관계형 데이터베이스에서 비정규화를 사용한 데이터베이스 설계

장영관* · 강맹규*

A Database Design Using Denormalization in Relational Database

Jang, Young Kwan · Kang, Maing Kyu

Abstract

Databases are critical to business information systems, and RDBMS is most widely used for the database system. Normalization has been designed to control various anomalies(insert, update, and delete anomalies). However, normalized database design does not account for the tradeoffs necessary for the performance.

In this research, we develop a model for database design by denormalization of duplicating attributes in order to reduce frequent join processes. In this model, we consider insert, update, and delete costs. The anomalies are treated by additional disk I/O which is necessary for each insert and update transaction.

We propose a branch and bound method for this model, and show considerable cost reduction.

1. 서론

데이터는 여러 부문에서 중요한 전략적 자원이 되어 가고 있다. 이전에는 데이터를 주로 화일로 관리하여 왔으나 정보처리 요구가 증가함에 따라 DBMS 사용이 급격히 증가하고 있다. DBMS 중에서 현재 가장 많이 쓰이고 있는 것은 관계형 DBMS이다.

데이터베이스 설계는 일반적으로 (1) 요구의 수집 및 분석, (2) 논리적 설계, (3) 물리적 설계, (4) 구현 등의 절차로 이루어진다. 물리적 설계 단계에서는 논리적 설계 단계에서 설계된 관계(relation)를 그대로 테이블(table)로 사상(mapping)하여 구현하는 것이 이상적이다. 그러나 이렇게 설계한 결과는 좋은 수행도(performance)를 기대하기가 어렵다. 좋은 수행도를 얻기 위해서는 많은 경우 비정규화(denormalization)가 필요하지만 아직까지 비정규화를 수리적으로 모델링하여 최적화하는 연구는 크게 미진하다.

Roger[14]는 논리적 데이터베이스 설계를 그대로 물리적 데이터베이스 설계로 전환하여 적절한 수행도를 얻기는 어려우므로 경우에 따라서 비정규화는 필요하다고 강조하였다. Rajiv[13]는 물리적 데이터베이스 설계와 논리적 데이터베이스 설계를 수리모형화하여 열거법으로 해를 구하였다. 물리적 설계에서는 정렬과 인덱스(indexing)를 고려하고, 논리적 설계에서는 universal 릴레이션(relation)을 분해(decomposition)하는 것을 고려하였다.

Westland[17]는 데이터베이스 정규화에 관련된 저장비용, 갱신이변비용(anomaly cost), 조회비용을 확률모형으로 분석하였다. Lee[12]는 무조건 고차의 정규화 형태로 설계하는 것 대신에 비용과 수익을 고려해서 의사결정나무 방법으로 정규화형을 결정하는 방법을 제시하였다. Lee는 비용요소로서 저장비용(storage cost), 갱신이변비용, 그리고 결합비용(join cost)을 고려하였다.

본 연구에서는 저장용량에 제약이 있는 경우에 데이터베이스의 수행도를 향상시키기 위해 정규화된 데이터베이스 설계를 비정규화된 데이터베이스 설계로 조율(tuning)하는 방법을 수리적으로 모델링하고 분지한계법을 사용하여 최적해를 구한다.

2. 물리적 데이터베이스 설계

물리적 데이터베이스 설계는 좋은 수행도를 얻을 수 있도록 저장구조와 액세스(access)경로를 설정하는 것이다. 수행도는 보통 트랜잭션(transaction)을 처리하기 위해 필요한 액세스 횟수에 의해 결정된다. 예를 들어, 높은 차수로 정규화되어 있는 테이블은 보통의 경우에 많은 튜플(tuple)을 액세스해야 한다[10].

이러한 문제를 해결하는 방법에는 하드웨어 시스템 조율, DBMS 조율, 데이터베이스 설계 조율, 응용프로그램 조율 등이 있다[5]. 이 중에서 데이터베이스 설계 조율과 응용프로그램 조율이 가장 효

*한양대학교 산업공학과

과가 크다. 응용프로그램 조율은 관계형 데이터베이스의 질의언어(SQL)를 조율하는 방법으로서 데이터베이스의 물리적 설계 결과에 크게 영향을 받는다. 그러므로 좋은 수행도를 갖는 데이터베이스를 설계하기 위해서는 물리적 설계가 가장 중요한 것이라 할 수 있다[7].

데이터베이스의 물리적 설계 방법은 인덱스, clustering, hashing, 비정규화, 수직 단편화(fragmentaion), 수평 단편화 등이 있다. 이 중에서 인덱스, clustering, hashing의 방법은 데이터베이스의 테이블 설계를 변경하지 않고 할 수 있는 방법이고, 비정규화, 단편화는 테이블 설계를 변경하여 수행도를 향상시키는 방법이다.

