인터넷 질의 처리를 위한 웨이블릿 변환에 기반한 통합 요약정보의 관리

(Wavelet Transformation based Management of Integrated Summary Data for Internet Query Processing)

조 문 증 † 황 규 영 †† 김 상 욱 ††† 심 규 석 ††††
(Moon Jeung Joe) (Kyu-Young Whang) (Sang-Wook Kim) (Kyuseok Shim)

요 약 최근, 인터넷 기술의 급격한 발전으로 인하여 다수의 정보원들을 처리 대상으로 하는 인터넷 질의의 사용이 점차 확대되고 있다. 인터넷 질의 처리를 위해서는 여러 정보원들에 분산된 전체 데이타분포를 함축적으로 표현한 통합 요약정보가 필요하다. 본 논문에서는 웨이블릿 변환을 기반으로 한 통합 요약정보의 관리 및 이를 이용한 인터넷 질의 최적처리에 관하여 논의한다. 통합 요약정보의 구성을 위한 가장 단순한 방법은 각 정보원에 분산된 데이타분포들을 합병한 후, 이를 기반으로 통합 요약정보를 구성하는 것이다. 그러나 이 방법은 큰 용량의 데이타분포를 전송, 저장, 통합하는 비용이 매우 크므로 실용적이지 않다. 본 논문에서는 이러한 문제점을 극복하기 위하여 웨이블릿 변환을 기반으로 요약정보들을 합병함으로써 통합 요약정보를 구성하는 새로운 방법과 이를 이용한 인터넷 질의 최적화 방안을 제시한다. 웨이블릿 요약정보는 합병 조건을 만족하도록 변환되며, 합병 과정이 웨이블릿의 특성으로 인하여 매우 단순하다는 장점을 갖는다. 본 논문에서는 제안된 방법으로 구성된 통합 요약정보의 오차 상한선을 정량적으로 유도한다. 제안된 방법에 대한 실험 결과에 의하면, 히스토그램 요약정보의 합병과 웨이블릿 요약정보의 합병을 비교한 선택률 추정 실험은 통합 히스토그램에 비해 통합 웨이블릿 요약정보가 1.6 ~ 5.5배 더 정확하다는 결과를 보였다. 또한, 56개의 정보원이 참여하는 인터넷 top-N 질의를 처리할 때, 통합 요약정보를 사용하지 않는 방법과 비교하여 이를 사용하는 경우 약 44배의 성능 개선 효과를 보였다.

Abstract As Internet technology evolves, there is growing need of Internet queries involving multiple information sources. Efficient processing of such queries necessitates the integrated summary data that compactly represents the data distribution of the entire database scattered over many information sources. This paper presents an efficient method of managing the integrated summary data based on wavelet transform, and addresses Internet query processing using the integrated summary data. The simplest method for creating the integrated summary data would be to summarize the integrated data distribution obtained by merging the data distributions in multiple information sources. However, this method suffers from the high cost of transmitting, storing, and merging a large amount of data distributions. To overcome the drawbacks, we propose a new wavelet transform based method that creates the integrated summary data by merging multiple summary data and effective methods for optimizing Internet queries using it. A wavelet transformed summary data is converted to satisfy conditions for merging. Moreover, the merging process is very simple owing to the properties of the wavelet transform. We formally derive the upper bound of the errors of the wavelet transformed integrated summary data. Compared with the histogram-based integrated summary data, the wavelet transformed integrated summary data proves to be 1.6 ~ 5.5 times more accurate when used for selectivity estimation in experiments. In processing Internet top-N queries involving 56 information sources, using the integrated summary data reduces the processing cost to 1/44 of the cost of not using it.

· 본 연구는 첨단정보기술연구센터를 통하여 한국과학재단의 지원을 받았음.

† 비 회 원 : LG전자기술원 정보기술연구소
joemoon@mozart.kaist.ac.kr
†† 종신회원 : 한국과학기술원 전자전산학과 교수
kywhang@cs.kaist.ac.kr
††† 정 회 원 : 강원대학교 컴퓨터정보통신공학부 교수
wook@kangwon.ac.kr
†††† 정 회 원 : 한국과학기술원 전자전산학과 교수
shim@cs.kaist.ac.kr

논문접수 : 2000년 7월 27일
심사완료 : 2001년 6월 26일

1. 서 론

최근, 인터넷 기술의 급격한 발전으로 인하여 인터넷을 통한 데이타베이스 검색이 가장 보편적인 정보 검색의 방법으로 자리를 잡아가고 있다[1]. 본 논문에서는 인터넷 환경에서 데이타베이스 질의를 처리하는 효과적인 방안에 관하여 논의하고자 한다.

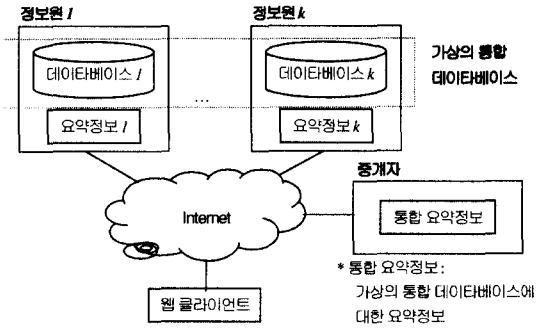


그림 1 인터넷 질의 처리를 위한 모델

그림 1은 인터넷에서의 질의 처리 모델을 나타낸 것이다[2]. 인터넷에는 다수의 **정보원**(*information source*)들이 존재한다. 각 정보원은 고유의 데이타베이스를 가지며, 고유의 질의 처리 기능을 지원한다. 사용자는 **웹 클라이언트**(*Web client*)를 통하여 미리 지정된 정보원에 대해 **인터넷 질의**(*internet query*)를 요청하고, 처리된 질의 결과를 받는다. 기존의 인터넷 응용에서는 단 하나의 정보원만을 처리 대상으로 하는 단순한 질의들이 주로 사용되었으나, 최근 응용 분야가 다양해짐에 따라 다수의 정보원들을 처리 대상으로 하는 질의의 사용이 점차 확대되고 있다[3]. 이러한 형태를 갖는 질의가 발생하는 예로는 여러 인터넷 쇼핑몰들로부터 주어진 조건을 만족하는 상품들을 찾는 응용을 생각할 수 있다.

다수의 정보원이 질의에 참여하는 경우에는 전체 질의 처리를 주관하는 **중개자**(*mediator*)가 필요하다[2]. 중개자는 (1) 웹 클라이언트로부터 인터넷 질의를 받고, (2) 이 질의를 수행시킬 정보원들을 선택하고, (3) 인터넷 질의를 선택한 정보원이 수행할 수 있는 로컬 질의(local query)로 변환하고, (4) 로컬 질의를 각 정보원에 전달한다. 또한, (5) 각 정보원이 수행한 질의의 결과를 취합하여, (6) 이를 통합한 질의 결과를 웹 클라이언트에게 반환한다[4]. 인터넷 질의는 다수의 정보원이 참여하게 되어 거대한 데이타베이스를 다루므로 이러한 인터넷 질의의 효과적인 처리는 매우 중요하다. 최근 들어 인터넷 질의에 관한 연구가 활발히 진행되고 있다. 이들 연구 중에는 인터넷 질의를 수행 시킬 정보원을 선택하는 문제, 인터넷 질의를 로컬 질의로 바꾸는 문제, 질의 결과를 통합하는 문제 등이 있다[5].

대부분의 데이타베이스 관리 시스템(Database Management System: DBMS)들은 효과적인 질의 처리를 위하여 **요약정보**(*summary data*)를 사용한다. 요약정보는 데이타베이스에 저장된 레코드들의 특정 속성 값에 대한 데이타분포를 함축적으로 표현한 것으로서, 해당 속성 값과 빈도의 쌍으로 이루어진다[6]. 요약정보는 질의 최적화(query optimization)[7]와 물리적 데이타베이스 설계(physical database design)[8]에서 선택률(selectivity)을 추정하는데 주로 사용되어 왔다. 최근에는 데이타베이스의 크기가 급격히 증가하면서 근사 질의(approximate query)[9], OLAP(on-line analytical processing)[10], Top-N 질의[11] 등 요약정보를 이용한 응용 영역이 확대되고 있다.

지금까지 제안된 데이타베이스로부터 요약정보를 생성하는 기법은 **매개 변수적 기법**(*parametric technique*)과 **비매개 변수적 기법**(*non-parametric technique*), **샘플링**(*sampling*)으로 나뉜다[6]. 매개 변수적 기법은 데이타베이스의 데이타분포에 대한 모델을 설정하고, 모델의 매개 변수를 추정하고, 매개 변수에 대한 연산으로 데이타분포를 추정하는 방법이다. 비매개 변수적 기법은 모델 없이 데이타분포를 일정한 조건으로 나누고, 나누어진 집단(cluster)을 합이나, 평균 같은 대표값으로 데이타분포를 요약하는 방식으로 히스토그램이 대표적인 방법이다. 샘플링은 데이타베이스 중 일정 양의 데이타를 추출하고, 추출된 데이타로부터 주어진 데이타베이스의 데이타분포를 추정하는 방식이다. 그러나 이들 연구들은 주로 단일 데이타베이스에서 요약정보를 구성하고 이를 이용하여 질의 처리를 수행하는데 초점을 맞추고 있다. 그러나, 하나가 아닌 다수의 정보원들이 관련된 인터넷 질의를 효율적으로 처리하는 방안에 관해서는 논의된 바 없다.

