów jednocześnie (tj. wykorzystania reprezentacji rozkładu łącznego i rozkładów brzegowych). Metoda, która zostanie zaproponowana poniżej, pozwala dla pewnej klasy zapytań skorzystać z obu typów rozkładu.

3. Opis metody wyznaczania selektywności z użyciem rozkładów brzegowych i rozkładu wielowymiarowego

3.1. M2HSE – przypadek dwuwymiarowy

Opisywana procedura dotyczy wyznaczania selektywności złożonych zapytań jednotablicowych z zakresowymi warunkami selekcji. Przykładem takiego zapytania jest „dwuwymiarowe" zapytanie Q zdefiniowane następująco:

$$Q(a_1 < x_1 < b_1 \;\&\; a_2 < x_2 < b_2). \tag{13}$$

Niech Δx_1, Δx_2 oznaczają tzw. szerokość warunku zapytania w wymiarach x_1, x_2, tj.:

$$\Delta x_1 = b_1 - a_1,\ \Delta x_2 = b_2 - a_2. \tag{14}$$

Załóżmy, że szerokość zapytania w wymiarze x_1 jest bardzo duża – tzn. przedział (a_1, b_1) obejmuje niemal całą dziedzinę wartości x_1. Takie zapytania można formalnie określić, wprowadzając pojęcie szerokości względnej (znormalizowanej do 1) w wymiarze x_1:

$$w_1 = \frac{\Delta x_1}{D_1}. \tag{15}$$

Przez zapytanie „szerokie" w wymiarze x_1 należy rozumieć zapytanie z takim warunkiem selekcji, dla którego szerokość względna dla x_1 spełnia warunek:

$$w_1 \geq T, \tag{16}$$

gdzie T jest pewną umowną wartością progową bliską jedności (np. $w_1 = 0{,}9$).

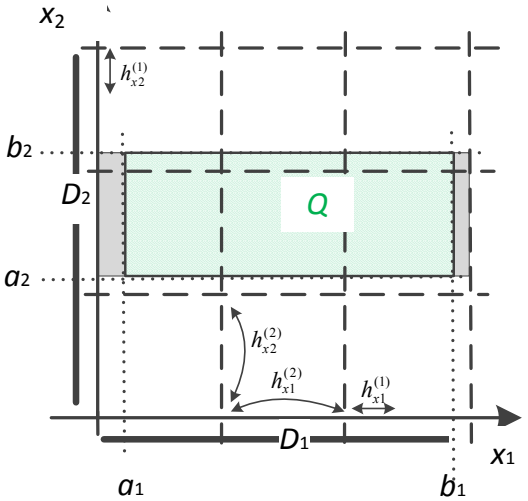
Załóżmy, że zapytanie Q jest „szerokie" w wymiarze x_1, czyli spełnia warunek 16 (natomiast w wymiarze x_2 nie jest „szerokie"). Przykład ilustrujący takie zapytanie przedstawia rys. 1 (dziedzina zapytania jest oznaczona ukośnymi zielonymi liniami).

Dla takiego zapytania można zaproponować pewien specyficzny sposób wyznaczania selektywności – ponieważ zapytanie obejmuje prawie całą dziedzinę x_1, a tylko częściowo obejmuje dziedzinę x_2, więc selektywność będzie można wyznaczyć głównie na podstawie dokładnego rozkładu brzegowego dla x_2.

Pierwsze, najmniej dokładne przybliżenie selektywności opiera się na wykorzystaniu jedynie rozkładu brzegowego dla x_2, tj.:

$$\mathrm{Sel}\,(Q(a_1 < x_1 < b_1 \;\&\; a_2 < x_2 < b_2)) \approx \mathrm{Sel}_{Hx2}\,(Q_2(a_2 < x_2 < b_2)), \tag{17}$$

gdzie: Sel oznacza selektywność dokładną, a Sel_H – selektywność przybliżoną, wyznaczoną na podstawie histogramu H (we wzorze (17) użyto histogramu H_{x2}).



Rys. 1. Przypadek dwuwymiarowy. Dziedziny zmiennych x_1, x_2 oraz zapytania Q. Granice kubełków histogramu dla $M = 9$

Fig. 1. The 2-dimensional case. Domains of x_1, x_2 and query Q. Boundaries of histogram buckets for $M = 9$

Oczywiście, rozwiązanie określone wzorem (17) nie jest prawidłowe, gdyż nie uwzględnia warunku $a_1 < x_1 < b_1$. Wartość uzyskaną w (17) można potraktować jako oszacowanie od góry szukanej selektywności zapytania Q.

