論文96-33B-9-3

계층 그리드 화일을 이용한 선택률 추정에서 발생되는 오차 분석

(Analyzing Errors in Selectivity Estimation Using the Multilevel Grid File)

金 尙 煜 *, 黃 煥 圭 *, 黃 奎 永 **

(Sang-Wook Kim, Whan-Kyu Whang, and Kyu-Young Whang)

요 약

본 논문에서는 계층 그리드 화일을 이용한 다차원 선택률 추정에서 발생되는 오차에 관하여 논의한다. 먼저, 선택률 추정시 계층 그리드 화일의 각 단계 디렉토리 엔트리가 나타내는 영역내에서 레코드들이 균일하게 분포한다는 균일 분포 가정이 추정 오차의 근본적인 원인이 됨을 보이고, 영역내에서의 데이타 분포, 레코드의 수, 페이지의 크기, 질의 영역의 크기, 계층 그리드 화일의 디렉토리 단계 등 추정 오차에 영향을 미치는 요소들을 파악한다. 또한, 다양한 실험을 통하여 각 요소의 변화에 따르는 추정 오차의 경향을 제시한다. 실험 결과에 의하면, (1) 영역내에서의 데이타 분포가 균일할수록, (2) 저장된 레코드의 수가 많을수록, (3) 페이지의 크기가 작을수록, (4) 선택률 추정을 위한 질의 영역의 크기가 클수록, (5) 선택률 추정시 데이타 분포 정보로서 사용되는 계층 그리드 화일의 디렉토리 단계가 낮을수록 추정 오차는 작아지는 것으로 나타났다. 끝으로 이러한 요소들과 추정 오차간의 근원적인 관계를 함축하는 Granule Ratio를 정의하고, 실험을 통하여 Granule Ratio 값의 변화에 따르는 추정 오차의 변화를 제시한다. 실험 결과에 의하면, 여러가지 요소들의 값의 변화에도 불구하고 같은 Granule Ratio의 값을 갖는 경우에는 추정 오차가 거의 유사한 경향을 가지는 것으로 나타났다.

Abstract

In this paper, we discuss the errors in selectivity estimation using the multilevel grid file(MLGF). We first demonstrate that the estimation errors stem from the uniformity assumption that records are uniformly distributed in their belonging region represented by an entry in a level of an MLGF directory. Based on this demonstration, we then investigate five factors affecting the accuracy of estimation: (1) the data distribution in a region, (2) the number of records stored in an MLGF, (3) the page size, (4) the query region size, and (5) the level of an MLGF directory. Next, we present the tendancy of estimation errors according to the change of values for each factor through experiments. The results show that the errors decrease when (1) the distribution of records in a region becomes closer to the uniform one, (2) the number of records in an MLGF increases, (3) the page size decreases, (4) the query region size increases, and (5) the level of an MLGF directory employed as data distribution information becomes lower. After the definition of the Granule Ratio, the core formula representing the basic relationship between the estimation errors and the above five factors, we finally examine the change of estimation errors according to the change of the values for the Granule Ratio through experiments. The results indicate that errors tend to be similar depending on the values for the Granule Ratio regardless of the various changes of the values for the five factors.

* 正會員, 江原大學校 情報通信工學科
(Dept. of Information and Telecom. Eng., Kangwon National University)
** 正會員, 韓國科學技術院 電算學科
(Dept of Computer Science, KAIST)

※ 본 논문은 인공지능연구센터(CAIR) 95년도 위탁과제와 한국과학재단 96년도 핵심전문과제(과제 번호: 961-0903-019-1)의 연구비 지원에 의한 결과임.
接受日字:1996年4月8日, 수정완료일:1996年7月26日

Ⅰ. 서 론

선택률(selectivity)이란 "저장된 전체 레코드 수에 대한 질의 조건을 만족하는 레코드 수의 비" 로 정의된다. 질의 최적화 과정(query optimization)[1][7] 및 물리적 데이타베이스 설계 과정(physical database design) [Wha84] 에서는 선택률을 이용하여 질의를 만족하는 레코드의 수를 추정하며, 추정된 결과를 가지고 여러가지 질의 처리 방식에 대한 각각의 응답 시간을 예측한다. 그러므로 선택률의 정확한 추정은 질의 최적화 및 물리적 데이타베이스 설계 과정에서 필수적이다.

System R[7]에서 사용된 최초의 선택률 추정 기법은 저장된 레코드들이 최대값과 최소값 사이에서 균일하게 분포한다는 가정하에 선택률 추정을 위하여 (질의 조건을 만족하는 구간의 크기) / (전체 구간의 크기) 라는 공식을 이용하였다. 그러나 이 기법은 데이타의 실제 분포를 전혀 반영하지 않으므로 레코드들이 균일하게 분포하지 않는 경우에는 추정 오차가 커진다.

이러한 문제점을 해결하기 위하여 데이타의 실제 분포를 추정하고자 하는 연구가 이루어져 왔다[6][5][2]. 이러한 연구의 핵심적인 개념은 전체 구간을 여러개의 부분 구간으로 분할하고 각 구간내의 레코드 수를 기록해 둠으로써 데이타 분포에 대한 정보를 유지하는 것이다. 히스토그램 기법(histogram method)이라고 불리우는 이러한 기법들은 데이타 분포 정보를 이용함으로써 보다 정확하게 선택률을 추정할 수 있으며, 구간의 경계를 정하는 방법에 따라 equal-width 기법, equal-depth 기법, variable-width 기법 등 크게 세 가지로 분류된다[4]. 그러나 데이타의 분포를 나타내는 정보가 어느 한 순간에 구성된 정적인 정보이므로, 레코드의 삽입과 삭제가 빈번히 발생되는 동적인 상황에서는 데이타 분포 정보를 주기적으로 재구성해야 한다는 문제점이 있다[10].

참고 문헌 [10] 에서는 이러한 문제점을 피하기 위하여 다차원 동적 화일(multidimensional dynamic file)의 하나인 계층 그리드 화일(multilevel grid file)[9]을 이용하였다. 계층 그리드 화일에서는 각 단계의 디렉토리에 저장되는 엔트리들이 다차원 데이타 공간에서 차지하는 자신의 영역을 표현하므로 전체 데이타 공간의 분할 상태가 항상 각 단계의 디렉토리에 반영된다. 또한 이렇게 디렉토리내에 저장되는 데이타 공간의 분할 정보는 레코드의 삽입, 삭제 알고리즘에 의하여 분할 상태의 변화를 지속적으로 반영한다. 참고 문헌 [10] 에서는 각 디렉토리 단계에서 제공하는 데이타 공간 분할 상태의 정보를 이용하여 데이타 분포를 추정하는 방법과 이를 기반으로 다차원 선택률을 추정하는 기법을 제안하였다. 계층 그리드 화일의 디렉토리는 동적으로 변화되는 데이타의 분포를 지속적으로 반영하므로 이 기법은 기존 히스토그램 기법의 문제점인 재구성의 오버헤드를 해결할 수 있었다.

본 논문에서는 계층 그리드 화일을 이용한 다차원 선택률 추정 기법에서 발생하는 추정 오차에 관하여 논의하고자 한다. 먼저, 추정 오차의 발생 원인을 분석하고, 데이타 분포, 질의 영역의 크기, 레코드의 수, 계층 그리드 화일의 디렉토리 단계, 페이지의 크기 등 추정 오차에 영향을 미치는 요소들을 파악한다. 또한, 각 요소와 추정 오차간의 관계를 규명하고, 실험을 통하여 각 요소의 변화에 따르는 추정 오차의 경향을 제시한다. 실험을 위하여 균일 분포(uniform distribution), 정규 분포(normal distribution), 지수 분포(exponential distribution) 등의 분포를 취하는 데이타, 서로 다른 키간의 상관 관계를 가지는 데이타 등을 사용하였다. 끝으로 이러한 요소들과 추정 오차간의 근원적인 관계를 함축하는 공식 Granule Ratio를 정의하고, 실험을 통하여 Granule Ratio 값의 변화에 따르는 추정 오차의 변화를 제시함으로써 같은 Granule Ratio의 값을 갖는 경우에는 다른 요소들의 값의 변화와는 관계없이 추정 오차가 거의 유사한 경향을 가짐을 보인다.