본 논문에서는 이 문제에 초점을 맞추어 인터넷 질의를 처리하기 위한 통합된 요약정보의 생성 및 관리 방안과 이를 이용한 효과적인 인터넷 질의 처리 방안을 제시하고자 한다. 본 연구에서는 정보원들의 데이타베이스들을 통합한 가상의 데이타베이스에 대한 요약정보를 **통합 요약정보**(*integrated summary data*)라 정의한다. 중개자는 이 통합 요약정보를 생성, 저장, 관리하며, 다수의 정보원들이 참여하는 인터넷 질의를 효과적으로 처리하기 위하여 이 정보를 사용하게 된다.

통합 요약정보의 생성을 위하여 고려할 수 있는 가장 단순한 방법은 데이타베이스 내의 특정 속성 값과 레코드 빈도 쌍으로 구성되는 데이타분포들을 각 정보원이 전송하고, 중개자가 이를 합병하여 통합 데이타분포를 만든 후, 이를 기반으로 통합 요약정보를 생성하는 것이다. 그러나 데이타분포는 매우 크므로 데이타분포의 전송, 저장, 통합 비용을 고려할 때 데이타분포를 이용한 합병 방법은 실용적이지 못하다. 또 다른 방법은 각 정보원이 전송한 요약정보들을 중개자가 합병하여 통합 요약정보를 생성하는 방법이다. 요약정보는 데이타분포와 비교하여 매우 작으므로 요약정보의 합병 방법은 전송, 저장, 통합 비용이 작다. 또한, 각 정보원은 질의 최적화의 목적으로 요약정보를 관리하고 있으므로 이를 활용할 수 있다는 장점을 갖는다. 이러한 실용성으로 인하여 본 논문에서는 통합 요약정보 관리를 위한 기본 전략으로서 요약정보를 이용한 합병 방법을 채택한다. 그러나 이 방법은 이미 오차를 포함하고 있는 요약정보를 합병하므로 데이타분포의 합병 방법에 비하여 통합 요약정보의 오차가 커지는 경향이 있다. 따라서, 통합 요약정보의 오차를 줄이기 위한 요약정보의 합병 방법이 필요하다.

본 논문은 웨이블릿 변환에 기반한 통합 요약정보를 관리하는 방법을 제안한다. 먼저, 임의의 요약정보들이 합병되기 위한 조건을 제시하고, 웨이블릿 변환(wavelet transform)[12]을 기반으로 하는 요약정보는 항상 이 조건을 만족되도록 변환됨을 보인다. 웨이블릿 기반 통합 요약정보의 합병은 웨이블릿 변환의 고유 특성인 선형성(linearity)과 이동성(shift)을 이용함으로써 통합 요약정보가 쉽게 생성될 수 있음을 보인다. 또한, 웨이블릿 요약정보 합병으로 인한 오차의 한계를 증명한다. 일반적으로 인터넷 환경에서 각 요약정보의 갱신 주기는 서로 다르다. 본 논문에는 통합 요약정보의 **점진적 갱신**(*incremental update*)을 제시함으로써 서로 다른 시점에 발생되는 요약정보의 변경을 통합 요약정보가 손쉽게 반영할 수 있는 기반을 제공한다. 제안된 방법에 대한 응용으로서 통합 요약정보를 이용한 인터넷 질의 최적화 방안과 인터넷 top-N 질의 처리 방안을 제시하고 통합 요약정보로 인해 얻는 효과에 대해 설명한다. 끝으로 다양한 실험을 통하여 제안된 방법의 우수성을 검증한다.

본 논문의 구성은 다음과 같다. 먼저, 제2절에서는 요약 정보에 관련된 용어를 정의하고, 요약정보의 합병에 관하여 설명한다. 제3절에서는 웨이블릿 요약정보를 정의한다. 제4절에서는 통합 웨이블릿 요약정보 생성 및 관리 방법을 제안하고, 오차에 관하여 논의한다. 제5절에서는 제안된 기법의 응용으로서 통합 요약정보를 이용한 인터넷 질의 처리 및 top-N 질의 처리 방안을 제시한다. 제6절에서는 실험 결과를 보이고, 이를 분석한다. 제7절에서는 논문의 공헌을 요약하고, 본 논문의 결론을 내린다.

2. 요약정보

2.1 용어 정의

요약정보는 데이타베이스 내에 저장된 특정 속성값의 데이타분포를 압축하여 표현한 정보이다. 본 논문에서는 정수형과 실수형 속성을 요약정보의 대상으로 한다. **도메인**(*domain*) D_X는 속성 X가 가질 수 있는 값들의 집합이다. **값 집합**(*value set*) V_X는 데이타베이스에 실제로 저장된 속성 X 값들의 집합이다($V_X \subseteq D_X$). $V_X = \{ v_i : 1 \le i \le |D| \}$라 할 때, **빈도**(*frequency*)) f_i는 값 v_i를 가지는 레코드의 수이고, **누적빈도**(*cumulative frequency*) c_i는 속성 X의 값이 v_i이하인 레코드의 수이다(즉, $c_i = \sum_{j=1}^{i} f_i$). **단순 데이타분포**(*simple data distribution*)는 v_i와 빈도 f_i의 쌍으로 이루어진 집합이고, **누적 데이타분포**(*cumulative data distribution*)는 v_i와 누적빈도 c_i의 쌍으로 이루어진 집합이다. **확장 누적 데이타분포**(*extended cumulative data distribution*)는 $D_X - V_X$에 대해 빈도를 0으로 할당하여 누적 데이타분포를 D_X로 확장한 것이다[13]. 이들 데이타분포들은 서로 변환이 가능하므로 통칭하여 데이타분포라 부른다. 본 논문에서는 데이타분포에 대한 요약정보를 다음과 같이 정의한다.

정의 1 : 데이타분포 T의 **데이타 분할**(*data partition*) P는 다음의 조건을 만족하는 요소 $P_i (1 \le i \le k)$의 집합이다:

1. $P_i \neq \phi$
2. $\bigcup_{i=1}^{k} P_i = T$
3. $(v_m, f_m) (\in P_i)$와 $(v_n, f_n) (\in P_j)$에 대해 $i \neq j$이면 $v_m \neq v_n$이다.

정의 2 : **요약범위**(*summarization range*) r_i는 P_i내의 요소 (v_m, f_m)들 중에서 v_m의 최소값과 최대값의 쌍이다.

정의 3 : **요약값**(*summarization value*) $\boldsymbol{c}_i$ $(\in C^n)$는 n차원의 값[1]이다. 여기서, C^n을 **요약값의 도메인** (*summarization value domain*) W라 한다.

1) 값은 실수 또는 복소수 도메인을 가진다.

정의 4 : 요약요소(*summarization element*)는 요약범위 r_i와 요약값 c_i의 쌍이다.

정의 5 : 요약함수(*summarization function*)는 데이타 분할의 요소 P_i를 요약요소 (r_i, c_i)로 사상하는 함수이다.

정의 6 : 요약정보(*summary data*)는 데이타 분할 P에 대해 요약함수를 적용하여 얻은 요약요소의 집합이다.

예를 들어, 요약정보의 일종인 *MaxDiff(V,A)* **히스토그램**(*histogram*)[14]은 데이타분포 T = {(1, 10), (2, 20), (3, 10), (4, 20), (5, 40), (6, 30)}를 P_1 = {(1, 10), (2, 20), (3, 10), (4, 20)}과 P_2 = {(5, 40), (6, 30)}으로 분할하고, 요약값으로 총빈도를 사용한다. 이 요약정보는 S = {([1:4], 60), ([5:6], 70)}으로 표기된다.

요약정보로 추정한 데이타분포를 **추정 데이타분포**(*approximate data distribution*)[15]라 한다. 요약정보는 데이타분포를 압축한 것이므로 추정 데이타분포와 실제 데이타분포 사이에는 오차가 존재한다.

2.2 요약정보의 합병

본 절에서는 요약정보의 합병을 정의하고, 합병 조건을 유도한다. **요약정보의 합병**은 다수의 요약정보들을 합하여 하나의 **통합 요약정보**를 만드는 연산이다. 본 논문에서는 설명의 편의상 두 요약정보 S_X와 S_Y가 합병된다는 가정 하에 논의를 전개한다.

$S_X = \{s_{X,1}, \ldots s_{X,m}\}, s_{X,i}$
$= (r_{X,i}, c_{X,i}), c_{X,i} \in W_X, 1 \le i \le m,$
$S_Y = \{s_{Y,1}, \ldots s_{Y,n}\}, s_{Y,j}$
$= (r_{Y,j}, c_{Y,j}), c_{Y,j} \in W_Y, 1 \le j \le n.$

요약요소의 합병은 요약범위와 요약값의 도메인이 같은 두 요약요소 $s_{X,i}$와 $s_{Y,j}$를 합병함으로써 통합 요약요소 $s_{XY,k} = (r_{XY,k}, c_{XY,k})$를 만드는 연산이다. **요약정보의 합병**은 $r_{X,i}$와 $r_{Y,j}$가 같은 $s_{X,i}$, $s_{Y,j}$는 요약요소 합병을 하여 통합 요약요소 $s_{XY,k}$를 통합 요약정보에 포함시키고, 동일한 요약범위를 갖는 대응되는 요약요소가 없는 경우, 합병 없이 통합 요약정보에 그대로 포함시킨다. 따라서 본 논문은 합병의 대상이 되는 두 요약정보 S_X와 S_Y의 **요약정보의 합병 조건**을 다음과 같이 제안한다.

1. 요약값에 대한 합병조건: S_X와 S_Y에 대한 요약값의 도메인 W_X, W_Y가 같다.

2. 요약범위에 대한 합병조건: 모든 i, j에 대해 요약범위 $r_{X,i} \in S_X$ 와 $r_{Y,j} \in S_Y$ 는 범위가 동일하거나, 서로소(disjoint)이다.