Lepsze przybliżenie selektywności można uzyskać, przyjmując proporcjonalny podział wyniku uzyskanego w (17) na podstawie szerokości względnej zapytania w wymiarze x_1, tj.:

$$\mathrm{Sel}\,(Q) \approx s_2 \cdot \mathrm{Sel}_{Hx2}\,(Q_2(a_2 < x_2 < b_2)) = \frac{b_1 - a_1}{D_1}\,\mathrm{Sel}_{Hx2}(Q_2(a_2 < x_2 < b_2))\,, \tag{18}$$

gdzie współczynnik s_2 (współczynnik skalujący selektywność określoną z użyciem rozkładu brzegowego dla x_2) jest zdefiniowany następująco:

$$s_2 = w_1 = \frac{b_1 - a_1}{D_1}\,. \tag{19}$$

Takie rozwiązanie, polegające na skalowaniu wyniku uzyskanego z histogramu brzegowego dla x_2, z zastosowaniem współczynnika proporcjonalności s_2, będącego szerokością względną dla x_1, odpowiednio poprawi wynik oszacowania selektywności.

Ten ostatni krok (zastosowanie formuły (18)) mógłby być już krokiem końcowym proponowanej metody. Przy takim skalowaniu przez współczynnik w_1 domyślnie zakłada się, że rozkład danych jest równomierny w obszarze dziedziny $X_1 \times X_2 = [\min\{X_1\}\ \max\{X_1\}] \times [a_2\ b_2]$ (szary pas na rys. 1), co na ogół nie będzie spełnione (lub będzie spełnione tylko z przybliżeniem).

Jednak możliwe jest zastosowanie innego, adaptacyjnego podejścia, opisanego poniżej. Zamiast skalowania Sel_{Hx2}, wynikającego jedynie z warunku zapytania $a_1 < x_1 < b_1$, skalujący współczynnik proporcjonalności można wyznaczyć, opierając się na aktualnym rozkładzie łącznym (reprezentowany przez H_{x1x2}), tj.:

$$s_{a2} = \frac{\mathrm{Sel}_{Hx1x2}(Q(a_1 < x_1 < b_1 \,\&\, a_2 < x_2 < b_2))}{\mathrm{Sel}_{Hx1x2}(Q_2(a_2 < x_2 < b_2))}, \tag{20}$$

gdzie litera „*a*" w nazwie s_{a2} pochodzi od słowa „adaptacyjny".

Stąd kolejne przybliżenie selektywności zapytania Q, z wykorzystaniem współczynnika skalującego zadanego wzorem (20), jest następujące:

$$\mathrm{Sel}\,(Q) \;\approx\; s_{a2} \cdot \mathrm{Sel}_{Hx2}(Q_2(a_2 < x_2 < b_2)) = \frac{\mathrm{Sel}_{Hx1x2}(Q)}{\mathrm{Sel}_{Hx1x2}(Q_2)} \mathrm{Sel}_{Hx2}(Q_2). \tag{21}$$

Podsumowując, przebieg procedury wyznaczania selektywności dla omawianego przypadku dwuwymiarowego opisuje sekwencja następujących czynności:

1. **Wstępna kwalifikacja zapytania.** Sprawdzenie, czy któryś ze składowych warunków selekcji spełnia kryterium nałożone na szerokość względną, czyli sprawdzenie, czy analizowane zapytanie należy do klasy zapytań objętych niniejszą procedurą wyznaczania selektywności. Załóżmy, że został spełniony warunek $w_1 \geq T = 0{,}9$ (i nie został spełniony taki sam warunek dla w_2).
2. **Wyznaczenie adaptacyjnego współczynnika skalowania**. Obliczenie adaptacyjnego współczynnika skalowania s_{a2} (wzór 20) i potwierdzenie, czy współczynnik adaptacyjny również spełnia kryterium takie jak kryterium nałożone na w_1, tzn. $s_{a2} \geq T$.
3. **Wyznaczanie selektywności na podstawie skalowania wartości selektywności wyznaczonej z rozkładu brzegowego**, czyli zastosowanie formuły (21).