본 논문의 구성은 다음과 같다. 제 2장에서는 관련 연구로서 계층 그리드 화일을 이용한 선택률 추정 기법에 대하여 논의한다. 제 3장에서는 추정 오차의 원인과 이에 영향을 미치는 요소들을 분석한다. 제 4장에서는 실험을 통하여 추정 오차와 요소간의 관계를 실험을 통하여 제시한다. 또한, Granule Ratio를 정의하고 이에 대한 추정 오차의 경향을 실험을 통하여 제시한다. 제 5장에서는 본 논문을 요약하고, 결론을 내린다.

Ⅱ. 계층 그리드 화일을 이용한 선택률 추정 기법

본 장에서는 참고 문헌 [10] 에서 제시한 계층 그리드 화일을 이용한 선택률 추정 기법에 관하여 기술

한다. 제 2.1절에서는 계층 그리드 화일(multilevel grid file)[9]의 특성에 대하여 소개하고, 제 2.2절에서는 계층 그리드 화일의 디렉토리 내에서 유지되는 데이타 분포 정보를 이용하여 선택률을 추정하는 방법에 대하여 설명한다.

1. 계층 그리드 화일

본 절에서는 계층 그리드 화일의 동적 특성과 구조적 특성에 대하여 간략히 설명한다.

1) 계층 그리드 화일의 동적 특성

계층 그리드 화일은 N 개의 키를 갖는 레코드들을 관리하는 다차원 동적 화일(multidimensional dynamic file)의 하나이며, 디렉토리와 데이타 페이지로 구성된다. 디렉토리 엔트리는 데이타 공간내의 영역과 일대일 대응 관계를 가지며, 이러한 디렉토리 엔트리들의 집합으로 구성되는 디렉토리는 여러 영역으로 분할된 데이타 공간의 상태를 반영한다. 데이타 페이지는 영역과 일대일 대응 관계를 가지며, 대응되는 영역내에 속하는 레코드들만을 저장한다. 따라서 계층 그리드 화일에서는 영역, 디렉토리 엔트리, 데이타 페이지가 각각 일대일의 대응 관계를 갖는다.

계층 그리드 화일은 레코드가 데이타 공간에 삽입되고 삭제되는 상황에 대하여 분할과 병합을 반복함으로써 동적 변화에 적응한다. 레코드가 삽입되는 경우, 레코드가 갖는 N 개의 키값을 해쉬하여 그 레코드가 속하는 영역을 찾게 되고, 그 영역에 할당된 데이타 페이지에 레코드를 삽입하게 된다. 이 결과로 데이타 페이지의 용량이 초과되면(overflow), 해당 영역은 같은 크기를 갖는 새로운 두 영역으로 분할되고 새로운 데이타 페이지가 하나 더 할당된다. 기존의 데이타 페이지에 있던 레코드들은 분할된 두 영역의 분할 경계값을 기준으로 각각의 영역에 할당된 두 데이타 페이지에 분산된다. 레코드의 삭제하는 경우에는 삽입 과정의 역에 해당되는 작업이 수행된다.

2) 계층 그리드 화일의 구조적 특성

계층 그리드 화일의 디렉토리내에 저장되는 엔트리의 구조에 대하여 살펴보자. 계층 그리드 화일에서는 영역들의 모양과 크기가 불규칙적이므로 각 영역에 대응되는 엔트리는 해당 영역의 위치뿐만 아니라 모양과 크기에 대한 정보도 유지해야 한다. N 차원 데이타 공간내의 한 영역을 표현하는 엔트리는 N+1 개의 필드로 구성된다. 각각의 키와 대응되는 N 개의 필드는 리전 벡터(region vector)라 정의되며, 대응되는 영역의 위치, 모양 및 크기에 대한 정보를 갖는다. 나머지 한 개의 필드는 이 영역에 속한 레코드들이 저장되어 있는 데이타 페이지에 대한 포인터이다. 리전 벡터는 각 키에 대응되는 N 개의 해쉬값으로 구성된다. 한 엔트리의 리전 벡터에서 i 번째 해쉬값은 그 엔트리의 영역내에 속하는 모든 레코드들의 i 번째 키를 해싱하였을 때 나타나는 해쉬값들의 공통 접두어(prefix)가 된다.

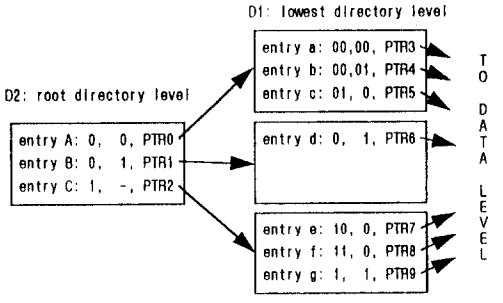
전체 디렉토리 엔트리들이 단순히 일차원의 배열 구조로 저장된다면, 하나의 레코드가 어느 영역에 속하는가를 알아보기 위해 최악의 경우 모든 엔트리들을 조사하여야 한다. 계층 구조는 이러한 전체 탐색의 비효율성을 제거하여 준다. 즉, 같은 영역에 속하는 레코드들을 같은 데이타 페이지에 넣고 이 영역을 표현하는 엔트리를 최하위 단계 디렉토리 D1에 유지하는 것과 마찬가지 방식으로 데이타 공간상에서 인접한 D1내의 몇 개의 엔트리들을 같은 페이지에 저장하고, 이 엔트리들이 나타내는 영역들을 모두 포함하는 보다 큰 영역의 엔트리를 D1의 상위 단계 디렉토리 D2에 유지시키는 것이다. 같은 방식으로 D2의 상위에 디렉토리 D3를 두게되며, 이러한 작업은 루트(root)라 정의되는 최상위 디렉토리의 엔트리들이 하나의 페이지내에 유지될 수 있을 때까지 반복된다[1].

그림 1(a)는 두개의 키를 갖는 이단계 계층 그리드 화일의 전체 디렉토리 구조를 표현한 것이다. 그림 1(b)와 그림 1(c)는 각각 그림 1(a)의 디렉토리 D1과 D2가 나타내는 데이타 공간의 분할 상태를 도면화한 것이다. 그림 1(b)와 그림 1(c)내의 사각형들은 각각 D1과 D2의 엔트리와 대응되는 영역들을 나타낸 것이며, 내부의 문자는 해당 엔트리를 의미한다. 최하위 단계 디렉토리 D1을 위한 디렉토리인 동시에 루트 디렉토리인 D2의 한 엔트리 A는 첫번째 키와 두번째 키의 해쉬값의 접두부가 각각 '0', '0'인 레코드들이 속한 영역 (0, 0)을 나타내며, 이 영역은 D1에서 디렉토리 엔트리 a, b, c가 나타내는 세개의 영역 (00, 00), (00, 01), (01, 0)으로 다시 세분된다. 따라서 상위 디렉토리에서 하위 디렉토리로 내려올수록 보다 구체화된 데이타 공간의 분할 상태가 나타난다[2].

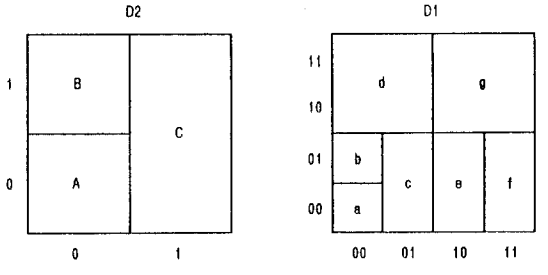
---

1) 이러한 구조적 특징은 계층 그리드 화일을 위한 삽입 및 삭제 알고리즘[9]에 의하여 자연스럽게 유지된다.