예제 1: 네 개의 요약정보 $S1$ = {([1:4], 10), ([5:6], 30)}, $S2$ = {([1:4], (20, 30)), ([5:6], (30,40))}, $S3$ = {([1:2], 30), ([3:4], 40)}, $S4$ = {([5:6], 30), ([7:8], 40)}가 있을 때에 $S1$은 요약값의 도메인이 1 차원 정수의 집합이고, $S2$는 요약값의 도메인이 2 차원 정수의 집합 이므로 요약값의 도메인이 다르므로 합병이 되지 않는다. 또한 $S1$의 요약범위 [1:2]와 S3의 요약범위 [1:4]가 동일하지도 않고, 서로소도 아니므로 합병이 되지 않는다. $S1$과 $S4$는 요약값과 요약범위의 합병 조건을 만족한다. 요약요소의 합병을 두 요약값의 덧셈으로 정의하면 통합 요약정보 $S14$ = {([1:4], 10), ([5:6], 60), ([7:8], 40)}를 얻을 수 있다.

같은 요약함수로 만들어진 요약정보들은 요약값에 대한 합병 조건을 만족하지만, 요약범위에 대한 합병 조건을 만족하지 않는 경우가 빈번하게 발생한다. 이 경우, 히스토그램의 균일 빈도(uniform frequency) 가정[13]을 이용하여 두 요약정보가 합병 조건을 만족하도록 변환된 후에 합병할 수 있다. 예를 들어, 예제 1의 요약정보 $S1$과 $S3$가 합병되려면 $S1$을 $S1'$ = {([1:2], 5), ([3:4], 5), ([5:6], 30)}으로 변환한 후, $S1'$와 $S3$를 합병하여 통합 요약정보 $S13'$ = {([1:2], 35), ([3:4], 45), ([5:6], 30)}을 얻는다. 이 방법은 균일 빈도 가정으로 인하여 통합 요약정보의 정확도가 떨어지는 것이 큰 단점이다. 본 논문에서는 이 문제점을 해결하기 위하여 웨이블릿 변환에 기반한 요약정보로 통합 요약정보를 관리하는 기법을 제안한다. 웨이블릿 요약정보는 합병 조건을 만족하도록 쉽게 변환이 되어 균일 빈도 가정 없이 간단하게 합병이 이루어진다.

3. 웨이블릿 요약정보

본 절에서는 웨이블릿 변환의 기본 개념을 설명하고, 웨이블릿 변환을 기반으로 하는 요약정보의 특성에 대해 설명한다.

3.1 웨이블릿 변환

웨이블릿 변환(*wavelet transform*)은 임의의 함수를 근사하는 함수인 **배율함수**(*scaling function*)와 근사로 생기는 오차를 보상하는 함수인 **웨이블릿**(*wavelet*)의 합으로 표현하는 기법으로 비주기적인 데이타에 대해 높은 데이타 압축 효과가 있어 컴퓨터 그래픽스와 신호처리 등에서 사용된다[12]. 웨이블릿의 정의에 의해 배율함수 $\varphi_{m,n}(t)$와 웨이블릿 $\phi_{m,n}(t)$는 다음과 같이 정의된다[16].

$$\varphi_{m,n}(t) = \sqrt{2^{-m}}\varphi(2^{-m}t - n),$$
$$\phi_{m,n}(t) = \sqrt{2^{-m}}\phi(2^{-m}t - n). \quad (1)$$

여기서, $\varphi(t)$는 $\varphi_{m,n}(t)$의 기저함수이고, $\phi(t)$는 $\phi_{m,n}(t)$의 기저함수이다.

위의 식에서 m은 기저함수의 주기를 변화시키는 계수로 **해상도**(*resolution*)라 하고, n은 기저함수의 오프셋(offset)으로 **인덱스**(*index*)라 부른다. 또한, $\sqrt{2^{-m}}$은 배율함수와 웨이블릿의 내적을 1로 만들기 위한 **정규화 계수**(*normalization factor*)이다.

배율함수와 웨이블릿을 사용하여 임의의 함수 $f(t)$는 웨이블릿 변환을 통해 다음 식으로 표현된다.

$$f(t) = \sum_{n=0}^{N-1} A[0,n]\varphi_{0,n}(t) \quad \text{해상도 } 0$$

$$= \sum_{n=0}^{\frac{N}{2}-1} A[1,n]\varphi_{1,n}(t) + \sum_{n=0}^{\frac{N}{2}-1} B[1,n]\phi_{1,n}(t) \quad \text{해상도 } 1$$

$$\vdots$$

$$= \sum_{n=0}^{\frac{N}{2^m}-1} A[m,n]\varphi_{m,n}(t) + \sum_{k=1}^{m} \sum_{n=0}^{\frac{N}{2^k}-1} B[k,n]\phi_{k,n}(t)$$

해상도 m

$$\text{where } A[m,n] = \int_{-\infty}^{\infty} \varphi_{m,n}(t) f(t)dt,\ B[k,n] = \int_{-\infty}^{\infty} \phi_{k,n}(t) f(t)dt. \quad (2)$$

여기서 $A[m,n]$은 **배율함수 계수**(*scaling function coefficient*)라 하고, $B[k,n]$은 **웨이블릿 계수**(*wavelet coefficient*)라 한다. 또한, 식 (2)에서 해상도를 증가 시키는 변환을 **분해**(*decomposition*)라 하고, 해상도를 감소시키는 변환을 **복원**(*reconstruction*)이라 한다.

웨이블릿 변환은 기저함수에 따라 Haar 웨이블릿, 선형(linear) 웨이블릿, Daubechies 웨이블릿 등으로 분류된다[12]. 이 중 본 연구에서 사용하는 Haar 웨이블릿은 단위함수 $u(t)$를 기저함수로 사용하는 웨이블릿으로 변환 연산과 압축이 간단하여 많이 사용된다.

웨이블릿 변환을 위하여 실수형 속성의 확장 누적 데이타분포 T는 함수 $T(t)$로 표현되고, 정수형 속성의 확장 누적 데이타분포 T는 단위함수 $u(t)$를 사용하여 함수 $T(t)$로 표현된다. 예를 들어, 데이타분포 T = {(1, 10), (2, 20), (3, 40)}는 함수 $T(t) = 10\ u(t) + 20\ u(t-1) + 40\ u(t-2)$로 표현된다.

예제 2: 정수형 속성의 데이타분포 {(1,20), (2,20), (3,70), (4,90),(5,100), (6,100), (7,120), (8,140)}을 Haar 웨이블릿 변환하면, 각 해상도에서 다음과 같은 계수들을 구할 수 있다.

해상도	배율함수 계수	웨이블릿 계수
0	1[20, 20, 70, 90, 100, 100, 120, 140]	
1	$\sqrt{2}$[20, 80, 100, 130]	$\sqrt{2}$[0, −10, 0, −10]
2	2[50, 115]	2[−30, −15]

해상도 0에서는 주어진 데이타분포의 빈도로 이루어진 시퀀스가 배율함수 계수가 되고, 해상도 1에서의 배율함수 계수와 웨이블릿 계수는 각각 연속된 두 시퀀스 값의 합의 절반과 차의 절반으로 표현된다. 위의 예에서 20와 20의 합의 절반인 20이 배율함수 계수로 계산되고, 차의 절반인 0이 웨이블릿 계수로 계산된다. 해상도 2에서는 20와 80에 대하여 50과 −30이 각각 배율함수 계수와 웨이블릿 계수로 계산된다. 각 해상도에 맨 앞에 나와있는 1, $\sqrt{2}$, 2는 식 (1)에 의한 정규화 계수이다. 결과적으로 주어진 시퀀스에 대한 해상도 2에서 배율함수 계수는 {100, 230}이고, 웨이블릿 계수는 {(−60, −30), (0, $-10\sqrt{2}$, 0, $-10\sqrt{2}$)}이다. 또한 연산을 반대로 적용하여 배율함수 계수와 웨이블릿 계수로부터 원래의 시퀀스를 복원할 수 있다. □

3.2 웨이블릿 요약정보의 생성

데이타분포를 웨이블릿 변환하여 얻어진 배율함수 계수와 웨이블릿 계수가 요약값인 요약정보를 **웨이블릿 요약정보**라 한다[17]. 웨이블릿 요약정보는 전체 데이타분포를 등간격(equi-width)으로 분할하여 같은 크기의 요약범위를 만든다. 요약범위의 크기는 해상도에 의해 결정된다. 즉, 웨이블릿 변환은 해상도 m에서 요약범위의 크기가 2^m인 등간격 요약정보를 생성한다. 웨이블릿 요약정보는 분해와 복원으로 요약범위의 크기를 조절할 수 있다. 분해는 요약범위의 크기를 2배 증가 시키고, 복원은 요약범위의 크기를 1/2로 감소 시킨다.

예제 2와 같이, 웨이블릿 변환으로 얻어진 배율함수 계수와 웨이블릿 계수의 총 수는 데이타분포의 길이와 동일하므로 압축 효과를 얻지 못한다. 따라서, 웨이블릿 요약정보는 웨이블릿 압축을 통해 저장해야 할 계수의 수를 줄인다. 압축과정은 다음과 같다[12]. (1) 계수들을 절대값의 역순으로 정렬하고, (2) 가용한 저장 공간에 의해 정해진 k값에 따라 k개의 계수를 선택하고, (3) 나머지 계수들은 값을 '0'으로 간주하여 버린다. 웨이블릿 압축으로 발생하는 오차의 분석은 제4.2절에서 다룬다.