Powyższą procedurę można jeszcze rozpatrywać w pewnym rozszerzonym wariancie. Ostatni, trzeci krok procedury może zostać tak zmodyfikowany, aby uwzględniał zarówno selektywność wyznaczaną wg zaproponowanej metody (z użyciem skalowanego rozkładu brzegowego), jak i wyznaczaną wg metody klasycznej (z użyciem jedynie histogramu dwuwymiarowego). W takim wariancie selektywność wynikowa może być liczona jako średnia ważona wyników obu wspomnianych metod wyznaczania selektywności, tj.:

$$\mathrm{Sel}\,(Q) \approx s_{a2}[s_{a2} \cdot \mathrm{Sel}_{Hx2}(Q_2(a_2 < x_2 < b_2))] + (1 - s_{a2})[\mathrm{Sel}_{Hx1x2}(Q))]. \tag{22}$$

Im większe jest s_{a2}, tym oczywiście istotniejszy jest „wkład" metody opartej na rozkładzie brzegowym. W szczególności dla $s_{a2} \to 1$ selektywność wynikowa jest liczona tylko na podstawie rozkładu brzegowego, co jest zgodne z oczekiwaniami.

3.2. M2HSE – przypadek wielowymiarowy

Poniżej opisano kroki procedury wyznaczania selektywności z użyciem rozkładów brzegowych dla tzw. zapytania *d*-wymiarowego ($d \geq 2$) o postaci:

$$Q(a_1 < x_1 < b_1 \,\&\, a_2 < x_2 < b_2 \,\&\, \ldots \,\&\, a_d < x_d < b_d). \tag{23}$$

1. **Wstępna kwalifikacja zapytania.** Sprawdzenie, czy któryś ze składowych warunków selekcji spełnia kryterium nałożone na szerokość względną – wszystkie składowe warunki selekcji (predykaty zakresowe określone na pojedynczej zmiennej x_j) są względnie szerokie, z wyjątkiem jednego z nich. Załóżmy, że składowy warunek selekcji, który nie spełnia kryterium, jest określony na zmiennej x_k, czyli $w_k < T = 0{,}9$ i $\underset{\substack{j=1\ldots d \\ j\neq k}}{\forall} w_j \geq T$.

Jeżeli zastosowano by podejście nieadaptacyjne, wówczas współczynnik skalowania byłby określony następująco:

$$s_k = \prod_{\substack{j=1\ldots d \\ j\neq k}} w_j = \prod_{\substack{j=1\ldots d \\ j\neq k}} \frac{\Delta x_j}{D_j}. \tag{24}$$

(Wzór (24) jest odpowiednikiem wzoru (19) dla przypadku *d*-wymiarowego). Stąd selektywność można byłoby obliczyć następująco:

$$\mathrm{Sel}\,(Q) \;\approx s_k \cdot \mathrm{Sel}_{Hxk}\,(Q_k(a_k < x_k < b_k)). \tag{25}$$

(Wzór (25) jest odpowiednikiem wzoru (18) dla przypadku *d*-wymiarowego). Dla uzyskania potencjalnie lepszego oszacowania selektywności, podobnie jak dla przypadku dwuwymiarowego, zostanie zaproponowane adaptacyjne skalowanie, opisane poniżej.

2. **Wyznaczenie adaptacyjnego współczynnika skalowania**. Obliczenie adaptacyjnego współczynnika skalowania s_{ak} i potwierdzenie, czy uzyskany współczynnik adaptacyjny spełnia kryterium: $s_{ak} \geq T$.

Adaptacyjny współczynnik skalowania s_{ak} można określić następująco:

$$s_{ak} = \frac{\mathrm{Sel}_{Hx1x2\ldots xd}(Q(a_1 < x_1 < b_1 \,\&\, a_2 < x_2 < b_2 \,\&\, \ldots \,\&\, a_d < x_d < b_d))}{\mathrm{Sel}_{Hx1x2\ldots xd}(Q_k(a_k < x_k < b_k))}. \tag{26}$$

(Wzór (26) jest odpowiednikiem wzoru (20) dla przypadku *d*-wymiarowego).