2) 그림 1(c)에서 나타나는 D2내 세번째 엔트리 C의 기



(a) 이단계 계층 그리드 화일의 디렉토리 구조



(b) D1내의 엔트리들이 나타내는 영역들

(c) D2내의 엔트리들이 나타내는 영역들

그림 1. 계층 그리드 화일의 구조
Fig. 1. Structural Characteristics of the Multi-level Grid File.

2. 계층 그리드 화일을 이용한 선택률 추정 기법

본 절에서는 계층 그리드 화일을 이용한 선택률 추정 기법[10]에 대하여 논의한다. 먼저 계층 그리드 화일의 각 단계 디렉토리가 동적인 환경에서도 지속적으로 데이타 분포 정보를 유지할 수 있음을 보이고, 이를 이용하여 선택률을 추정하는 방법을 제시한다.

1) 데이타 분포의 추정

계층 그리드 화일에서는 레코드의 양과 분포의 변화에 따라 자신의 디렉토리 구조를 동적으로 적응시키므로, 항상 삽입된 레코드들에 의한 데이타 공간의 분할 상태가 각 단계 디렉토리에 반영된다. 이러한 데이타 공간의 분할 상태 정보를 근거로 분할이 많이 발생되어 크기가 작은 영역에는 많은 수의 레코드들이 분포됨을 예측할 수 있고, 반대로 분할이 적게 발생되어 크기가 큰 영역에는 적은 수의 레코드들이 분포됨을 예측할 수 있다. 따라서 계층 그리드 화일에서는 각 단계 디렉토리에서 반영되는 데이타 공간의 분할 상태를 이용하여 저장된 데이타의 분포를 추정할 수 있다.

최하위 디렉토리 D1을 도면화시킨 그림 1(b)를 보자. 영역을 나타내는 사각형은 디렉토리 엔트리와 일대일 대응되고, 각 엔트리에는 하나의 데이타 페이지가 할당되어 있으므로, 각 영역에는 하나의 데이타 페이지가 할당되어 있다. 실제 이 영역에 속하는 레코드 수를 대응되는 엔트리에 유지하면, 영역의 조밀도는 영역내의 레코드 수에 비례하고 그 영역의 크기에 반비례하게 된다. 따라서 최하위 단계 디렉토리 엔트리들은 데이타의 분포를 나타내 주는 척도로서 사용될 수 있다. 다음은 루트 디렉토리 D2를 도면화시킨 그림 1(c)를 살펴보자. 영역을 나타내는 사각형은 루트 디렉토리 D2의 엔트리와 일대일 대응되고, 각 엔트리에는 하위 단계 디렉토리 D1의 페이지가 하나씩 할당되어 있다. 그러므로 최하위 단계 디렉토리에서의 방식과 마찬가지로 실제 이 영역에 속하는 레코드 수를 대응되는 엔트리에 유지함으로써 루트 디렉토리내의 엔트리들도 데이타의 분포 상태를 나타내 주는 척도로서 사용될 수 있다. 일반적으로 계층 그리드 화일이 L 단계의 디렉토리를 가질 때, 시스템에 저장된 데이타의 분포는 L개의 디렉토리 단계에 모두 동적으로 반영된다.

2) 선택률의 추정

먼저, 선택률 추정 공식에서 사용되는 두가지 용어를 정의한다. 디렉토리 엔트리가 나타내는 영역들 중에서 질의 영역(query region) [10] 내에 완전히 포함되는 것을 full 영역(full region)이라 정의하고, 엔트리가 나타내는 영역들 중에서 질의 영역내에 일부분만이 포함되는 것을 partial 영역(partial region)이라 정의하자. 그림 2는 질의 영역, full 영역, 그리고 partial 영역간의 관계를 보여준다.

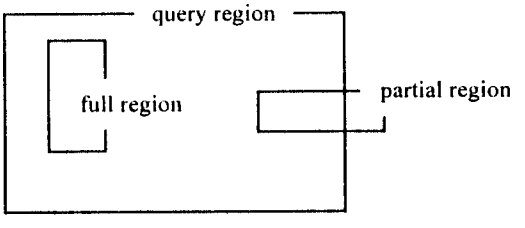


그림 2. 질의 영역, full 영역, partial 영역간의 관계
Fig. 2. A Full Region and Partial Region versus a Query Region.

호 '-'는 두번째 키 차원이 아직 분할되지 않은 전체 구간을 의미한다[9].

디렉토리내에서 제공되는 정보만으로는 한 영역내에서의 데이타 분포 파악이 불가능하므로 선택률의 추정을 위하여 다음과 같은 가정을 세운다.

가정 A: 하나의 디렉토리 엔트리가 표현하는 영역내에서 레코드들은 균일하게 분포한다.

참고 문헌 [10]에서 제안하는 선택률 추정 기법에서는 가정 A에 의하여 다음과 같은 공식으로 선택률을 계산한다.

$$Selectivity(query) = \frac{\sum_{i=1}^{f} count(i) + \sum_{j=1}^{p} (count(j) \times fraction(j))}{N}$$

공식에서 사용된 변수들의 의미는 다음과 같다. 변수 f 와 p는 각각 질의 영역에 포함되는 full 영역과 partial 영역의 수를 의미하며, N은 시스템에 저장된 전체 레코드의 수를 의미한다. 또한 count(i)와 count (j)는 각각 디렉토리 엔트리 i와 j가 나타내는 영역내에 포함되는 레코드의 수를 의미한다. 끝으로, fraction(i)는 다음과 같은 식으로 정의된다.

$$fraction(i) = \frac{size\ of\ (query\ region \cap i\text{-}th\ partial\ region)}{size\ of\ (i\text{-}th\ partial\ region)}$$

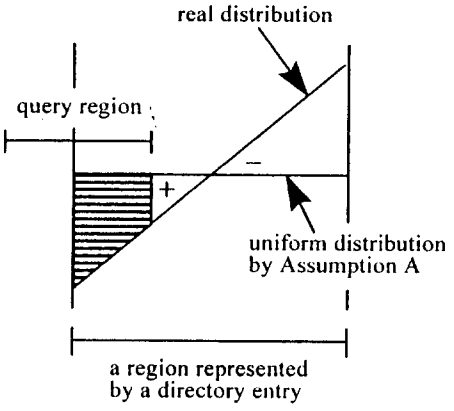
Ⅲ. 추정 오차 분석

본 절에서는 계층 그리드 화일을 이용한 다차원 선택률 추정 기법으로 구한 선택률 추정값의 오차에 관하여 논의하고자 한다. 먼저, 제 3.1절에서는 추정 오차의 발생 원인을 지적하고, 제 3.2절에서는 추정 오차에 영향을 미치는 요소들을 제시한다.

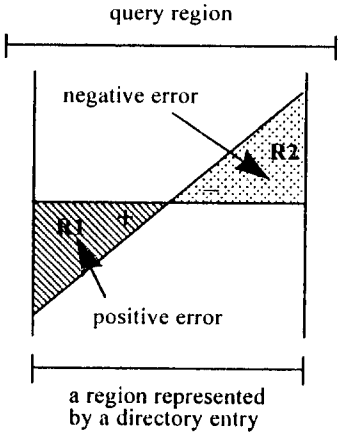
1. 추정 오차의 발생 원인

제 2.2절에서 세운 가정 A에 의하여 선택률 추정시 엔트리가 나타내는 각 영역내에서는 레코드들의 균일 분포를 가정한다. 그러나 실제로 레코드들이 한 영역내에서 항상 균일하게 분포되지는 않는다. 따라서 영역내에서 가정된 균일 분포와 실제 분포 사이에는 차이가 존재하는데 이것이 선택률 추정 오차의 원인이 된다.