예제 3: 예제 2의 데이타분포 T에 대한 웨이블릿 요약정보 S는 2개의 요약요소 s_1, s_2로 구성된다. 요약값을 배율함수 계수와 웨이블릿 계수의 순으로 표시하면 s_1 = ([1:4],(100, −60, 0, $-10\sqrt{2}$))이고, s_2 = ([5:8], (230, −30, 0, $-10\sqrt{2}$))가 된다. 이때 4개의 계수를 위한 저장 공간만 있다면 요약정보는 s_1 = ([1:4], (100, −60, 0, 0))와 s_2 = ([5:8], (230, −30, 0, 0))으로 압축된다. □

Matias 등은 웨이블릿에 기반한 히스토그램을 이용한

선택률 추정 기법을 제안하였다[17]. 본 논문에서는 참고문헌 [17]에서 논의된 개념을 확장하여 웨이블릿 요약정보를 기반으로 통합 요약정보 관리에 대하여 논의하고자 한다. 통합 요약정보의 표현 수단으로서 채택한 웨이블릿 요약정보의 주요 장점은 다음과 같다.

1. 웨이블릿 변환은 $O(n)$의 복잡도를 가지므로 계산 시간이 빠르다[16].
2. 웨이블릿 요약정보는 히스토그램에 비해 적은 양의 정보로도 데이타분포의 정확한 추정이 가능하다[17].
3. 웨이블릿 변환은 해상도의 조정이 가능하고, 이동성을 가지므로 웨이블릿 요약정보는 항상 합병 조건을 만족하도록 변환이 가능하다(제4.1절 참조).
4. 웨이블릿 변환의 선형성으로 웨이블릿 요약정보의 합병은 매우 단순하고 효율적으로 수행된다(제4.1절 참조).
5. 통합 웨이블릿 요약정보의 오차 한계가 증명 되어진다(제4.2절 참조).

4. 통합 웨이블릿 요약정보의 관리

본 절에서는 웨이블릿 요약정보를 기반으로 하는 통합 요약정보의 관리 기법에 관하여 논의한다. 먼저, 요약정보의 합병 알고리즘을 제안하고, 합병 결과로 생성된 통합 요약정보의 오차를 분석한다. 또한, 각 웨이블릿 요약정보의 개별적인 변화를 통합 요약정보에 반영하기 위한 점진적 갱신 방법을 제안한다.

4.1 요약정보의 합병

웨이블릿 요약정보들이 합병되려면 제2.2절에서 제시한 요약값과 요약범위에 대한 합병 조건을 만족해야 한다. 먼저, 요약값에 대한 합병 조건을 만족하려면 웨이블릿 요약정보들의 해상도가 같아야 한다. 데이타분포를 해상도 m으로 웨이블릿 변환하면 총 2^m개의 배율함수 계수와 웨이블릿 계수로 구성된 요약값이 생성된다. 압축은 계수의 값을 0으로 간주하여 저장하지 않는 것이므로 압축 정도와 계수의 수와는 무관하다. 따라서 같은 해상도 m의 두 웨이블릿 요약정보의 요약값은 2^m차원의 실수 도메인을 가지므로 요약값에 대한 합병 조건을 만족한다.

요약범위에 대한 합병 조건을 만족하려면, 웨이블릿 요약정보들의 해상도가 같고, 모든 요약범위의 최소값인 요약정보의 시작값이 같아야 한다. 요약정보들의 해상도가 같으면 요약범위의 크기가 같아지고, 시작값이 같으면 요약범위가 같거나 서로소가 된다. 만일 시작값이 다르다면 값집합의 범위를 이동하여 시작값을 같게 한다. 예를 들어, 두 웨이블릿 요약정보 S_X와 S_Y가 합병될 때 요약정보 S_X의 요약범위는 [1:4], [5:8]이고, 요약정보 S_Y의 요약범위는 [3:6], [7:10]이라면 요약범위의 크기는 같지만 시작값이 다르므로 요약범위에 대한 합병 조건을 만족하지 않는다. 따라서 요약정보 S_Y의 시작값을 이동하여 요약범위가 [1:4], [5:8], [9:12]인 요약정보 S'_Y로 변환하여야 한다. 즉, 값집합의 범위가 [3:10]인 요약정보 S_Y는 범위가 [1:12]인 요약정보 S'_Y로 조정된다. 이러한 값집합의 범위 조정을 위하여 웨이블릿의 이동성을 사용한다. **이동성**(*shift*)은 임의의 함수 $X'(t)$ (= $X(t - \tau)$)의 배율함수 계수와 웨이블릿 계수와의 관계를 나타내는 다음 식으로 정의된다[16].

$$A_X[m,n] = A_{X'}[m, n - \tau 2^m],$$
$$B_X[k,n] = B_{X'}[k, n - \tau 2^k], \quad \text{where } 1 \le k \le m. \tag{3}$$

위 식에서 A_X 와 B_X는 $X(t)$의 배율함수 계수와 웨이블릿 계수이고, $A_{X'}$와 $B_{X'}$는 $X'(t)$의 배율함수 계수와 웨이블릿 계수이다. 식 (3)에 의해 값집합의 범위 조정은 계수 값의 재계산 없이 인덱스의 수정만으로 쉽게 구할 수 있다.

합병 조건을 만족하도록 변환된 두 웨이블릿 요약정보들은 웨이블릿의 선형성을 이용하여 합병된다. **선형성**(*linearity*)은 같은 해상도 m에서 두 함수 $X(t)$, $Y(t)$의 배율함수 계수와 웨이블릿 계수의 합이 두 함수의 합인 $X(t)+Y(t)$의 배율함수 계수와 웨이블릿 계수와 같다는 성질로 다음 식으로 표현된다[16].

$$A_{X+Y}[m,n] = A_X[m,n] + A_Y[m,n],$$
$$B_{X+Y}[k,n] = B_X[k,n] + B_Y[k,n], \text{ where } 1 \le k \le m. \tag{4}$$

위 식에서 A_{X+Y}와 B_{X+Y}는 $X(t)+Y(t)$의 배율함수 계수와 웨이블릿 계수이고, A_X, A_Y와 B_X, B_Y는 $X(t)$, $Y(t)$의 배율함수 계수와 웨이블릿 계수이다. 선형성으로 인하여 웨이블릿 요약정보의 합병은 해상도와 인덱스가 같은 계수들을 더하는 연산으로 단순화된다. 이러한 특성을 이용한 합병 알고리즘은 그림 2와 같다.

이 알고리즘은 전처리 단계, 통합 요약정보 생성 단계, 그리고 후처리 단계로 구성된다. 전처리 단계에서 알고리즘의 스텝 1.1은 두 요약정보 S_X와 S_Y가 요약값에 대한 합병 조건을 만족하도록 동일한 해상도 m으로 분해 또는 복원한다. 스텝 1.2는 요약범위에 대한 합병 조건을 만족하도록 웨이블릿의 이동성을 이용하여 값집합의 시작값을 같게 만든다. 통합 요약정보 생성 단계는 웨이블릿의 선형성인 식 (4)를 이용하여 통합 요약정보를 생성한다. 후처리 단계는 생성된 통합 요약정보를 주어진 크기 l로 압축한다. 배율함수 계수와 웨이블릿 계

알고리즘 MergeSummaryData
입력: (1) 요약정보 S_X와 S_Y
(2) 통합 요약정보의 해상도 m
(3) 통합 요약정보의 크기 l
출력: 통합 요약정보 S_{XY}
알고리즘:
1 전처리단계: 합병조건을 만족하도록 다음 단계를 수행한다.
1.1 요약정보 S_X와 S_Y의 해상도가 m이 되도록 분해 또는 복원을 한다.
1.2 이동성(식 (3))을 사용하여 S_X와 S_Y의 값집합의 시각값을 같게 한다.
2 통합 요약정보 생성 단계: 선형성(식 (4))을 사용하여 통합 요약정보 S_{XY}를 생성한다.
3 후처리 단계: 통합 요약정보를 압축하기 위해 다음 단계를 수행한다.
3.1 S_{XY}의 요약값 A'_{XY}와 B'_{XY}을 절대값이 큰 순서로 정렬한다.
3.2 정렬된 요약값 중에서 상위 l개만 취하고 나머지는 버린다.

그림 2 웨이블릿 요약정보의 합병 알고리즘

수를 절대값 순으로 정렬하여 절대값이 큰 l개의 계수만 저장하고, 나머지 계수는 저장하지 않는다.

4.2 오차분석

웨이블릿 요약정보는 압축에 의한 정보의 손실로 오차를 갖는다. 본 절에서는 제안된 합병 기법으로 생성된 통합 요약정보의 오차는 항상 합병에 참여한 요약정보들의 오차를 합한 값보다 작거나 같음을 보인다.

