# 다차원 히스토그램에서 범위 질의의 선택도에 대한 오차 추정

정지훈[0], 홍석진, 배진욱, 안성준, 송병호[+], 이석호
서울대학교 전기·컴퓨터공학부
[+]상명대학교 소프트웨어학부
{combat[0], jinny, oblody}@db.snu.ac.kr, junny@bawi.org, bhsong@sangmyung,ac,kr, shlee@cse.snu.ac.kr

# Error Estimation about Selectivity of Approximate Range Queries in Multi-Dimensional Histogram

Jihoon Jung[0], Seokjin Hong, Jinuk Bae, Seongjoon Ahn, Byoungho Song[+], Sukho Lee
School of Electrical Engineering and Computer Science, Seoul National University
[+]Division of Software Science, Sangmyung University

## 요 약

히스토그램은 질의 최적화를 위해 사용되는 통계 정보 중 하나이다. 최근에는 방대한 데이타에 대한 범위 질의의 선택도 추정 방법의 하나로 사용되기도 한다. 히스토그램을 통한 범위 질의의 선택도 추정 결과는 항상 오차를 포함한다. 따라서 결과의 신뢰성을 보장하기 위해 선택도에 대한 오차를 추정하는 방법이 요구된다. 추정된 선택도의 오차 추정에 대한 기존 방법은 1차원 히스토그램만을 고려하여 하나의 애트리뷰트의 값에 따라 빈도의 분포를 반영하므로 애트리뷰트가 많은 다차원 히스토그램에 바로 적용시키는데 문제가 있다.

이 논문에서는 기존의 추정된 선택도에 대한 오차 추정 기법들을 다차원에 적용할 수 있게 확장한 M-Max, M-Sum 기법을 제안하고, 두 기법을 합친 하이브리드 기법을 제안한다. 실험을 통해 M-Sum 기법과 하이브리드 기법이 M-Max 기법보다 정확한 오차 추정 기법임을 보이고, 또한 작은 기억 공간에서도 두 기법이 오차를 보다 정확하게 추정함을 보인다.

## 1 서 론

히스토그램(*histogram*)은 질의 최적기에서 특정 애트리뷰트 값의 분포를 파악하기 위한 통계 자료 중 하나이다. 하지만 최근에는 방대한 데이타에 대한 범위 질의의 선택도(selectivity) 추정을 처리하기 위한 방법 중 하나로 사용되고 있다.

히스토그램에 대한 연구는 Equi-width 히스토그램을 이용한 질의 최적화 기법을 시작으로 최근에는 기존 히스토그램의 구성 방식을 정리하고 히스토그램의 정확성을 보다 높일 수 있는 기법[1]이 소개되었다.

1차원 히스토그램에 대한 기법을 다차원 데이타에 적용하는 방법으로 Equi-width 히스토그램을 다차원에 적용시킨 다차원 히스토그램 기법[2]이 제안되었고, 이를 좀더 발전시켜 기존의 버킷 분할 방식(Equi-depth, Max-diff 등)을 모두 적용할 수 있는 MHIST 기법[3]이 소개되었다.

히스토그램이 범위 질의의 선택도 추정에 사용되면서 이에 대한 오차를 추정하기 위한 기법으로 실제 빈도와 평균 빈도의 최대차를 이용한 기법(이하 Max 기법)[4]과 실제 빈도 누적 합과 평균 빈도 누적 합의 최대차를 이용한 기법(이하 Sum 기법)[5]이 소개되었다.

그러나, 이 기법들은 1차원의 경우만을 고려하여 한 애트리뷰트 값의 빈도에 대한 분포를 사용하기 때문에 다차원 데이타를 사용한 범위 질의의 선택도에 대한 오차 추정에 그대로 적용시킬 수 없다. 특히 Sum 기법의 경우 다차원에 적용 시 빈도 누적의 방향성, 질의 형태에 따른 오차의 수정 문제들이 더불어 발생한다. 이 논문에서는 기존의 기법을 다차원에 적용할 때 발생하는 문제들을 모두 해결한 M-Max, M-Sum 기법을 제안하고, 덧붙여 하이브리드(hybrid) 기법을 제안한다.

이 논문의 구성은 다음과 같다. 2절에서는 1차원 히스토그램을 이용한 기존 기법들에 대한 설명을 하고, 3절에서는 다차원에서 사용할 수 있는 M-Max, M-Sum 기법을 제안한다. 4절에서는 실험을 통한 성능 분석과 결과를 정리하고, 5절에서는 결론을 맺는다.