보다 쉬운 이해를 위하여 일차원 영역으로 표현된 그림 3을 예로 살펴보자. 그림 3(a)에서 나타나는 partial 영역내에는 실제 분포와 가정된 균일 분포간의 분포차가 존재하게 되며, 이중 질의 영역에 포함되는 빗금친 부분이 질의 영역에 대한 오차에 해당된다.



(a) 추정 오차가 발생하는 경우



(b) 추정 오차가 발생하지 않는 경우

그림 3. 한 영역에서 균일 분포를 가정했을 경우의 추정 오차

Fig. 3. Estimation Errors Caused by Uniform Distribution Assumption.

그러나 엔트리가 나타내는 영역 전체가 질의 영역내에 완전히 포함되어 full 영역이 되면, 이 영역내에서의 분포차로 인한 오차는 0이 된다. 그림 3(b)는 이러한 이유를 설명한다. 그림 3(b)에서 엔트리가 나타내는 영역은 R1과 R2 두 부분으로 구성된다. 각 부분에서 발생되는 오차를 보면, R1에서는 실제 분포보다 추정 분포가 큰 값을 가지므로 양의 오차가 발생하게 된다. 반면, R2에서는 실제 분포보다 추정 분포가 작은 값을 가지므로 음의 오차가 발생하게 된다. 추정된 분포는 실제 분포의 평균값을 취하게 되므로 한 영역에서의 양의

오차와 음의 오차의 절대값의 크기는 같다. R1, R2 모두를 포함하는 질의 영역에서 이 엔트리가 나타내는 영역은 크기가 같은 양의 오차와 음의 오차를 동시에 가지므로 전체 오차는 0이 되는 것이다. 따라서 이러한 full 영역에서의 분포차로 인한 오차는 0이 되므로 질의 영역에 대한 추정 오차는 전적으로 partial 영역으로부터 기인한 것이다.

본 논문에서는 추정된 선택률의 오차의 정도를 측정하기 위한 성능 평가 지수로서 다음과 같이 정의되는 상대 추정 오차(relative estimation error)를 사용하였다. 앞으로 이 공식으로 구해진 추정 상대 오차를 간략히 추정 오차라 부르고자 한다.

$$\text{상대 추정 오차} = \frac{(\text{추정된 선택률} - \text{실제 선택률})}{\text{실제 선택률}}$$

2. 추정 오차에 영향을 미치는 요소

다음은 추정 오차에 영향을 미치는 요소들에 대하여 살펴본다. 이러한 요소들은 영역내에서의 데이타 분포, 저장된 레코드의 수, 페이지의 크기, 질의 영역의 크기, 디렉토리의 단계 등이다. 각각의 요소와 추정 오차간의 관계를 보면 다음과 같다.

요소 1: 영역내에서의 데이타 분포

영역내에서의 레코드의 분포가 균일할수록 추정 오차는 작아진다. 영역내에서의 분포가 균일하다면, 가정 A를 만족하게 되므로 추정 오차 발생의 근본적인 원인이 제거되는 셈이 된다. 따라서 영역내에서의 레코드들의 분포가 균일 분포에 가까울수록 추정 오차는 작아진다.

요소 2: 저장된 레코드의 수

시스템에 저장된 레코드의 수가 많을수록 추정 오차는 작아진다. 삽입되는 레코드의 수가 많아질수록 계층 그리드 화일에 할당되는 데이타 페이지의 수가 많아지고, 이에 대응되는 디렉토리 엔트리의 수도 많아진다. 그 결과, full 영역과 partial 영역의 수가 증가한다. 이 때, 질의 영역의 둘레에 걸치는 partial 영역의 수는 일차원적으로 증가하는 반면, 질의 영역의 내부에 존재하는 full 영역의 수는 이차원적으로 증가한다. 따라서 질의 영역이 커질수록 추정 오차를 전혀 발생시키지 않는 full 영역이 partial 영역과 비교하여 상대적으로 질의 영역에 많이 포함되므로 추정 오차는 작아진다.

요소 3: 페이지의 크기

페이지의 크기가 작을수록 추정 오차는 작아진다. 페이지의 크기가 작아지면 하나의 페이지내에 저장 가능한 레코드의 수가 적어진다. 이 결과, 같은 수의 레코드들을 저장하기 위하여 필요한 페이지의 수가 많아지므로 페이지들과 대응되는 디렉토리 엔트리들의 수도 함께 증가한다. 따라서 요소 (2)에서 설명한 것과 같은 이유로 질의 영역에 포함되는 full 영역의 수가 partial 영역의 수에 비하여 상대적으로 많아지므로 추정 오차는 작아진다.

요소 4: 질의 영역의 크기

질의 영역의 크기가 커질수록 추정 오차는 작아진다. 이것은 질의 영역의 크기가 커짐에 따라 여기에 포함되는 full 영역의 수가 partial 영역의 수보다 요소 (2)에서 설명한 것과 같은 이유로 인하여 상대적으로 많아지기 때문이다.

요소 5: 디렉토리의 단계

데이타 분포 정보로서 하위 단계 디렉토리를 이용할수록 추정 오차가 작아진다. 하위 단계의 디렉토리로 내려갈수록 데이타의 분포를 제공해 주는 디렉토리 엔트리의 수가 디렉토리 페이지의 블럭킹 인수(blocking factor)[11]를 밑으로 하는 디렉토리 단계수의 지수 함수로 커지므로 질의 영역에 포함되는 full 영역의 수가 partial 영역의 수보다 상대적으로 많아진다. 따라서 추정 오차는 작아진다.

Ⅳ. 실 험

본 장에서는 실험을 통하여 제 3.2절에서 논의한 요소들이 추정 오차에 미치는 영향을 분석한다. 본 실험의 목적은 저장된 레코드의 수, 레코드의 분포, 질의 영역의 크기, 계층 그리드 화일의 디렉토리 단계, 페이지의 크기 등 여러가지 요소의 변화에 대하여 제안된 선택률 추정 기법의 추정 오차의 경향을 구체적으로 제시하는 것이다. 제 4.1절에서는 성능 평가를 위하여 사용된 실험 모델에 대하여 기술하고, 제 4.2절에서는 실험 결과를 제시한다.

1. 실험 모델

추정 오차 분석을 위한 실험은 다음의 세 단계를 걸쳐 수행된다.