3. **Wyznaczanie selektywności na podstawie skalowania wartości selektywności wyznaczonej z rozkładu brzegowego.**

Wykorzystując wzór (26), selektywność można obliczyć następująco:

$$\mathrm{Sel}\,(Q) \;\approx\; s_{ak} \cdot \mathrm{Sel}_{Hxk}\,(Q_k(a_k < x_k < b_k)) = \frac{\mathrm{Sel}_{Hx1x2\ldots xd}(Q)}{\mathrm{Sel}_{Hx1x2\ldots xd}(Q_k)}\,\mathrm{Sel}_{Hxk}\,(Q_k). \tag{27}$$

(Wzór (27) jest odpowiednikiem wzoru (21) dla przypadku *d*-wymiarowego).
Jako alternatywę dla wzoru (27) można rozważyć adaptacyjny wariant ważony, tj.:

$$\text{Sel}(Q) \approx s_{ak}[s_{ak} \cdot \text{Sel}_{Hxk}(Q_k(a_k < x_k < b_k))] + (1 - s_{ak})[\text{Sel}_{Hx1x2\ldots xd}(Q))]. \quad (28)$$

(Wzór (28) jest odpowiednikiem wzoru (22) dla przypadku *d*-wymiarowego).

4. Podsumowanie

Jedną z kwestii związanych z zaproponowaną metodą wyznaczania selektywności jest określenie zakresu jej praktycznego zastosowania, w szczególności oszacowanie potencjalnej liczby zapytań, których selektywność mogłaby być wyznaczana za pomocą tej metody. W ogólnym wypadku nie da się jednoznacznie odpowiedzieć na to pytanie, m.in. z powodu nieznanego rozkładu warunków selekcji zadawanych zapytań. Jednak gdy poczyni się pewne założenia dotyczące rozkładu granic zakresów (a_i, b_i) występujących w składowych warunkach selekcji, możliwe staje się określenie, jaki ułamek wszystkich zapytań stanowią zapytania, których selektywność może być wyznaczona omawianą metodą.

Wprowadźmy nowe zmienne y_i (znormalizowane do 1):

$$y_i = \frac{x_i - \min\{x_i\}}{\overline{\overline{D_i}}} \text{ dla } \quad i = 1,\ldots, d. \quad (29)$$

Z formuły (29) wynika, że $y_i \in [0,1]$.

Podobnie zdefiniujmy granice α_i i β_i składowego warunku selekcji określonego dla y_i:

$$\alpha_i = \frac{a_i - \min\{x_i\}}{\overline{\overline{D_i}}}, \beta_i = \frac{b_i - \min\{x_i\}}{\overline{\overline{D_i}}} \text{ dla } i = 1,\ldots, d. \quad (30)$$

Granice α_i i β_i spełniają warunki:

$$0 \le \alpha_i \le 1, \alpha_i \le \beta_i \le 1. \quad (31)$$

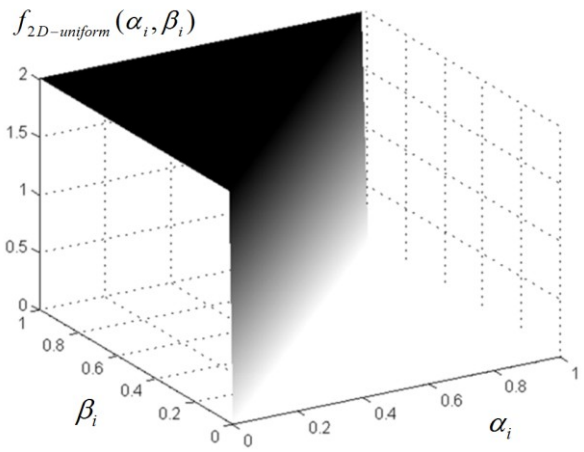
Uwzględniając zależności (29) i (30), można stwierdzić tożsamość wyników zapytań Q_x i Q_y, co oznacza także równość selektywności:

$$\text{Sel}(Q_x(a_1 < x_1 < b_1 \,\&\, \ldots \,\&\, a_d < x_d < b_d)) = \text{Sel}(Q_y(\alpha_1 < y_1 < \beta_1 \,\&\, \ldots \,\&\, \alpha_d < y_d < \beta_d)). \quad (32)$$

Przyjmijmy założenie, że rozkład par (α_i, β_i) jest dwuwymiarowym rozkładem równomiernym, opisanym dwuargumentową funkcją gęstości prawdopodobieństwa:

$$f_{2D-uniform}(\alpha_i, \beta_i) = \begin{cases} 2 & \text{dla} \quad 0 \le \alpha_i \le 1 \wedge 0 \le \beta_i \le 1 \wedge \beta_i \le \alpha_i \\ 0 & \text{w przeciwnym wypadku} \end{cases}, \quad (33)$$

pokazaną na rys. 2.