## 2 관련 연구

이 절에서는 1차원 히스토그램에서의 범위 질의에 대한 선택도 추정 방법과 기존에 제안된 오차 추정 기법(Max, Sum)에 대해 알아본다.

### 2.1 1차원 히스토그램의 구성과 범위 질의에 대한 선택도 추정 방법

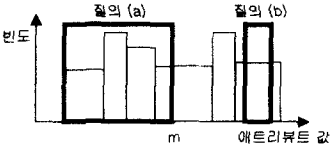


그림 1. 1차원 히스토그램의 범위 질의

히스토그램의 각 버킷( $b_i$)은, 버킷에 포함되는 애트리뷰트 값 ( $v_{i,j} \in b_i$, $v_{i,j}$는 $b_i$에 속하는 $j$ 번째 애트리뷰트 값) 중 최소값 ( $v_i^{min}$), 최대값( $v_i^{max}$)그리고 버킷 안의 평균 빈도( $f_i^{avg}$)로 구성된다.[4] 사용되는 범위 질의의 형태는 다음과 같다.

SELECT ATTR FROM R WHERE ATTR <= m

범위 질의에 대한 선택도는 질의에 속하는 애트리뷰트 값들의 빈도 합을 뜻한다. ( $\sum f_{i,j}$ , $f_{i,j}$는 $v_{i,j}$의 빈도 , $i=1,\cdots,n$, $n$은 버킷 수, $j=1,\cdots,n_i$ , $n_i$는 $i$ 번째 버킷의 애트리뷰트 값의 수)

위 질의를 도시한 그림 1의 질의 (a)를 보면 첫 세 버킷은 질의에 완전히 포함되지만 m을 포함한 버킷은 부분적으로 포함됨을 알 수 있다. 앞의 세 버킷은 ( $v_i^{max} - v_i^{min}$)× $f_i^{avg}$ 연산으로 정확한 선택도가 계산되며, 질의에 버킷의 일부만 포함되는 부분은 ( $m - v_i^{min}$)× $f_i^{avg}$ 연산으로 선택도를 추정하게 된다. 이 경우 추정된 선택도는 오차를 포함하게 되며, 이 오차를 추정하는 기법으로 Max 기법[4]과 Sum 기법[5]이 있다.

### 2.2 선택도에 대한 오차 추정 기법

#### 2.2.1 Max 기법

[4]에서는 각 버킷에 실제 빈도와 평균 빈도의 최대 차, $E_i = MAX\{|f_{i,j} - f_i^{avg}| \mid j=1,\cdots,n_i\}$ ($f_{i,j}$는 $v_{i,j}$의 빈도, $n_i$는 $v_{i,j}$의 수)를 추가하고, 이를 이용하여 선택도 추정에 대한 오차 추정을 하였다. 그림 1의 질의 (a)의 네 번째 버킷과 같이 버킷 일부가 질의에 포함된 경우 $E_i$를 이용한 오차 추정 식[4]은 다음과 같다.

$$\min(m - v_i^{\min} + 1,\ v_i^{\max} - m + 1) \times E_i$$

그림 2의 (a)는 $b_i$ 버킷의 $E_i$를 구하고 범위 질의에 대한 선택도 추정과 오차 추정 과정을 나타낸 것이다. 하지만 이 기법은 각 버킷이 실제 빈도와 평균 빈도의 최대차만을 고려하고 있기 때문에 그림 2의 (a)처럼 한 버킷 안에 평균 빈도와 큰 차이를 보이는 값이 있을 경우 $E_i$값이 커지게 되어 신빙성 있는 오차 추정이 어렵다는 문제점이 있다.

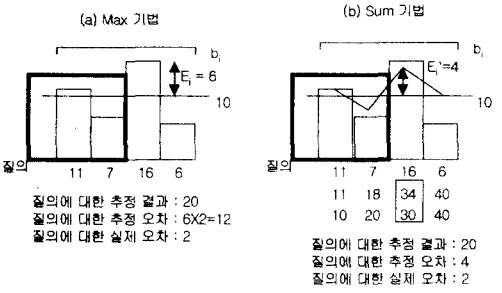


그림 2. 1차원 히스토그램에서의 Max, Sum 기법

#### 2.2.2 Sum 기법

[5]에서는 버킷 내의 분포를 고려하여 실제 빈도와 평균 빈도의 최대차, $E_i$ 대신 실제 빈도의 누적 합과 평균 빈도의 누적 합의 최대차, $E_i' = MAX\left\{\sum_{i=1}^{j}(f_{i,j} - f_i^{avg}) \mid j=1,\cdots,n_i\right\}$ ($n_i$는 $v_{i,j}$의 수) 를 사용하였다. Max 기법과 달리 이 값은 어떤 범위 질의라도 $E_i'$ 이상의 오차가 날 수 없음을 보장하기 때문에 $E_i'$를 바로 오차 추정 값으로 사용한다.