먼저, 웨이블릿 요약정보의 오차를 구하기 위해 다음과 같은 표기법을 사용한다. 도메인이 D이고 길이가 N인 데이타분포 T를 웨이블릿 변환을 하여 얻은 배율함수 계수들과 웨이블릿 계수들을 A, B라 하고, A, B를 압축하여 얻은 배율함수 계수들과 웨이블릿 계수들을 A', B'라 하자. 그러면, T에 대한 웨이블릿 요약정보 S는 A'과 B'이다. 압축시 손실된 배율함수 계수 $\tilde{A}$는 $\tilde{A}[m,n] = A[m,n] - A'[m,n]$이며 압축시 손실된 웨이블릿 계수 $\tilde{B}$는 $\tilde{B}[m,n] = B[m,n] - B'[m,n]$이다. 요약정보 S로 복원한 추정 데이타분포를 T'이라 할 때에 오차 $\tilde{T}(t)$는 실제 데이타분포와 추정 데이타분포의 차가 된다. (t)는 식 (2)와 (4)에 의해 식 (5)로 구해진다.

$$\begin{aligned}\tilde{T}(t) &= T(t) - T'(t)\\ &= \sum_{n=0}^{\frac{N}{2^m}-1}(A[m,n]-A'[m,n])\varphi_{m,n}(t)\\ &\quad + \sum_{k=1}^{m}\sum_{n=0}^{\frac{N}{2^k}-1}(B[k,n]-B'[k,n])\phi_{k,n}(t)\\ &= \sum_{n=0}^{\frac{N}{2^m}-1}\tilde{A}[m,n]\varphi_{m,n}(t) + \sum_{k=1}^{m}\sum_{n=0}^{\frac{N}{2^k}-1}\tilde{B}[k,n]\phi_{k,n}(t).\end{aligned} \quad (5)$$

식 (5)에서 $\tilde{T}(t)$는 $\tilde{A}$와 $\tilde{B}$로 복원한 함수를 의미한다. 이때에 실수형 속성은 $\tilde{T}(t)$를 사용하고, 정수형 속성은 $\tilde{T}(t)$를 샘플링하여 얻은 시퀀스 $\{\tilde{T}[i]\}_{1\le i\le |D|}$를 사용하여 오차를 다음 식으로 정의한다.

오차	속성의 자료형	
	실수형	정수형
절대오차 E_1	$\int_D \lvert\tilde{T}(t)\rvert dt$	$\sum_{i=1}^{\lvert D\rvert}\lvert\tilde{T}[i]\rvert$
제곱근오차 E_2	$\sqrt{\int_D \tilde{T}(t)^2 dt}$	$\sqrt{\sum_{i=1}^{\lvert D\rvert}\tilde{T}[i]^2}$
최대오차 E_∞	$\max_{t\in D}\{\lvert\tilde{T}(t)\rvert\}$	$\max_{1\le i\le \lvert D\rvert}\{\lvert\tilde{T}[i]\rvert\}$

Haar 웨이블릿의 배율함수 $\varphi_{m,n}(t)$와 웨이블릿 $\phi_{m,n}(t)$는 수직정규(orthonormal) 함수를 구성하므로 다음 식이 성립한다[16].

$$\int_D \tilde{T}(t)^2 dt = \sum_{n=0}^{\frac{N}{2^m}-1}\tilde{A}[m,n]^2 + \sum_{k=1}^{m}\sum_{n=0}^{\frac{N}{2^k}-1}\tilde{B}[k,n]^2. \quad (6)$$

식 (6)은 제3.2절에서 설명한 절대값이 큰 순서로 m개의 계수를 선정하는 압축 방법은 웨이블릿 요약정보의 제곱근오차를 최소화함을 의미한다[12].

요약정보의 합병에서 합병에 참여하는 요약정보의 오차와 통합 요약정보의 오차와의 관계에는 정리 1이 성립한다. 즉, 합병된 요약정보를 후처리 과정에서 추가로 압축하지 않은 경우, 통합 요약정보의 오차는 각 요약정보의 오차 합보다 크지 않음을 보장하는 것이다.

정리 1: 속성 X의 웨이블릿 요약정보 S_X의 오차 E_1^X, E_2^X, E_∞^X 속성 Y의 웨이블릿 요약정보 S_Y의 오차 E_1^Y, E_2^Y, E_∞^Y가 주어질 때에 S_X와 S_Y를 합병하여 얻은 통합 웨이블릿 요약정보 S_{XY}의 오차 E_1^{XY}, E_2^{XY}, E_∞^{XY}에 대하여 다음 부등식이 성립한다.

$$E_1^{XY} \le E_1^X + E_1^Y, E_2^{XY} \le E_2^X + E_2^Y, \text{ and}$$
$$E_\infty^X \le E_\infty^X + E_\infty^Y. \quad (7)$$

증명 : 식 (2)에 의해 요약정보의 해상도가 증가하거나 감소하더라도 복원된 함수, 즉, 추정 데이타분포는 변화되지 않는다. 따라서 S_X와 S_Y는 같은 해상도로 변환되었다고 가정하고 증명한다. 속성 X, Y의 데이타분포를 T_X, T_Y라 하고, 통합 요약정보 S_{XY}를 복원하여 얻은 추정 데이타분포를 T'_{XY}라 하면, 통합 요약정보의 오

차 $_{XY}$는 식 (8)로 표현된다.

$$\begin{aligned}\tilde{T}_{XY}(t) &= T_X(t) + T_Y(t) - T'_{XY}(t) \\ &= \sum_{n=0}^{\frac{N}{2^m}-1} (A_X[m,n] + A_Y[m,n] - A'_{XY}[m,n])\phi_{m,n}(t) \\ &\quad + \sum_{k=1}^{m} \sum_{n=0}^{\frac{N}{2^k}-1} (B_X[k,n] + B_Y[k,n] - B'_{XY}[k,n])\psi_{k,n}(t) \\ &= \sum_{n=0}^{\frac{N}{2^m}-1} (\tilde{A}_X[m,n] + \tilde{A}_Y[m,n])\phi_{m,n}(t) \\ &\quad + \sum_{k=1}^{m} \sum_{n=0}^{\frac{N}{2^k}-1} (\tilde{B}_X[k,n] + \tilde{B}_Y[k,n])\psi_{k,n}(t) \\ &= \tilde{T}_X(t) + \tilde{T}_Y(t).\end{aligned} \tag{8}$$

따라서, 통합 요약정보의 절대오차 $E_1^{XY} = \int_D |\tilde{T}_X(t) + \tilde{T}_Y(t)| dt$ 이다.

여기서 $\int_D |\tilde{T}_X(t) + \tilde{T}_Y(t)| dt \le \int_D |\tilde{T}_X(t)| dt + \int_D |\tilde{T}_Y(t)| dt$ 의 성질을 사용하면 식 (9)가 유도된다.

$$E_1^{XY} \le E_1^X + E_1^Y. \tag{9}$$

또한, $\sqrt{\int_D |\tilde{T}_X(t) + \tilde{T}_Y(t)|^2 dt} \le \sqrt{\int_D |\tilde{T}_X(t)|^2 dt} + \sqrt{\int_D |\tilde{T}_Y(t)|^2 dt}$ 의 성질을 사용하면 식 (10)이 유도된다.

$$E_2^{XY} \le E_2^X + E_2^Y. \tag{10}$$

최대오차 E_∞^{XY}는 $\max_{t \in D}\{|\tilde{T}_X(t) + \tilde{T}_Y(t)|\}$ 이므로 여기서 $\max(|a+b|) \le \max(|a|) + \max(|b|)$ 의 성질을 이용하면 식 (11)을 얻는다.

$$\begin{aligned}E_\infty^{XY} &\le \max_{t \in D}\{|\tilde{T}_X(t)|\} + \max_{t \in D}\{|\tilde{T}_Y(t)|\} \\ &= E_\infty^X + E_\infty^Y.\end{aligned} \tag{11}$$

□

전체적인 오차의 추세를 표현하는 절대오차와 제곱근 오차와 달리 최대오차는 정확한 값의 범위를 한정하는 데 사용한다. 예를 들어, 최대 오차로 5가 주어지면 추정빈도 10에 대한 실제 빈도는 5보다 크고 15보다 작다는 것을 알 수 있다. 최대오차를 이용하여 요약정보의 오차로 인해 잘못된 질의 결과가 생기는 것을 방지할 수 있다[14]. 통합 웨이블릿 요약정보의 최대오차는 중개자가 요약정보를 합병하는 과정에서 합병에 참여한 요약정보들의 최대오차의 합을 저장함으로써 쉽게 통합 요약정보의 최대오차를 유지할 수 있다. 만일, 후처리 과정에서 통합 요약정보의 계수 중 $\hat{A}_{XY}$와 $\hat{B}_{XY}$가 압축되어 삭제되면 식 (11)은 식 (12)로 수정된다. 여기서 $\hat{T}_{XY}$는 $\hat{A}_{XY}$와 $\hat{B}_{XY}$로 복원한 데이타분포이다.

$$E_\infty^{XY} \le E_\infty^X + E_\infty^Y + \max_{t \in D}\{|\hat{T}_{XY}(t)|\}. \tag{12}$$

식 (11)은 전체 도메인 D에 대한 최대오차이므로 이를 이용하여 구하는 오차의 한계가 너무 커진다는 문제가 있다. 참고문헌 [15]에서는 이를 해결하기 위하여 요약범위의 최대오차를 개별적으로 저장하는 방법을 제안한 바 있다. 주어진 질의 조건의 범위에 따라 해당 요약범위의 최대오차를 사용하면 오차 범위를 줄일 수 있다. 예를 들어, 요약범위 [0:16]의 최대오차는 10이고, 요약범위 [17:32]의 최대 오차는 20이라면 전체 도메인에 대한 최대오차는 20 이지만 속성값 8에 대한 추정빈도의 최대오차는 10이 된다. 본 논문의 인터넷 top-N 질의 실험에서도 이 방식을 채택하여 각 요약범위마다 별도로 최대오차를 저장한다. 이 최대오차를 사용하여 통합 요약정보의 오차로 인한 잘못된 질의 결과가 생기는 것을 방지한다.

4.3 점진적 갱신

각 정보원의 데이타베이스 내에서 발생하는 데이타의 추가와 삭제는 데이타분포에 영향을 미치게 된다. 요약정보는 어떤 순간에 구성된 정적인 정보이므로 실제 데이타분포와 요약정보에 의하여 추정된 데이타분포 간의 오차는 시간이 경과함에 따라 커지게 된다. 따라서 각 정보원에서는 이러한 오차가 어느 한계를 넘지 않도록 요약정보를 재구성하는 방식을 사용한다[18].