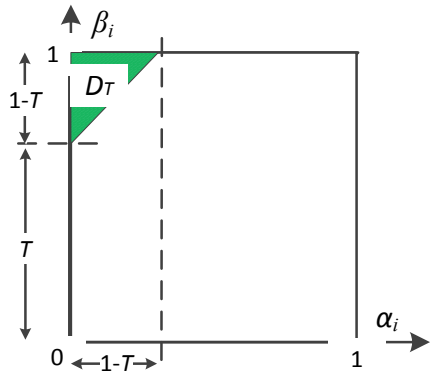
Rys. 2. Dwuargumentowa funkcja gęstości prawdopodobieństwa, określająca równomierny rozkład par (α_i, β_i), tzn. rozkład granic warunku opartego na y_i
Fig. 2. Bivariate probability density function of uniform distribution of pairs (α_i, β_i), i.e. distribution of boundary of query condition based on y_i

Załóżmy, że rozkłady granic warunków składowych dla poszczególnych zmiennych y_i (dla $i = 1,\ldots, d$), czyli rozkłady par (α_i, β_i), są niezależne.

Na początku rozważmy przypadek dwuwymiarowy ($d = 2$). Chcielibyśmy wyznaczyć prawdopodobieństwo P_{M2HSE}, takie że selektywność dowolnego zakresowego zapytania dwuwymiarowego będzie mogła być wyznaczana z wykorzystaniem zaproponowanej metody, czyli że złożony warunek selekcji spełnia kryterium podane w kroku 1. procedury przedstawionej w podrozdziale 3.1. W tym celu wyznaczmy następujące prawdopodobieństwo:

$$P(w_1 \geq T \wedge w_2 < T) = P(w_1 \geq T)P(w_2 < T) = P(w_1 \geq T)(1 - P(w_2 \geq T)). \tag{34}$$

Sposób wyznaczania $P(w_i \geq T)$ zostanie zilustrowany z użyciem rys. 3. Zbiór wszystkich możliwych wartości par (α_i, β_i) to jasnoszary trójkąt. Zbiór par (α_i, β_i) spełniających warunek $w_i \geq T$, czyli takich, że $\beta_i - \alpha_i \geq T$, to mały trójkąt o powierzchni zakreskowanej ukośnymi zielonymi liniami. Ten obszar dziedziny $\alpha_i \times \beta_i$ oznaczono symbolem D_T.



Rys. 3. D_T – podzbiór dziedziny par (α_i, β_i), spełniających warunek $w_i \geq T$
Fig. 3. D_T – the domain subset of (α_i, β_i) pairs that satisfying the condition $w_i \geq T$

Uwzględniając (33), można określić $P(w_i \geq T)$:

$$P(w_i \geq T) = \int_{D_T} f_{2D-uniform}(\alpha_i, \beta_i) d\alpha_i d\beta_i = 2 \cdot \frac{(1-T)(1-T)}{2} = (1-T)^2. \quad (35)$$

Ostatecznie, uwzględniając (34) i (35), otrzymamy:

$$P(w_1 \geq T \wedge w_2 < T) = (1-T)^2(1-(1-T)^2). \quad (36)$$

Prawdopodobieństwo określone formułą (34) nie jest jeszcze szukanym prawdopodobieństwem P_{M2HSE}. Formuła określająca P_{HHSE} jest następująca:

$$P_{M2HSE} = P(w_1 \geq T \wedge w_2 < T) + P(w_1 < T \wedge w_2 \geq T) \quad (37)$$

i dotyczy sytuacji, w której pierwszy warunek jest „szeroki", a drugi jest „wąski" lub pierwszy warunek jest „wąski", a drugi jest „szeroki".

Na podstawie (36) i (39) dla przypadku dwuwymiarowego P_{M2HSE} można wyznaczyć następująco:

$$P_{M2HSE} = 2(1-T)^2(1-(1-T)^2). \quad (38)$$

Na przykład dla $T = 0{,}8$ wartość P_{HHSE} będzie wynosić 0,0768, co oznacza, że selektywność średnio co trzynastego zapytania będzie wyznaczana z użyciem metody M2HSE.