그림 2의 (b)는 $E_i'$를 구하는 과정과 이를 이용하여 보다 정확한 오차 추정이 이루어짐을 보이고 있다.

이 기법은 범위 질의의 선택도가 결국 빈도의 누적 합을 사용한다는 것에 착안한 방법으로 Max 기법에 비해서 버킷의 분포를 더 잘 나타낸다고 할 수 있다. 하지만 이 기법은 그림 1의 질의 (b)와 같이 질의가 하나의 버킷 안에 포함되는 경우에 대해서는 고려하고 있지 않다.

Max 기법은 한 애트리뷰트의 값에 대한 빈도를 사용하고 Sum 기법은 한 애트리뷰트의 값을 따라 빈도를 누적하고 있다. 하지만 이렇게 한 애트리뷰트의 값에 대한 분포를 사용하는 기법들은 애트리뷰트가 많은 다차원의 경우에 바로 적용시킬 수 없다. 다음절에서는 위 기법들을 다차원에 적용시키는 방법들을 제안한다.

## 3 다차원 히스토그램을 이용한 근사값에 대한 오차 추정

### 3.1 다차원 히스토그램의 범위 질의

다차원 히스토그램에 대한 범위 질의는 다음과 같이 각 애트리뷰트들이 각각의 범위 조건을 가진다.

SELECT A1,A2 FROM R WHERE A1<=m AND A2>=n

다차원의 경우 1차원 히스토그램과는 다르게 그림 3의 (a)에 진하게 표시된 부분들처럼 질의에 일부분만 포함되는 버킷들이 많아진다. 그래서 이런 각 부분들의 오차를 모두 합하여 오차 추정 값으로 사용한다.

그리고 애트리뷰트 도메인의 조합, 예를 들어 A1의 도메인이 {a,b,c} 이고 A2의 도메인이 {1,2,3} 일 때, (a,1),(a,2),…(c,3) 이들 각각을 셀(cell)이라고 하겠다.

### 3.2 M-Max 기법

M-Max 기법은 Max 기법의 $E_i$을 변경하여 적용했다. $E_i$를 구하는 방법은 각 버킷 안의 셀의 빈도와 셀의 평균 빈도의 최대차를 사용했다. $Q_{b_i}$는 질의와 $b_i$ 버킷이 겹치는 부분에 있는 셀의 수이고, $b_i^{count}$는 $b_i$ 버킷이 가지고 있는 셀의 수라고 할 때 각 버킷의 오차 추정 식은 다음과 같다.

$$\min(Q_{b_i},\ b_i^{count} - Q_{b_i}) \times E_i$$

이렇게 구해진 각 버킷의 오차 추정 값들을 모두 합치면 범위 질의의 선택도 추정에 대한 오차 추정 값이 된다.

그림 3의 (b)는 2차원에서 이 오차 추정 방법을 사용한 예이다. 버킷 $b_i$의 평균 빈도 10과 차이가 가장 큰 15의 차를 $E_i$로 사용하여 범위 질의의 선택도 추정과 오차 추정과정을 보이고 있다.

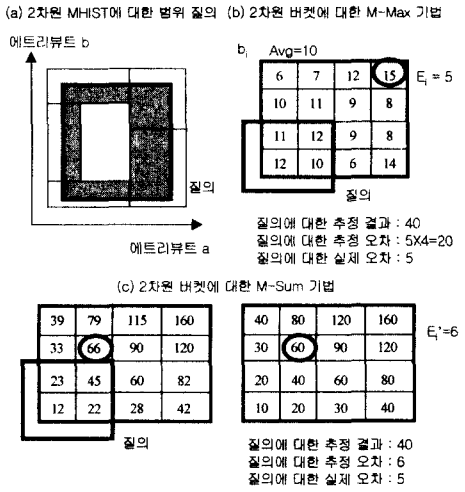


그림 3. 2차원에서의 M-Max, M-Sum 기법

### 3.3 M-Sum 기법

기존 Sum 기법은 애트리뷰트의 값들의 빈도를 그 애트리뷰트 축을 따라 누적해 가는 방법을 사용했다. 하지만 이 방법은 다차원에 적용하기 위해서는 두 가지 큰 문제점을 가지고 있다. 첫째는 한 애트리뷰트을 따라 누적해 나가는 방향성 때문에 애트리뷰트 수가 많은 다차원의 경우 Sum 기법을 그대로 적용 할 수 없다는 것이고, 둘째는 그림 1의 (b)처럼 버킷에 완전히 포함되는 질의 혹은 다차원의 경우 누적 방향과는 다른 방향성을 지닌 질의가 들어오는 경우 이에 대한 처리를 할 수 없다는 것이다.