단계 1: 레코드 생성 단계

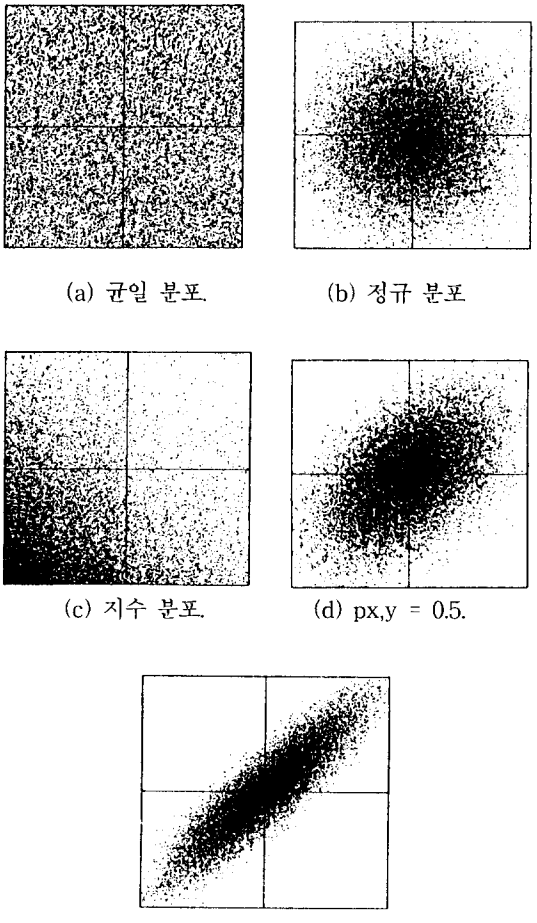
단계 2: 계층 그리드 화일 구성 단계

단계 3: 선택률 추정 및 오차율 계산 단계

단계 (1)에서는 데이타의 분포와 레코드의 수 등을 입력받아 해당되는 레코드들을 생성한다. 따라서 이 단계에서는 제 3.2절에서 논의한 요소 (1)과 요소 (2)에 변화를 줄 수 있다. 본 실험에서 사용된 레코드의 분포는 상관 관계가 전혀 없는 분포와 상관 관계가 있는 분포로 분류할 수 있다. 상관 관계가 없는 분포는 서로 다른 데이타 분포가 계층 그리드 화일을 이용한 선택률 추정시 추정 오차에 어떠한 영향을 미치는가를 분석하기 위한 것이다. 본 실험에서 사용된 레코드의 키 갯수는 두개이며, 차원간에 종속성 없이 [-2^{31}, $2^{31}-1$] 구간내의 균일 분포와 N(0, $2^{31}\times1/3$)의 정규 분포, 그리고 평균이 ($2^{32}\times1/4$)인 지수 분포를 취하도록 하였다[3]. 상관 관계가 있는 분포는 두개의 서로 다른 키간의 상관 관계가 계층 그리드 화일을 이용한 선택률 추정 기법의 추정 오차에 어떠한 영향을 미치는가를 분석하기 위하여 사용되었다. 상관 계수(correlation coefficient) $p_{x,y}$ 란 두 확률 변수 X와 Y간의 상관 관계를 나타내는 척도이다[13]. 본 실험에서는 표준 편차가 모두 $2^{31}\times1/3$를 갖는 두 정규 분포를 취하는 두 키간의 상관 계수가 각각 0.5와 0.9인 경우에 대하여 실험하였다. 그림 4는 본 실험에서 사용한 레코드들의 다섯가지 분포를 나타낸 것이다. 또한, 본 실험에서 사용한 레코드의 갯수로는 10,000, 20,000, 30,000, 40,000, 50,000의 5가지를 사용하였다.

단계 (2)에서는 특정한 페이지 크기를 갖는 계층 그리드 화일을 생성하고, 이곳에 단계 (1)에서 만들어진 레코드들을 삽입한다. 따라서 이 단계에서는 요소 (3)의 값에 변화를 줄 수 있다. 본 실험에서는 페이지의 크기로서 각각 128, 256, 512, 1024 바이트의 네가지를 사용하였다.[4]

단계 (3)은 단계 (2)에서 구성한 계층 그리드 화일을 대상으로 주어진 질의의 선택률을 추정하고, 이에 대한 추정 오차를 검증하는 단계이다. 따라서 이 단계에서는 질의 생성시 요소 (4)의 값에 변화를 줄 수 있으며, 선택률 추정시 요소 (5)의 값에 변화를 줄 수 있다. 사용된 질의 영역은 정방형이며, 포함하는 레코드의 수에 따라 선택률이 1/10인 대 영역(Large), 1/20인 중 영역(Medium), 1/100인 소 영역(Small), 1/1000인 극소 영역(Tiny)의 네가지로 분류하였다. 질의 영역의 생성 방법은 데이타 공간에서 한 점을 무작위로 선택하여 그 점을 중심으로 초기 질의 영역을 생성하고, 이것이 원하는 수의 95-105% 사이의 실제 레코드들을 포함할 때까지 영역의 크기를 조정하는 작업을 반복하게 된다. 질의 영역이 확정되면, 계층 그리드 화일의 각 단계 디렉토리내의 데이타 분포 정보를 이용하여



(a) 균일 분포. (b) 정규 분포

(c) 지수 분포. (d) px,y = 0.5.

(e) px,y = 0.9.

그림 4. 실험에 사용된 다섯가지 데이타 분포.
Fig. 4. Five Dota Distributions Used in the Experiments.

3) 원래의 지수 분포에서는 모든 데이타가 양의 값만을 가진다. 본 실험에서는 음의 값도 고려하기 위하여 전체 분포를 음의 방향을 231 만큼 이동시킨 분포를 사용하였다.

4) 본 실험에서 사용한 레코드의 크기와 디렉토리 엔트리의 크기가 각각 16, 18 바이트이므로 페이지내의 메타 정보를 고려한 블러킹 인수(blocking factor) [11]는 네가지 경우에서에서 각각 (7, 6), (15, 14), (31, 28), (63, 56)로 나타난다.

생성된 질의 영역의 선택률을 추정한다. 본 실험에서 사용된 디렉토리 단계는 최하위인 단계 1에서 최상위인 단계 7까지의 일곱 단계를 이용하였다.

2. 실험 결과

본 절에서는 저장된 레코드의 수, 레코드의 분포, 질의 영역의 크기, 계층 그리드 화일의 디렉토리 단계, 페이지의 크기 등 여러가지 요소의 변화에 대하여 제안된 선택률 추정 기법의 추정 오차의 경향을 실험을 통하여 제시한다.

실험 1은 계층 그리드 화일내에 저장된 레코드 수의 변화에 따라 추정 오차가 어떻게 변화하는가를 규명하기 위한 것이다. 그림 5는 다섯 가지 데이타 분포 각각에 대하여 10,000개에서 50,000개까지의 레코드들을 갖는 다섯 가지 계층 그리드 화일을 구성한 후 실험한 결과를 나타낸 것이다. 페이지의 크기는 128 바이트를 사용하였으며, 구성된 계층 그리드 화일의 최하위 단계를 이용하여 각각 1,000개의 중 영역(Medium) 질의에 대한 선택률을 추정하였다. 가로 축은 레코드의 수를 나타내고, 세로 축은 평균 상대 오차를 % 단위로 나타낸 것이다. 그림에서 나타난 uu, nn, ee는 각각 균일 분포, 정규 분포, 지수 분포를 취하는 데이타의 집합을 의미한다. 또한, co_0.5와 co_0.9는 두 키간의 상관 계수가 각각 0.5와 0.9인 데이타의 집합을 의미한다.

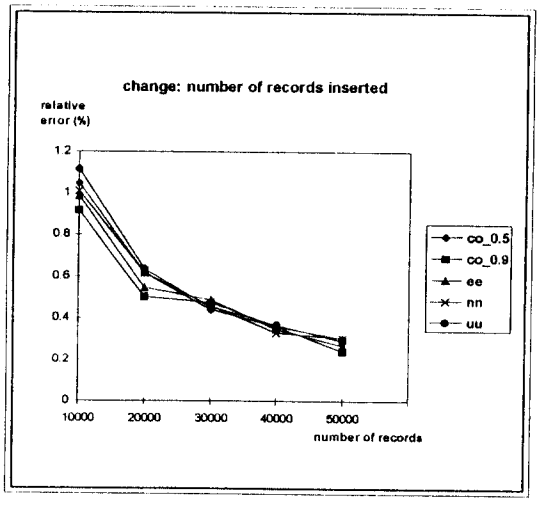


그림 5. 레코드 수의 변화에 따르는 상대 오차의 변화

Fig. 5. Tendancy of Estimation Errors According to the Change of the Number of Records.

10,000개의 레코드 집합에 대해서는 데이타 분포에 따라 약 0.9%에서 1.1%의 상대 오차를 보였다. 각 데이타 분포에 따라 다른 결과를 보이는 것은 질의 영역과 교차되는 partial 영역내에서의 분포와 가정 A의 균일 분포의 차이의 정도로 인한 것이다. 레코드의 수가 증가할수록 추정 오차가 점점 감소하여 50,000개의 레코드 집합에 대해서는 데이타 분포에 따라 약 0.25%에서 0.3%의 상대 오차를 보였다. 이것은 제 4.2절에서 언급한 바와 같이 레코드 수가 증가할수록 같은 질의 영역내에 포함되는 full 영역의 수가 partial 영역의 수에 비하여 상대적으로 증가하기 때문이다. 특히 레코드의 수가 30,000 이상인 경우에서는 데이타 분포차로 인한 상대 오차의 변화가 거의 없는 것으로 나타났다.