같은 이유로 인터넷 상의 중개자는 정확한 데이타분포의 추정을 위해 정보원에서 주기적으로 발생하는 요약정보의 변경을 통합 요약정보에 반영할 수 있어야 한다. 특히, 인터넷 환경에서 각 정보원의 요약정보의 변경 시점은 일정하지 않으므로 하나의 요약정보가 변경될 때마다 통합 요약정보를 합병 연산을 통하여 매번 새롭게 구성하는 경우, 재구성에 필요한 비용이 매우 크다. 따라서 전체 요약정보들을 합병하지 않고 통합 요약정보에 변경된 요약정보만을 새롭게 반영하는 **점진적 갱신**(*incremental update*)을 필요로 한다.

점진적 갱신의 처리는 정보원의 요약정보 갱신과 중개자의 통합 요약정보 갱신으로 나누어진다. 정보원에서 요약정보의 갱신은 다음과 같다. (1) 요약정보의 갱신을 요청 받고, (2) 현재 요약정보와 이에 대한 최대오차를 임시 저장한다. (3) 요약정보를 갱신하고, 최대오차를 계산한다[2]. (4) 정보원은 요약정보와 최대오차의 변화를 중계자에게 전송한다. 즉, 저장해 놓은 이전 요약정보와

2) 최대오차를 관리하는 웨이블릿 요약정보의 점진적 갱신 방법은 제안되어 있지 않다. 따라서, 본 논문에서는 웨이블릿 요약정보가 재구성되는 것으로 가정하였다. 각 정보원의 요약정보를 점진적으로 갱신하는 문제는 본 논문의 범위가 아니므로 자세히 다루지 않는다.

갱신된 요약정보의 차와 최대오차의 차를 중계자에게 전송한다. 중개자는 전송 받은 요약정보의 변화와 최대오차의 변화를 받아 통합 요약정보를 갱신한다. 중개자의 점진적 갱신 알고리즘은 그림 2의 알고리즘 Merge SummaryData와 동일하다. 정보원의 데이타분포 변경으로 인한 요약정보 S_X의 변화를 S_X라 표현하면 알고리즘은 (1) 통합 요약정보 S_{XY}와 S_X (2) S_{XY}의 해상도, (3) S_{XY}의 크기를 입력으로 받아, 변경된 통합 요약정보 S'_{XY}를 생성한다.

S_X의 최대오차의 변화를 ΔE_∞^X 라 표현하면, 정리 1에 의해 S'_{XY}의 최대오차는 (1) 갱신 전 요약정보 S_X의 오차, (2) ΔE_∞^X, 그리고 (3) 갱신과정의 후처리에서 발생된 최대오차의 합으로 계산된다. 일반적으로 점진적 갱신은 시간이 지남에 따라 오차가 계속적으로 증가한다. 따라서, 특정 임계치 이상이 되는 시점에 전체 요약정보를 재구성하는 방법을 채택하여 오차가 너무 커지는 것을 방지하여야 한다[18].

5. 응 용

본 절에서는 선택률 추정과 인터넷 top-N 질의 처리를 응용 사례로 이용하여 통합 요약정보를 이용한 인터넷 질의의 효과적인 처리 방안에 관하여 간략히 논의한다.

5.1 인터넷 질의 최적화

선택률(*selectivity*)은 데이타베이스의 크기에 대한 질의 결과 크기의 비율로 정의되며, 질의 최적화(query optimization) 및 물리적 데이타베이스 설계(physical database design) 과정에서 중요한 요소로서 사용된다[7, 8]. 인터넷 질의 처리 비용에서 가장 우세한 비용은 전송 비용이므로 중개자는 최소의 레코드 전송이 필요한 질의 처리 계획을 선정한다[19][3]. 중개자는 통합 요약정보를 기반으로 추정한 선택률을 이용하여 각 질의 처리 계획의 비용을 계산한다.

통합 요약정보의 유용성을 설명하기 위해 112개의 정보원들에 분산된 인구조사 데이타베이스를 예로 사용한다. 인구조사 데이타베이스는 ID, INCOME의 속성으로 이루어진 CENSUS1 데이타베이스가 주별로 56개의 INCOME 정보원에 저장되어 있고, ID, AGE의 속성으로 이루어진 CENSUS2 데이타베이스가 주별로 56개의 AGE 정보원에 저장되어 있다. 전국에서 20세 미만이면서 수입이 10,000불 이상인 사람의 ID를 찾기 위하여 다음 질의를 사용한다고 하자.

SELECT CENSUS1.ID **FROM** CENSUS1, CENSUS2
WHERE CENSUS1.INCOME >=10000 **AND**
CENSUS2.AGE < 20 **AND** CENSUS1.ID=CENSUS2.ID

이 질의는 CENSUS1에 대한 질의(간단히, CENSUS1 질의라 부른다) **SELECT** CENSUS1.ID **FROM** CENSUS1 **WHERE** CENSUS1.INCOME >= 10000와 CENSUS2에 대한 질의(간단히, CENSUS2 질의라 부른다) **SELECT** CENSUS2.ID **FROM** CENSUS2 **WHERE** CENSUS2.AGE < 20가 죠인으로 결합된 것이다. 분산 데이타베이스에서 널리 사용하는 세미-죠인(semijoin)[20]을 이용한다고 가정하면, 위 질의를 중개자가 수행하는 방법으로 다음의 세 가지 방법을 고려할 수 있다.

방법 1 : 중개자가 CENSUS1 질의를 56개의 INCOME 정보원에게 전송하여 수행시키고, 질의 결과로 ID를 받는다. 중개자는 질의 결과로 얻는 ID와 CENSUS2 질의를 56개의 AGE 정보원에게 전송한다. AGE 정보원은 CENSUS2 질의를 수행하고, 결과로 얻어진 ID와 중개자가 보내 온 ID를 죠인하여 질의 결과를 중개자에게 반환한다.

방법 2 : CENSUS2 질의를 먼저 수행하는 방법으로서 방법 1에서 CENSUS1와 CENSUS2를 서로 대치하고, INCOME 정보원과 AGE 정보원을 대치한 방법이다.

방법 3 : 중개자가 CENSUS1 질의를 INCOME 정보원에 전송하고, CENSUS2 질의를 AGE 정보원에 전송한다. 중개자는 CENSUS1 질의 결과 ID와 CENSUS2 질의 결과 ID를 받는다. 중개자는 전달 받은 ID들을 죠인하여 질의 결과를 얻는다.

이 세가지 방법의 처리 비용은 CENSUS1 질의와 CENSUS2 질의의 선택률로 예측된다. 이 과정에서 INCOME과 AGE에 대한 통합 요약정보를 사용한다. 중개자는 통합 요약정보를 기반으로 추정한 선택률로 각 방법의 질의 처리 비용을 계산하고, 최소의 비용을 가지는 처리 방법을 선정할 수 있다. 이 결과 인터넷 환경에서 효과적인 질의 처리를 수행할 수 있다.

5.2 인터넷 top-N 질의

Top-N **질의**는 주어진 조건을 만족하는 레코드들을 주어진 순서로 정렬한 후 상위 N개의 레코드만 추출하는 질의이다[21]. 기존의 DBMS가 모든 질의 결과를 추출한 후에 결과를 정렬하여 결과를 얻는 것과 달리, top-N 질의를 지원하는 DBMS는 결과 레코드 수가 N이라는 사실을 반영하여 질의를 최적화하므로 질의 수

3) 이러한 가정은 SDD-1[19] 같은 분산 질의 최적화 기법에서 사용되고 있으며 광역 네트웍을 사용하는 인터넷 상에서 잘 적용된다.

행 비용이 크게 감소한다[11].

인터넷 *top-N* **질의**는 중개자가 인터넷에 분산되어 있는 여러 정보원에 대해 수행하는 top-N 질의이다. 여러 인터넷 쇼핑몰에서 주어진 조건을 만족하는 최저 가격의 물품 100개를 찾는 질의가 그 예이다. 중개자는 인터넷 top-N 질의를 통합 요약정보를 사용하여 실렉션 질의(selection query)로 변환하여 각 정보원에 전달하고, 각 정보원들은 주어진 실렉션 질의를 수행한다. 이 과정에서 전송될 레코드의 수를 감소시켜 인터넷 top-N 질의의 처리 성능을 크게 향상시킬 수 있다.

앞서 사용한 인구조사 데이타베이스를 이용하여 인터넷 top-N 질의 처리에서 통합 요약정보의 유용성을 보인다. 다음 질의는 INCOME이 큰 순서로 100명의 ID를 반환하는 인터넷 top-N 질의이다[21].

SELECT ID **FROM** CENSUS1 **ORDER** 100 BY INCOME

통합 요약정보가 없는 경우에는 중개자는 주어진 질의를 **SELECT** ID, INCOME **FROM** CENSUS1 **ORDER** 100 **BY** INCOME으로 변환하여 56개 정보원에 전달한다. 중개자는 질의 결과로 56개 정보원으로부터 100개의 레코드, 총 5600개의 레코드를 전송 받는다. 중개자는 질의 결과를 INCOME에 대해 정렬하여 상위 100개의 레코드를 웹 클라이언트에게 반환한다.

반면에, 통합 요약정보가 있는 경우에는 중개자는 통합 요약정보를 이용하여 인터넷 top-N 질의에 대응하는 실렉션 질의의 조건 범위를 구한다. 예를 들어, 통합 요약정보로 속성 INCOME의 값이 10,000 이상인 레코드가 전체 데이타베이스 내에 100개 이상 존재한다는 것이 확인되면 주어진 top-N 질의를 다음의 실렉션 질의로 변환하고, 각 56개 정보원에 전달한다.