Opierając się na powyższych rozważaniach, łatwo można znaleźć ogólną formułę wyznaczającą P_{M2HSE} dla przypadku d-wymiarowego, tj.:

$$P_{M2HSE} = d(1-T)^{d-1}(1-(1-T)^2). \quad (39)$$

Na podstawie (39) można stwierdzić, że stosowalność metody (mierzona wartością P_{M2HSE}) spada wraz ze wzrostem wymiarowości (ale generalnie spadek nie dotyczy samej dokładności metody).

Powyższe rozważania dotyczące stosowalności zostały przeprowadzone dla przyjętego równomiernego dwuwymiarowego rozkładu par (α_i, β_i) (wobec niesprecyzowanych założeń co do postaci rozkładu granic zakresów warunków zapytania). Podobne rozważania można przeprowadzić dla innych funkcji gęstości rozkładu (α_i, β_i) (podstawienie we wzorze (33) zamiast $f_{2D-uniform}(\alpha_i, \beta_i)$).

Dalsze prace związane z metodą M2HSE najprawdopodobniej będą dotyczyć eksperymentalnej weryfikacji metody pod kątem jej praktycznej stosowalności oraz oceny dokładności estymacji selektywności. „Zmiennymi" eksperymentów będą m.in.: wartość progu T, rodzaj i parametry łącznego rozkładu wartości atrybutów $X_1 \times \ldots \times X_d$., rodzaj i parametry rozkładu (α_i, β_i) granic zakresów w warunku selekcji zapytania. Jednym z celów może być próba wyznaczenia optymalnej wartości parametru T przy zadanym uśrednionym błędzie estymacji selektywności dla określonych klas rozkładów.

BIBLIOGRAFIA

1. Ioannidis Y.: The History of Histograms (abridged). Proc. of VLDB Conference, 2003.
2. Poosala V., Ioannidis Y. E.: Selectivity Estimation Without the Attribute Value Independence Assumption. Proc. of the 23rd VLDB Conference, Athens, Greece 1997, s. 486÷495.
3. Gunopulos D., Kollios G., Tsotras V. J.: Approximating Multi-Dimensional Aggregate Range Queries Over Real Attributes. ACM SIGMOD 2000, Dallas 2000, s. 137÷154.
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, s. 205÷214.
5. Yan F., Hou W.-C., Jiang Z., Luo C., Zhu Q.: Selectivity estimation of range queries based on data density approximation via cosine series. Data & Knowledge Engineering 63(3), SienceDirect, 2007, s. 855÷878.
6. Augustyn D. R.: Asymptotically error-optimal shape of sampling zone for query selectivity estimation method based on discrete cosine trans form. Theoretical and Applied Informatics, Vol. 24 , No. 1, Versita, Warsaw 2012, s. 3÷22.
7. 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, s. 199÷223.
8. 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, s. 461÷472.
9. Augustyn D. R., Kostrzewa D.: Szacowanie selektywności zapytań oparte na transformacie Hougha i metodzie PCA. Studia Informatica, Vol. 33, No. 2A (105), Gliwice 2012, s. 211÷227.
10. 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.
11. 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, s. 25÷42.

Wpłynęło do Redakcji 16 stycznia 2013 r.

Abstract

Selectivity is a parameter obtained by database query optimizer for early estimation of size of data that satisfying a query condition. This is needed for finding the optimal query execution plan.

Most often, a selectivity is estimated using histograms that are non-parametric estimators of attribute values distribution. Obtaining a selectivity for a query with a selection condition bases on a few attributes requires a multimensional histogram estimating joint distribution. Accuracy of multidimensional histograms decreases for high dimensions. It is well-known as the curse of dimensionality problem. For a given number of all buckets, one-dimensional histograms describing marginal distributions are more accurate than multidimensional histograms. Those "marginal" histograms are built on a single attribute only so a dependency between many attributes is not described at all.

In this paper we propose a method of selectivity estimation based on both types of histograms describing either a multidimensional joint distribution or marginal ones. The method (named M2HSE) may be used for some kind of queries with a range selection condition based on many attributes. For such kind of queries, this method may give more accurate selectivity estimations than classical methods based on multidimensional histogram only or marginal histograms only (where the AVI rule is assumed).

The paper describes details of the proposed procedure of selectivity estimation. The problem of scope of applying the method is also considered.

Adres

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