실험 2는 계층 그리드 화일을 위하여 사용되는 페이지 크기의 변화에 따라 추정 오차가 어떻게 변화하는가를 규명하기 위한 것이다. 그림 6은 다섯 가지 데이타 분포 각각에 대하여 128 바이트에서 1,024 바이트까지의 페이지 크기를 갖는 네 가지 계층 그리드 화일을 구성한 후 실험한 결과를 나타낸 것이다. 레코드의 수는 50,000개를 사용하였으며, 구성된 계층 그리드 화일의 최하위 단계를 이용하여 각각 1,000개의 중 영역(Medium) 질의에 대한 선택률을 추정하였다.

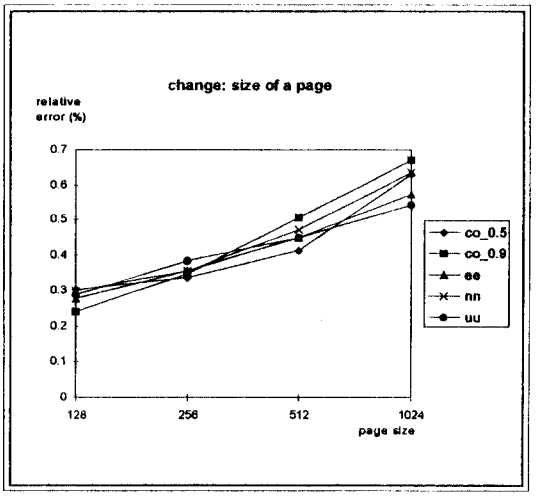


그림 6. 페이지 크기의 변화에 따르는 상대 오차의 변화

Fig. 6. Tendancy of Estimation Errors According to the change of the Page Size.

페이지의 크기가 128 바이트인 경우에서는 데이타 분포에 따라 약 0.24%에서 0.3%의 평균 상대 오차를

보였다. 페이지의 크기가 증가할수록 추정 오차도 점점 증가하여 1,024 바이트인 경우에는 데이타 분포에 따라 약 0.54%에서 0.67%의 상대 오차를 보였다. 이것은 페이지의 크기가 증가할수록 같은 질의 영역내에 포함되는 full 영역의 수가 partial 영역의 수와 비교하여 상대적으로 감소하기 때문이다. 페이지의 크기가 512 바이트인 경우와 1,024 바이트인 경우에서는 데이타 분포차로 인한 상대 오차의 변화가 비교적 큰 것으로 나타났다.

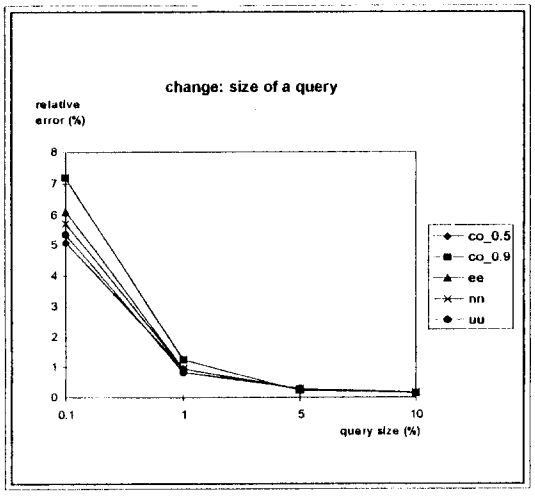


그림 7. 질의 영역 크기의 변화에 따르는 상대 오차의 변화

Fig. 7. Tendancy of Estimation Errors According to the Change of the Query Size.

실험 3은 선택률 추정시에 사용된 질의 영역 크기의 변화에 따라 추정 오차가 어떻게 변화하는가를 규명하기 위한 것이다. 그림 7은 다섯 가지 데이타 분포 각각에 대하여 128 바이트의 페이지 크기를 갖는 네 가지 계층 그리드 화일을 구성한 후 포함하는 레코드의 수에 따라 선택률이 1/10인 대 영역(Large), 1/20인 중 영역(Medium), 1/100인 소 영역(Small), 1/1000인 극소 영역(Tiny) 각 1,000개의 질의에 대하여 선택률을 추정한 실험 결과를 나타낸 것이다. 레코드의 수는 50,000개를 사용하였으며, 구성된 계층 그리드 화일의 최하위 단계를 이용하였다.

질의 영역이 극소인 경우에서는 데이타 분포에 따라 약 5.1%에서 7.2%의 상대 오차를 보였다. 질의 영역의 크기가 증가할수록 추정 오차는 점점 감소하여 대 영역인 경우에는 데이타 분포에 따라 약 0.16%에서 0.17%의 작은 상대 오차를 보였다. 이것은 질의 영역의 크기가 증가할수록 여기에 포함되는 full 영역의 수가 partial 영역의 수와 비교하여 상대적으로 증가하기 때문이다. 특히, 중 영역과 대 영역의 경우에서는 데이타 분포차로 인한 상대 오차의 변화가 거의 없는 것으로 나타났다.

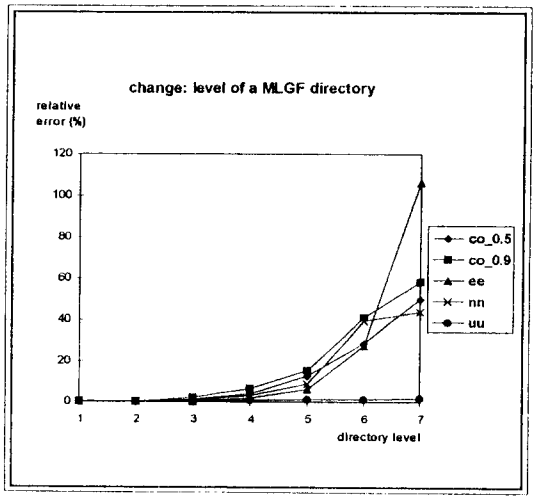


그림 8. 디렉토리 단계수 변화에 따르는 상대 오차의 변화

Fig. 8. Tendancy of Estimation Errors According to the Change of the directory level of the Multilevel Grid File.

실험 4는 선택률 추정시에 사용된 계층 그리드 화일의 디렉토리 단계 수의 변화에 따라 추정 오차가 어떻게 변화하는가를 규명하기 위한 것이다. 그림 8은 다섯 가지 데이타 분포 각각에 대하여 50,000개의 레코드들을 128 바이트의 페이지 크기를 갖는 계층 그리드 화일에 삽입한 후, 각 1,000개의 중 영역 질의에 대하여 계층 그리드 화일의 각 단계 디렉토리내의 데이타 분포 정보를 이용함으로써 선택률을 추정한 실험 결과를 나타낸 것이다.

단계 1인 최하위 단계 디렉토리를 이용한 실험에서는 데이타 분포에 따라 약 0.28%에서 0.3%의 작은 상대 오차를 보였다. 사용된 디렉토리 단계 수가 증가할수록 추정 오차는 점점 증가하여 최상위 단계인 단계 7의 루트 디렉토리를 이용한 경우에는 데이타 분포에 따라 약 1.63%에서 106.72%의 큰 상대 오차를 보였다. 이것은 상위 단계의 디렉토리를 이용할수록 같은 질의 영역내에 포함되는 full 영역의 수가 partial 영역

의 수에 비하여 상대적으로 감소하기 때문이다. 특히, 루트 단계를 이용한 경우에서는 분포로 인한 차이가 매우 큰 것으로 나타났다. 이것은 루트 단계 디렉토리 내에는 데이타 분포 정보로서 사용되는 디렉토리 엔트리의 수가 매우 작기 때문이다. 그러나 이 경우에서도 가정 A와 일치하는 균일 분포 데이타의 집합을 이용한 실험에서는 1.63%의 매우 작은 상대 오차를 보였다.