첫 번째 문제인 누적 방향에 대한 해결책으로 이 논문에서는 그림 3의 (b)에 있는 버킷을 (c)처럼 실제 빈도와 평균 빈도를 각 애트리뷰트 값의 증가 방향으로 모두 누적했고, 둘의 최대차를 $E_i'$로 사용했다. $f_i(a_1, a_2, \cdots, a_d)$가 $b_i$안의 각 빈도들을 나타낸다고 할 때, 이 기법의 오차 추정 식은 다음과 같다.

$$E_i' = MAX\left\{x \mid x = \left| \sum_{k_1=a_1^{min}}^{a_1} \cdots \sum_{k_2=a_2^{min}}^{a_2} \sum_{k_i=a_i^{min}}^{a_i} (f_i(k_1, k_2, \cdots, k_d) - f_i^{min}) \right| , \ a_i^{min} \le a_i \le a_i^{max}, \ a_i^{min} \in b_i, \ a_i^{max} \in b_i \right\}$$

그림 3의 (c)는 2차원에서 이 기법으로 실제 빈도, 평균 빈도의 누적 합을 이용하여 $E_i'$를 구하는 과정을 보이고 있다.

두 번째 문제인 누적 방향과는 다른 방향의 질의 혹은 하나의 버킷에 완전히 포함되는 질의에 대한 오차 추정은 다음과 같이 처리한다. 앞에서 첫 번째 문제의 해결 방법인 모든 애트리뷰트를 따라 각 애트리뷰트의 빈도를 누적해 가는 방법은 그림 4의 (a)에 있는 질의 1 형태의 질의들만 고려 한 것이다. 하지만 질의 2, 3, 4의 경우 질의 1과 거의 같은 확률로 나타날 수 있는 질의 형태이다. 따라서 이런 형태도 고려하여 그림 4의 (b)와 같이 애트리뷰트들의 증가·감소 방향의 모든 조합 방향으로 각 $E_i'$를 구하고 이 중 가장 큰 값만을 M-Sum의 실제 $E_i'$로 사용하였다.

하지만 질의 5, 6의 경우 여전히 누적 방향과는 다른 방향성을 지닌 질의이다. 그러나 이런 경우 기존의 질의 중 하나의 형태로 합, 차 연산을 통해 그 형태를 만들 수 있다. 예를 들어 질의 6은 그림 4의 (c)처럼 질의 I 형태 질의들의 합·차 연산을 통해 구해질 수 있다. 연산으로 인한 오차의 중복으로 질의 6의 오차는 $4\times E_i'$가 된다. 마찬가지로 질의 5의 경우 오차는 $2\times E_i'$가 된다. 위 기법에 대한 오차 추정 식은 다음과 같다.

$$E_i'' = 2^{d-m} \times E_i'$$

(질의에 포함되는 경계 수=m, 데이타 차원=d)

위 식은 그림 4의 경우 d 값은 2가 되고, m 값은 질의 1, 2, 3, 4의 경우 2, 질의 5의 경우 1, 질의 6의 경우 0이 된다. 위 방법은 집합에서 교집합의 크기를 구하는 방법과 동일하다.

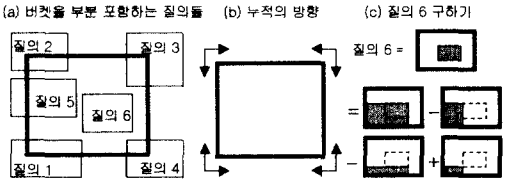


그림 4. M-Sum의 질의들과 누적 방향

### 3.4 하이브리드 오차 추정 기법

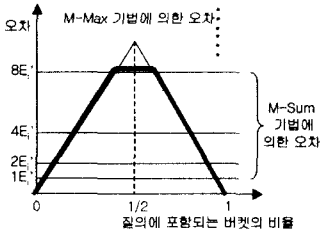


그림 5. M-Max, M-Sum 오차 그래프

차원이 높아짐에 따라 그림 4의 질의 6과 같은 형태는 경우에 따라 오차 값이 커질 수 있다. 하지만 그런 경우는 극히 드물기 때문에 무시할 수 있다. 그러나 질의의 범위가 매우 작고 포함된 버킷의 분포가 극단적인 경우 M-Max가 다소 정확한 오차를 추정하는 경우가 발생할 수 있다. 따라서 두 기법으로부터 얻어진 값 모두를 버킷에 저장하여 결과가 작은 것을 사용하는 하이브리드 기법을 생각할 수 있다. 그림 5는 두 가지 기법이 만들어 내는 오차를 그래프로 나타낸 것이다. 굵은 선은 하이브리드 기법에서 사용하게 되는 오차 값을 나타낸다.