SELECT ID, INCOME **FROM** CENSUS1 **WHERE** INCOME >= 10000

중개자는 각 정보원이 질의 결과로 보내온 레코드들을 받아 INCOME에 대해 정렬하여 100개의 레코드를 웹 클라이언트에게 반환한다. 이 방법은 통합 요약정보를 이용하지 않은 방법과 비교하여 전송되는 레코드의 수를 최대 1/56로 감소시킨다. 또한, 조건 INCOME >= 10000이 추가되므로 각 정보원 내에서의 질의도 보다 효과적으로 처리될 수 있다. 따라서, 통합 요약정보를 사용함으로써 top-N 질의 처리비용을 크게 줄일 수 있음을 알 수 있다.

6. 성능 평가

통합 요약정보의 유용성을 보이기 위해 1) 인터넷 top-N 질의 처리와 2) 선택률 추정 실험을 수행하였다. 본 절에서는 인터넷 top-N 질의 처리 실험을 통해 통합 요약정보가 인터넷 질의 최적화에 유용하게 사용될 수 있음을 보이고, 선택률 추정 실험을 통해 통합 웨이블릿 요약정보가 통합 히스토그램에 비해 더 정확한 통합 요약정보를 생성함을 보인다. 실험에서 각 데이타베이스는 인터넷 상에 분산된 정보원에 저장되어 있고, 각 정보원은 확장 누적 데이타분포에 대한 웨이블릿 요약정보를 관리한다고 가정한다. 또한, 중개자는 통합 요약정보를 생성하고, 통합 요약정보를 사용하여 사용자의 인터넷 질의를 처리한다.

6.1 실험 환경

실험에서는 합성된 데이타와 실제 데이타를 사용하였다. 합성된 데이타 *ZIPF-DATA*는 참고문헌 [13]에서 사용한 Zipf 분포에 의해 만들어진 데이타분포의 집합이다. 값 집합은 *uniform*(등 간격), *zipf_inc*(Zipf 분포로 간격이 증가), *zipf_dec*(Zipf 분포로 간격이 감소), *cusp_min*(Zip 분포로 간격이 감소 후 증가), *cusp_max*(Zip 분포로 간격이 증가 후 감소), *zipf_ran*(램덤 간격) 등 6가지 분포를 사용하였고, 빈도는 Zipf 계수를 0에서 4까지 1 단위로 증가시켜 5가지를 생성하였다. 값과 빈도의 상관관계로는 양의 상관(positive correlation), 음의 상관(negative correlation)과 영의 상관(zero correlation)의 3가지를 사용하였다[13]. 이 세가지 요소들의 모든 가능한 조합을 이용하여 총 90개의 데이타베이스를 생성하였다. 각 데이타베이스는 총 500,000개의 레코드로 구성되며, 데이타분포의 크기는 4 바이트 정수를 사용하면, 평균 약 16킬로 바이트이다. 실제 데이타 *CENSUS-DATA*는 미국의 인구조사 데이타[4] 중에서 고유번호(ID), 주(state), 소득(income)을 추출하고, 주별로 분리하여 총 56개 데이타베이스를 생성하였다. 각 데이타베이스 내의 평균 레코드 수는 98,600개이고, 데이타분포의 크기는 약 211킬로 바이트이다.

6.2 실험 결과

인터넷 top-*N* 질의

본 실험의 목적은 인터넷 top-N 질의 처리에서의 통합 요약정보의 유용성을 보이는 것이다. 실험은 통합 웨이블릿 요약정보를 사용한 경우와 사용하지 않은 경우에 대해 다음의 인터넷 top-N 질의의 처리비용을 비교하였다.

SELECT ID **FROM** CENSUS1 **ORDER** *N* **BY** INCOME

4) 이 데이타는 http://www.census.gov/DES/www/welcome.html에서 얻을 수 있다.

주어진 인터넷 top-N 질의의 처리는 제5.2절에서 설명한 방법으로 통합 요약정보를 사용하여 결과 크기가 N이 되는 INCOME의 조건 범위를 구한 후, 제4.2절에서 언급한 통합 요약정보의 최대오차를 보상하도록 조건 범위를 확장한다. 이 조건 범위로 각 정보원 내에서 실렉션 질의를 수행하고, 그 결과를 중개자에게 전송한다.

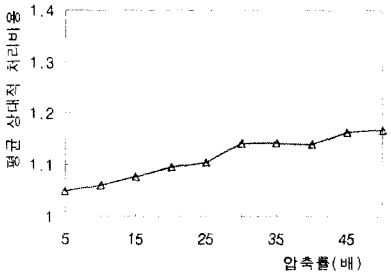


그림 3 압축률에 따른 인터넷 top-N 질의의 상대적 처리 비용

그림 3은 통합 요약정보의 압축률에 따르는 인터넷 top-N 질의의 상대적 처리 비용을 나타낸 것이다. 압축률은 평균 데이타분포의 크기에 대한 통합 요약정보의 크기로 압축률을 5부터 50까지 5씩 증가시키면서 통합 요약정보를 생성하였다. 실험은 각 통합 요약정보에 대해 N을 1,000부터 20,000까지 1,000씩 증가시키면서 인터넷 top-N 질의를 수행하고, 상대적 처리비용의 평균을 계산하였다. 여기서, 1,000은 전체 레코드 수의 0.018%이다. 상대적 처리비용은 정보원들이 중개자에게 전송한 총 레코드의 수를 N으로 나눈 것으로 하한값은 1이며 상한값은 정보원의 수이다. 따라서 CENSUS-DATA의 상대적 처리비용의 상한값은 56이다. 요약정보의 오차는 요약정보의 압축률이 높아질수록 증가하는 경향이 있다. 따라서, 요약정보의 압축률이 높아짐에 따라 질의의 조건 범위가 커지게 되고, 추출되는 레코드의 수가 많아진다. 그림 3에서 50배로 압축한 통합 요약정보를 사용한 경우, 1.16의 평균 상대적 처리비용을 나타내는데 이는 요약정보를 사용하지 않는 인터넷 top-N 질의 처리 방법과 비교하여 약 48배(=56/1.16)의 성능 개선 효과를 얻은 것을 의미한다. 이러한 실험 결과는 통합 요약정보가 인터넷 질의 처리비용을 크게 줄일 수 있음을 의미한다.

선택률 추정

본 실험의 목적은 통합 요약정보로 추정한 선택률의 정확도를 검증하는 것이다. 웨이블릿 요약정보와 비교 대상으로는 웨이블릿 요약정보와 히스토그램 기법들 중 참고문헌 [13]에서 구성 시간과 오차에서 가장 좋은 성능을 보인 MaxDiff(V,A) 히스토그램 기반 요약정보(이를 간단히 *MD* **요약정보**라 부른다)를 선정하였다. 실험에 사용한 질의는 $a < X \leq b$ 형태의 실렉션 질의를 사용하였다. 질의 집합은 값 집합의 범위에서 무작위로 a, b를 구하여 만든 총 1,000개의 질의들로 구성된다.

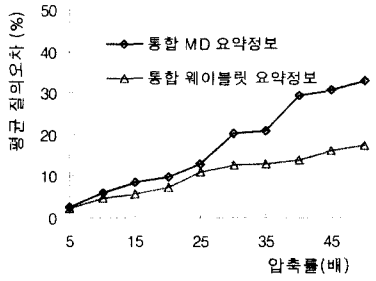
먼저, 각 요약정보들은 요약정보 합병으로 통합 요약정보를 생성한다. 통합 MD 요약정보를 구성을 위한 히스토그램의 합병은 아직 제안된 연구결과가 없으므로 제2.2절에서 설명한 균일 빈도 가정을 사용하여 추정한 데이타분포를 합병하여 통합 MD 요약정보를 구성하였다. 웨이블릿 요약정보와 MD 요약정보의 공정한 비교를 위해 동일한 크기의 저장 공간을 사용하였다. 웨이블릿 요약정보는 계수 당 인덱스와 계수값을 저장하기 위해 8 바이트를 사용하였다. MD 요약정보는 버킷 당 1) 범위의 끝값과 총 빈도수를 저장하기 위해 8 바이트를 사용하거나, 2) 시작값을 추가로 저장하여 12 바이트를 사용한다. 실험 결과에 의하면 동일한 저장 공간을 사용할 때에 방법 1) 보다 방법 2)가 더 정확한 통합 요약정보를 생성하므로 방법 2)로 생성한 통합 히스토그램에 대한 실험 결과만 보인다.

선택률 추정의 오차는 평균 추정오차를 사용하였다. **평균 추정오차**(*average estimation error*) J는 질의 집합 Q에 속하는 어떤 질의 q에 대한 실제 결과크기가 N_q이고, 추정 결과 크기가 N'_q일 때에 식(13)으로 구한다[13].