3. Granule Ratio

제 4.2절의 네가지 실험을 통하여 추정 오차의 크기는 영역내에서의 데이타 분포, 레코드의 수, 계층 그리드 화일의 디렉토리 단계, 질의 영역의 크기, 페이지의 크기 등의 다섯 가지 요소에 의하여 영향을 받음을 볼 수 있었다. 본 절에서는 추정 오차에 영향을 미치는 이 요소들의 근원적인 특성을 함축하는 Granule Ratio를 정의하고, Granule Ratio와 추정 오차와의 관계를 실험을 통하여 규명하고자 한다.

제 3.2절에서 제시한 다섯 가지 요소들을 다음과 같은 세가지 범주로 분류할 수 있다. 첫번째 범주는 영역내에서의 데이타 분포로서 요소 (1)이 여기에 해당된다. 두번째 범주는 질의 영역의 크기로서 요소 (4)가 여기에 해당된다. 세번째 범주는 디렉토리 엔트리가 나타내는 영역의 크기로서 요소 (2), (3), (5)가 모두 여기에 해당된다. 제 4.2절의 실험 결과는 "영역내에서의 데이타 분포가 균일할수록, 디렉토리 엔트리가 나타내는 영역의 크기가 작을수록, 그리고 질의 영역의 크기가 클수록 추정 오차의 크기는 작아진다" 로 요약될 수 있다. 본 절의 목적은 이러한 특성들을 함축하는 정량화된 공식을 정의함으로써 오차에 미치는 요소들의 효과를 보다 일관된 형태로 표현하기 위한 것이다.

이를 위하여 엔트리 영역과 질의 영역의 크기간의 관계를 정립하는 Granule Ratio라는 다음과 같은 공식을 정의한다. Granule Ratio에는 데이타 분포를 제외한 네가지 요소들의 특성이 모두 포함되어 있다. 여기서 데이타 분포를 제외한 이유는 이에 대한 정량화에 특별한 기준이 마련되어 있지 않기 때문이다.

$$\textit{Granule Ratio} = \frac{\textit{average number of records in a region represented by a directory entry}}{\textit{number of records in a given query region}}$$

Granule Ratio의 값과 추정 오차의 크기와의 관계를 살펴본다. Granule Ratio이 작아지면 하나의 질의 영역내에 속하는 디렉토리 엔트리의 수가 많아진다. 엔트리의 수가 많아지면, 제 3.2절에서 기술한 바와 같이 full 영역의 수가 partial 영역의 수보다 상대적으로 많아진다. 따라서 Granule Ratio가 작아지면, 추정 오차도 작아진다.

Granule Ratio와 연관시켜 요소 (2), (3), (4), (5)의 의미를 해석해 본다. 요소 (2)인 레코드의 수가 많아지면, Granule Ratio의 분모 부분의 값이 커진다. 따라서 Granule Ratio의 값이 작아지므로 추정 오차도 작아진다. 요소 (3)인 페이지 크기가 작아지면, Granule Ratio의 분자 부분의 값이 작아진다. 따라서 Granule Ratio의 값이 작아지므로 추정 오차도 작아진다. 요소 (4)인 질의 영역의 크기가 커지면, Granule Ratio의 분모 부분의 값이 커진다. 따라서 Granule Ratio의 값이 작아지므로 추정 오차도 작아진다. 요소 (5)인 선택률 추정에 사용되는 계층 그리드 화일의 디렉토리 단계가 하위로 내려가면, Granule Ratio의 분자 부분의 값이 작아진다. 따라서 Granule Ratio의 값이 작아지므로 추정 오차도 작아진다.

실험 5는 요소 (2)에서 (4)까지의 특성을 함축하는 Granule Ratio의 값의 변화에 따라 추정 오차가 어떻게 변화하는가를 규명하기 위한 것이다. 여기서는 0.5%, 1%, 5%, 10%의 서로 다른 Granule Ratio 값에 대하여 각각의 소실험 ex1, ex2, ex3, ex4를 수행하였다.

ex1은 ***50,000개***의 정규 분포를 갖는 레코드들을 ***128 바이트***의 페이지를 사용하는 계층 그리드 화일내에 삽입한 후, ***최하위 단계 디렉토리***를 사용하여 선택률을 추정한 실험이다. 소실험 ex2, ex3, ex4는 소실험 ex1의 각 요소값을 변화시킨 것이다. ex2는 50,000개의 정규 분포를 갖는 레코드들을 128 바이트의 페이지를 사용하는 계층 그리드 화일내에 삽입한 후, ***최하위 단계 바로 상위***의 디렉토리를 사용하여 선택률을 추정한 실험이다. ex3은 ***10,000개***의 정규 분포를 갖는 레코드들을 128 바이트의 페이지를 사용하는 계층 그리드 화일내에 삽입한 후, 최하위 단계 디렉토리를 사용하여 선택률을 추정한 실험이다. ex4는 50,000개의 정규 분포를 갖는 레코드들을 ***256 바이트***의 페이지를 사용하는 계층 그리드 화일내에 삽입한 후, 최하위 단계 디렉토리를 사용하여 선택률을 추정한 실험이다. 각 소실험에서 0.5%, 1%, 5%, 10%의 Granule Ratio 값으로의 고정을 위하여 질의 영역의 크기를 조정함으로써 원하는 Granule Ratio 값을 취하도록 하였다.

그림 9는 실험 5의 결과를 나타낸 것이다. Granule

Ratio 값이 0.5%인 경우에는 데이타 분포에 따라 약 0.51%에서 0.58%의 작은 상대 오차를 보였다. Granule Ratio 값이 점차 증가할수록 예측한 바와 같이 상대 오차는 점점 증가하였으며, Granule Ratio 값이 10%인 경우에는 분포에 따라 약 5.3%에서 5.9%의 상대 오차를 보였다. 특히, 네가지 소실험에서 사용한 요소 (2), (3), (4), (5)가 서로 다름에도 불구하고, Granule Ratio의 값이 같은 경우에는 거의 같은 추정 오차를 보였다. 따라서 이 Granule Ratio는 네가지 요소의 특성을 모두 합축하는 공식임을 알 수 있다.

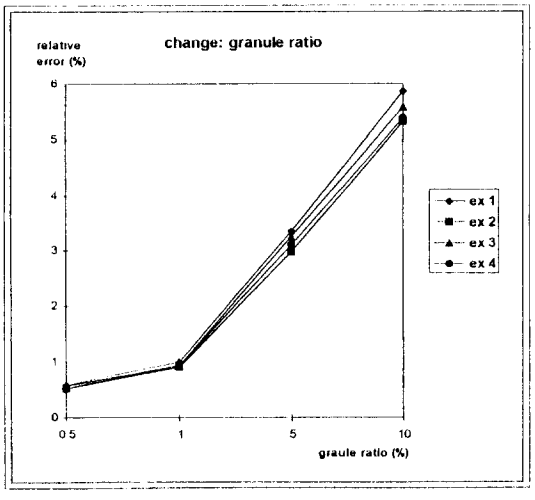


그림 9. Granule Ratio 변화에 따르는 상대 오차의 변화

Fig. 9. Tendancy of Estimation Errors According to the Change of the Granule Ratio.