이 기법은 위의 기법들보다 정확한 오차를 추정할 수 있다는 점에서 우수하지만 두 가지 값을 모두 저장해야 하므로, 유지할 수 있는 버킷의 수가 줄어든다는 단점이 있다.

## 4 성능 평가

실험 데이타는 [6]에서 제공하는 유체역학 데이타를 사용하였고, 히스토그램은 2차원 MHIST에 Max-diff를 사용하여 구성하였다.

그림 6의 (a)는 임의의 범위 질의 10000개에 대해 오차들의 평균을 나타내고 있다. 그림에서 실제 오차는 실제 데이타를 이용한 범위 질의 결과와 M-Max, M-Sum 기법에 사용된 히스토그램을 이용한 범위 질의 결과의 차를 나타내고 있다. 그래프에서는 Hybrid 실제 오차는 하이브리드 기법의 경우 다른 방법들과 버킷의 크기가 달라 실제 오차도 다르므로 따로 표시해 주었다. (a)에서는 M-Max 기법에 비해 다른 기법들이 더 정확한 결과를 보이고 있다. (b)는 M-Max 기법을 제외한 결과를 보인 것으로 M-Sum 기법이 하이브리드 기법보다 좀 더 정확한 오차 추정 값을 보이고 있다. 하지만 하이브리드 기법의 결과가 M-Sum 기법의 결과에 근접한 결과를 보이고 하이브리드 기법의 실제 오차 역시 실제 오차와 큰 차이를 보이고 있지 않다.

실험 결과와 같이 M-Max 기법보다는 M-Sum 기법을 이용하는 것이 같은 공간을 사용하면서 보다 효율적인 오차 추정을 할 수 있는 방법이다. 또, 하이브리드 기법이 M-Sum 기법 결과에 크게 벗어나지 않으므로 극단적으로 작은 질의가 대부분인 경우 하이브리드 기법을 사용한다면 경우에 따라 정확한 오차 추정을 할 수 있을 것이다.

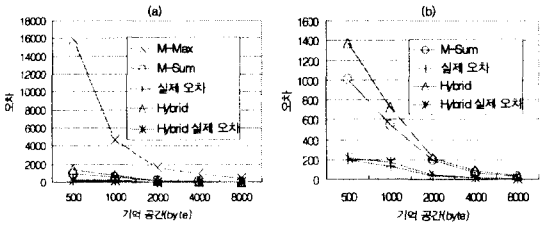


그림 6. 2차원 히스토그램에서 기억 공간 크기에 따른 오차 추정값

## 5 결 론

이 논문에서는 다차원 히스토그램에서 범위 질의의 선택도에 대한 오차 추정 기법을 제시하였다. 이 중 M-Sum 기법은 작은 저장 공간에서도 데이타의 분포를 정확하게 반영한다. 실험 결과에서 알 수 있듯이, 다른 기법들에 비해 M-Sum 기법이 범위 질의에 대한 실제 오차에 더 가까운 오차 추정값을 보이고, 또 저장 공간이 작은 경우에도 매우 정확한 결과를 보인다.

## 참고 문헌

[1] Viswanath Poosala, Yannis E. Ioannidis, Peter J. Haas, and Eugene J. Shekita. "Improved Histograms for Selectivity Esitmation of Range Predicates," SIGMOD, pp. 294-305, May 1996.

[2] M. Muralikrishna and David J. DeWitt. "Equi-Depth Histograms For Estimating Selectivity Factors For Multi-Demensional Queries," SIGMOD, pp. 256-276, 1988.

[3] Viswanath Poosala and Yannis E. Ioannidis. "Selectivity Estimation Without the Attribute Value Independence Assumption," VLDB, pp. 486-495, August 1997.

[4] H.V. Jagadish, Nick Koudas, S. Muthukrishnan, Viswanath Poosala, Ken Sevcik, and Torsten Suel. "Optimal Histograms with Quality Guarantiees," VLDB, pp. 275-286 1998.

[5] 안성준, 배진욱, 심마로, 이석호. "히스토그램을 이용한 근사적 집단 연산과 효과적인 오차 추정," 정보과학회 가을 학술발표논문집(I), 1999.

[6] CFD data sets, "http://www.cs.du.edu/~leut/MultiDimData.html"