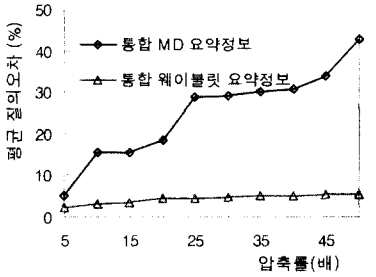
$$J = \frac{100}{|Q|} \times \sum_{q \in Q} \frac{|N_q - N'_q|}{N_q}. \quad (13)$$

그림 4는 요약정보 합병으로 생성된 통합 웨이블릿 요약정보와 통합 MD 요약정보를 사용하여 선택률 추정 실험을 한 결과이다. 그림 4에서 (a)는 ZIPF-DATA에 대한 실험 결과이고, (b)는 CENSUS-DATA에 대한 실험 결과이다. 그래프의 가로축은 통합 요약정보의 압축률이며 세로축은 평균 추정오차이다. 실험결과에 의하면 통합 웨이블릿 요약정보는 압축률이 증가함에 따라 오차는 미미하게 증가하고 오차의 크기도 크지 않다. 반면에 통합 MD 요약정보는 오차가 크며 압축률이 증가함에 따라 오차가 급격하게 커짐을 알 수 있다. 이러한 문제점은 MD 요약정보의 합병 과정에서 실제 상황에서는 거의 성립하지 않는 균일 빈도 가정을 사용하는 데서 기인한 것이다. 요약정보의 합병으로 얻은 통합 웨이블릿 요약정보의 평균 추정오차는 통합 MD 요약정보와

비교하여 ZIPF-DATA의 경우 1.6배, CENSUS-DATA의 경우 5.5배 작은 것으로 나타났다. 이 실험 결과는 웨이블릿 요약정보 합병이 효과적인 통합 요약정보의 생성 방법임을 의미한다.



(a) ZIPF-DATA



(b) CENSUS-DATA

그림 4 통합 요약정보의 선택률 추정과 오차

7. 결 론

본 논문은 다수의 정보원이 참여하는 인터넷 질의를 효과적으로 처리하기 위한 통합 요약정보의 생성, 관리, 그리고 이를 이용한 질의 처리 기법에 관하여 다루었다. 통합 요약정보의 구성을 위한 방법으로 데이타분포의 합병 방법은 큰 용량의 데이타분포를 전송, 저장, 통합하는 비용이 매우 크므로 실용적이지 않다. 본 논문에서는 이러한 문제점을 극복하기 위하여 웨이블릿 변환을 기반으로 요약정보들을 합병함으로써 통합 요약정보를 구성하는 새로운 방법을 제안하였다. 또한, 구성된 통합 요약정보를 이용한 인터넷 질의 최적화 방안을 제시하였다.

본 논문의 공헌은 다음과 같다.

1. 웨이블릿의 특성인 이동성과 선형성을 이용한 요약정보의 합병 방법을 제안하였다. 웨이블릿 요약정보는 요약정보들이 합병 조건을 만족하도록 쉽게 변환되며 합병 과정이 매우 단순하다.
2. 절대오차, 제곱오차, 최대오차에 대해 통합 요약정보의 오차는 항상 합병에 참여한 각 요약정보들의 오차 합을 넘지 않음을 증명하였다.
3. 통합 요약정보의 점진적 갱신은 요약정보의 합병을 이용하여 적은 비용으로도 수행될 수 있음을 보인다.
4. 제안한 통합 요약정보의 응용으로서 선택률 추정을 통한 인터넷 질의 최적화와 인터넷 top-N 질의 처리에 적용하여 질의 처리 비용이 크게 감소할 수 있음을 보였다.

본 논문은 선택률 추정과 인터넷 top-N 질의 처리 실험을 수행하였다. 선택률 추정 실험은 통합 웨이블릿 요약정보와 균일 빈도 가정을 사용하여 생성한 통합 히스토그램으로 추정 선택률의 정확도를 비교하였다. 그 결과 히스토그램의 합병에 비해 웨이블릿 요약정보의 합병이 1.6 ~ 5.5배 더 정확한 통합 요약정보를 생성한다는 것을 확인하였다. 또한 56개 정보원이 참여하는 인터넷 top-N 질의 실험에서 통합 요약정보를 사용하는 경우, 사용하지 않는 경우에 비하여 약 44배의 성능 개선 효과를 얻는 것으로 나타났다.

이와 같은 결과는 통합 요약정보가 다양한 인터넷 질의 처리에 매우 유용하게 활용될 수 있음을 보이는 것이며, 또한, 제안한 요약정보의 합병 기법이 인터넷 환경에서 통합 요약정보의 생성과 관리를 위한 효과적인 방법임을 나타내는 것이다.

참 고 문 헌

[1] Bernstein, P. et al., The Asilomar Report on Database Research, *SIGMOD Record*, Vol. 27, No. 4, pp. 74-80, 1998.

[2] Papakonstantinou, Y., Garcia-Molina, H., and Ullman, J., Medmaker: A Mediation System Based on Declarative Specifications, In *Proc. Int'l Conf. on Data Engineering(ICDE)*, pp. 132-141, 1996.

[3] Gravano, L., Garcia-Molina, H., and Tomasic, A., The Effectiveness of GlOSS for Text Database Discovery Problem, In *Proc. Int'l Conf. on Management of Data*, ACM SIGMOD, pp. 126-137, 1994.

[4] Papakonstantinou, Y., Abiteboul, S., and Garcia-Molina, H., Object Fusion in Mediator Systems, In *Proc. Int'l Conf. on Very Large Data Bases (VLDB)*, pp. 413-424, 1996.

[5] Florescu, D., Levy, A., and Mendelzon, A., Database Techniques for the World-Wide Web: A Survey, *SIGMOD Record*, Vol. 27, No. 3, pp. 59-74, 1998.

[6] Barbara, D. et al., The New Jersey Data Reduction Report, *IEEE Data Engineering Bulletin*, Vol. 20, No. 4, pp. 3-45, 1997.

[7] Selinger, P. et al., Access Path Selection in a Relational Database Management System, In *Proc. Int'l Conf. on Management of Data*, ACM SIGMOD, pp. 23-34, 1979.
[8] Whang, K., Wiederhold, G., and Sagalowicz, D., SeparabilityAn Approach to Physical Database Design, *IEEE Trans. on Computers*, Vol. c-33, No. 3, pp. 209-222, Mar. 1984.
[9] Vrbsky, S. and Liu, J., APPROXIMATE - A Query Processor that Produces Monotonically Improving Approximate Answers, *IEEE Trans. on Knowledge and Data Engineering*, Vol. 5, No. 6, pp. 1056-1068, Dec. 1993.
[10] Widom, J., Research Problems in Data Warehousing, In *Proc. Int'l Conf. on Information and Knowledge Management(CIKM)*, pp. 25-30, 1995.
[11] Chaudhuri, S. and Gravano, L., Evaluating Top-k Selection Queries,' In *Proc. Int'l Conf. on Very Large Data Bases(VLDB)*, pp. 397-410, 1999.
[12] Stollnitz, E., DeRose, T., and Salesin, D., *Wavelets for Computer Graphics: Theory and Applications*, Morgan Kaufmann, 1996.
[13] Poosala, V. et al., Improved Histograms for Selectivity Estimation of Range Predicates, In *Proc. Int'l Conf. on Management of Data*, ACM SIGMOD, pp. 294-305, 1996.
[14] Piatetsky-Shapiro, G. and Connell, C., Accurate Estimation of the Number of Tuples Satisfying a Condition, In *Proc. Int'l Conf. on Management of Data*, ACM SIGMOD, pp. 256-276, 1984.
[15] Jagadish, H. et al., Optimal Histograms with Quality Guarantees, In *Proc. Int'l Conf. on Very Large Data Bases(VLDB)*, pp. 275-286, 1998.
[16] Vetterli, M. and Kovacevic, J., *Wavelets and Subband Coding*, Prentice Hall, 1995.
[17] Matias, Y., Vitter, J., and Wang, M., Wavelet-Based Histograms for Selectivity Estimation, In *Proc. Int'l Conf. on Management of Data*, ACM SIGMOD, pp. 448-459, 1998.
[18] Whang, K., Kim, S., and Wiederhold, G., Dynamic Maintenance of Data Distribution for Selectivity Estimation, *VLDB Journal*, Vol. 3, No. 1, pp. 29-51, 1994.
[19] Bernstein, P. et al., Query Processing in a System for Distributed Databases(SDD-1), *ACM Trans. on Database Systems*, Vol. 6, No. 4, pp. 602-625, Dec. 1981.
[20] Ozsu, M. and Valduriez, P., *Principles of Distributed Database Systems*, Prentice Hall, 1999.
[21] Carey, M. and Kossmann, D., On Saying Enough Already! in SQL, In *Proc. Int'l Conf. on Management of Data*, ACM SIGMOD, pp. 219-230, 1997.

조 문 증
1989년 한양대학교 전기공학과 졸업(학사). 1991년 포항공과대학교 전자전기공학과 졸업(석사). 2001년 한국과학기술원 전자전산학과 전산학전공 졸업(박사). 1991년 1월 ~ 현재 LG전자기술원 정보기술연구소. 관심분야는 웹 데이타베이스, 분산 데이타베이스, ORDBMS

황 규 영
정보과학회논문지 : 데이타베이스
제 28 권 제 1 호 참조

김 상 욱
정보과학회논문지 : 데이타베이스
제 28 권 제 2 호 참조

심 규 석
1986년 서울대학교 전기공학과 졸업(B.S.). 1988년 University of Maryland, College Park (M.S.). 1993년 University of Maryland, College Park (Ph.D). 1994년 Federal Reserve Board, Research Staff. 1995년 ~ 1996년 IBM Almaden Research Center, Research Staff. 1996년 ~ 2000년 Bell Laboratories, Member of Technical Staff. 2001년 Micorosoft Research, Visiting Scientist. 1999년 ~ 현재 KAIST 전자전산학과 전산학전공 조교수. 2001년 ~ 현재 Program Committe Member: ICDE'02, VLDB'02, EDBT'02, SIGKDD'01, VLDB'00 SIGMOD'99 등 Editor: The VLDB Journal. 2000년 ~ 현재 Advisory Committee Member: ACM SIGKDD. 2001년 ~ 현재 한국 데이타마이닝학회 국제학술이사. 관심분야는 데이타마이닝, 데이타베이스, XML.