V. 결 론

선택률이란 "저장된 전체 레코드 수에 대한 질의 조건을 만족하는 레코드 수의 비"로 정의되며, 질의 최적화 과정 및 물리적 데이타베이스 설계에서 필수적인 요소로서 사용된다. 참고 문헌 [10]에서는 다차원 동적 화일의 하나인 계층 그리드 화일을 이용하여 선택률을 추정하는 기법을 제안하였다. 이 기법에서는 계층 그리드 화일의 각 단계 디렉토리에서 제공하는 데이타 분포 정보를 기반으로 선택률을 추정하므로 정확성을 높일 수 있으며, 특히 이러한 데이타 분포 정보가 레코드의 삽입 및 삭제시의 상황에 따라 지속적으로 유지되는 동적인 정보이므로 재구성의 오버헤드가 없다는 것이 장점이다.

본 논문에서는 계층 그리드 화일을 이용한 다차원 선택률 추정 기법에서 발생되는 추정 오차에 관하여 논의하였다.

먼저, 추정 오차의 발생 원인을 분석하였다. 계층 그리드 화일을 이용한 선택률 추정 기법에서는 선택률 추정시 각 단계 디렉토리 엔트리가 나타내는 영역내에서 레코드들이 균일하게 분포한다는 가정을 이용한다. 따라서 실제 분포와 가정한 균일 분포간에는 차이가 존재하는데 이것이 추정 오차의 근본적인 원인이 됨을 보였다.

또한, 영역내에서의 데이타 분포, 레코드의 수, 페이지의 크기, 질의 영역의 크기, 계층 그리드 화일의 디렉토리 단계 등 추정 오차에 영향을 미치는 요소들을 파악하고 각 요소와 추정 오차간의 관계를 규명하였으며, 실험을 통하여 각 요소의 변화에 따르는 추정 오차의 경향을 제시하였다. 실험 결과를 요약하면 다음과 같다. (1) 영역내에서의 데이타 분포가 균일할수록, (2) 저장된 레코드의 수가 많을수록, (3) 페이지의 크기가 작을수록, (4) 선택률 추정을 위한 질의 영역의 크기가 클수록, (5) 선택률 추정시 데이타 분포 정보로서 사용되는 계층 그리드 화일의 디렉토리 단계가 낮을수록 추정 오차는 작아진다.

끝으로 이러한 요소들과 추정 오차간의 근원적인 관계를 합축하는 Granule Ratio를 정의하고, 실험을 통하여 Granule Ratio 값의 변화에 따르는 추정 오차의 변화를 제시하였다. 실험 결과에 의하면, 여러가지 요소들의 값의 변화에도 불구하고 같은 Granule Ratio의 값는 경우에는 추정 오차가 거의 유사한 경향을 가지는 것으로 나타났다.

Acknowlegment

※ 본 논문은 인공지능연구센터(CAIR) 95년도 위탁과제와 한국과학재단 96년도 핵심전문과제(과제 번호: 961-0903-019-1)의 연구비 지원에 의한 결과임.

참 고 문 헌

[1] Blasgen, M. W. and Eswaran, K. P., "Storage and Access in Relational Databases," *IBM Systems Journal*, Vol. 16, No. 4, pp. 363-377, 1977.

[2] Chen, M. C., McNamee, L., and Matloff, N.,

"Selectivity Estimation Using Homogeneity Measurement," In *Proc. Intl. Conf. on Data Engineering,* IEEE, Los Angeles, pp. 304-310, Feb. 1990.

[3] Law, A. M. and Kelton, W. D., *Simulation Modeling and Analysis,* McGraw-Hill Company, New York, 1982.

[4] Mannino, M. V., Chu, P., and Sagar, T., "Statistical Profile Estimation in Database Systems," *ACM Computing Surveys,* Vol. 20, No. 3, pp. 191-221, Sept. 1988.

[5] Muralikrishna, M. and DeWitt, D., "Equi-Depth Histograms for Estimating Selectivity Factors for Multi- Dimensional Queries," In *Proc. Intl. Conf. on Management of Data,* ACM SIGMOD, Chicago, pp. 28-36, June 1988.

[6] Piatetsky, S. G. and Connell, G., "Accurate Estimation of the Number of Tuples Satisfying a Condition," In *Proc. Intl. Conf. on Management of Data,* ACM SIGMOD, Boston, pp. 256-276, June 1984.

[7] Selinger, P. G. et al., "Access Path Selection in a Relational Database Management System," In *Proc. Intl. Conf. on Management of Data,* ACM SIGMOD, Boston, pp. 23-34, May 1979.

[8] Whang, K.-Y., Wiederhold, G. and Sagalowicz, D., "Seperability-An Approach to Physical Database Design," *IEEE Trans. on Computers,* Vol. C-33, No. 3, pp. 209-222, Mar. 1984.

[9] Whang, K. Y. and Krishnamurthy, R., "The Multilevel Grid File-A Dynamic Hierarchical Multi-dimensional File Structure," In *Proc. 2nd Intl. Conf. on Database Systems for Advanced Applications,* Tokyo, pp. 449-459, Apr. 1991.

[10] Whang, K. Y., Kim, S. W., and Wiederhold, G., "Dynamic Maintenance of Data Distribution for Selectivity Estimation," *The VLDB Journal,* Vol. No. pp. 1994.

[11] Wiederhold, G., *Database Design,* McGraw-Hill Book Company, New York, 1983, Second Edition.

저 자 소 개

金 尙 煜(正會員)

1966년 8월 8일생. 1989년 서울대학교 컴퓨터 공학과 공학사. 1991년 한국과학기술원 전산학과 공학석사. 1994년 한국과학기술원 전산학과 공학박사. 1991년 7월 ~ 8월 미 스탠포드 대학 방문 연구원. 1994년 2월 ~ 1995년 2월 한국과학기술원 정보전자연구소 Post ~ Doc. 1995년 3월 ~ 현재 강원대학교 정보통신공학과 전임강사. 주 관심 분야는 데이타 베이스 시스템, DBMS, 트랜잭션 프로세싱, 다차원 동적 화일, 공간 데이타베이스/GIS, 객체지향 데이타베이스, 멀티미디어 데이타베이스 등임

黃 奎 永(正會員)

1951년 3월 2일생. 1973년 서울대학교 전자과 졸업. 1975년 한국과학기술원 전기 및 전자공학과 졸업(M.S). 1982년 Stanford University (전산학과, M.S.). 1983년 Stanford University (전산학과, Ph.D.). 1975년 ~ 1978년 국방과학연구소 선임연구원. 1983년 ~ 1990년 IBM T. J. Watson Research Center, Research Staff Member. 1991년 Visiting Professor, Stanford University. 1992년 Visiting Associate Professor, Georia Insititute of Technology. 1993년 Hewlett~Packard Laboratories 기술자문 (Palo Alto). 1992년 ~ 1994년 한국정보과학회 데이터베이스연구회(SIGDB) 운영위원장. 1993년 ~ 현재 체신부 통신진흥협의회 DB산업육성분과 위원장. Editor The VLDB Journal. Editor: Distributed and Parallel Databases An International Journal. Associate Editor The IEEE Data Engineering Bulletin (1990 ~ 1993). Editor International Journal of Geographical Information Systems. 1990년 ~ 현재 인공지능연구센터 데이타베이스 및 멀티미디어 연구실장. 1990년 ~ 현재 한국과학기술원 전산학과 교수. 주 관심분야는 데이타 베이스 시스템, 멀티미디어 데이타베이스, 객체지향 데이타베이스, 하이퍼미디어, GIS, 분산데이타베이스 및 Client/Server기술, 연역 데이타베이스, 공학 데이타베이스, 사무자동화, CASE, 전문가 시스템 등임

黃 煥 圭(正會員)

1552년 5월 26일생. 1976년 서울대학교 공업교육학과(전자전공)공학사. 1987년 플로리다 대학 전기공학과 공학석사. 1992년 플로리다 대학 전기공학과 공학박사. 1994년 3월 ~ 현재 강원대학교 정보통신공학과 조교수. 주 관심 분야는 데이타베이스 시스템, 분산 데이타 베이스 등임