비정규화의 일반적인 절차는 가장 중요하고 빈도가 높은 트렌젝션을 대상으로 하여 액세스 경로를 분석한다. 이러한 액세스 경로를 향상시키기 위해 결합 데이블을 정의하고 삽입, 갱신, 삭제, 조회, 저장 비용을 평가하고 갱신이변(anomaly)에 의한 부작용을 고려한다[16].

논리적 데이터베이스 설계의 결과인 정규화된 테이블 설계를 비정규화하여 설계하는 방법은 다음과 같다[14].

- 두 개의 릴레이션이 1:1의 관계에 있을 때: 두 개의 테이블을 1개의 테이블로 만든다.
- 두 개의 릴레이션이 M:N의 관계에 있을 때: 세 개의 테이블을 두 개의 테이블로 만든다.
- 두 개의 릴레이션이 1:N의 관계에 있을 때: E-R 모델에서 가장 일반적인 경우로서 참조되는 1의 관계의 속성을 N의 관계 테이블에 복사하여 결합(join)을 줄일 수 있다[9].
- 매우 자세한 데이터를 가지는 릴레이션: 부모(parent) 테이블에 자식(child)을 포함시킨다.
- 파생된(derived) 속성 : 한 개의 속성을 추가하여 반복적인 계산을 피한다.
- 보고서용으로 속성을 발췌하여 새로운 테이블을 만든다.

3. 연구 내용

2장에서의 비정규화 방법은 특수한 상황에서만 적용할 수 있고 일반적으로 정형화하기는 매우 힘들다 또한 두 개의 테이블을 한 개의 테이블로 만들 경우에 삽입, 갱신, 삭제 이변(anomaly)이 발생하여 부정확한 정보가 되거나 정보를 표현할 수 없는 경우가 발생한다. 본 연구에서는 더 일반적이고 범용성이 있는 모델을 만들어서 초기의 물리적 설계단계뿐만 아니라 데이터베이스 운용 중에 데이터베이스를 재설계할 경우에도 사용할 수 있도록 한다.

본 연구에서는 실제적으로 시스템을 구축할 때 널리 쓰이는 제 3 정규형으로 정규화된 테이블 설계가 주어져 있다고 가정한다. 두 개의 테이블을 i, j라고 하고 테이블 i는 테이블 j의 기본키(primary key)를 참조하는 참조 무결성 제약(referential integrity constraint)을 가지는 것을 i->j 라고 하면 테이블 i에 테이블 j의 속성들을 복사해 둠으로써 빈번한 결합을 방지하여 조회 트렌젝션의 수행도를 향상시키고 전체적인 수행도를 향상시킨다. 이 결과는 다음의 절충(trade-off) 관계가 있다.

- 조회 비용 감소
- 보조기억장치(디스크)의 사용량 증가
- 삽입, 갱신, 삭제 비용 증가

데이터베이스 설계와 구현의 과정에서 디스크 저장용량에 대한 투자 결정이 이루어지고 단기적으로는 디스크 저장용량은 제한되어 있으며, 디스크 저장용량을 늘리는 것은 장기적인 의사결정 사항이다[17]. 본 연구에서는 디스크 저장용량이 제한되어 있는 경우에 삽입 비용, 갱신 비용, 삭제 비용, 조회 비용을 수리 모형에서 고려한다.

본 연구의 비용 요소는 디스크의 액세스 횟수이다. 본 수리 모형은 삽입, 갱신이변(anomaly)에 기인한 데이터 일관성 유지 비용을 물리적인 액세스 방법에 따른 비용으로 수리 모형에 포함시켜 수식화한다.

본 연구에서의 액세스 방법은 순차 액세스(segment scan access)와 인덱스 액세스(indexed scan access)의 두 가지를 고려한다. 그 이외의 액세스 방법인 clustering, hashing 은 본 연구를 확장하여 쉽게 적용할 수 있다. 순차 액세스에서는 테이블의 모든 튜플을 액세스해야 하고 또한 blocking에 의해 한 번의 액세스로 여러 block을 액세스할 수 있기 때문에 $\frac{T_i \cdot R_i}{W}$ 로 표현되고, 또, 인덱스 액세스에서의 액세스 횟수는 질의를 처리하기 위해 액세스가 필요한 튜플의 개수로 추정할 수 있다[7]. 인덱스는 주로 B*-tree 혹은 B+-tree를 사용한다. 인덱스 액세스에서는 먼저. 인덱스를 인덱스 높이(height) 만큼을 액세스한 다음 실제 데이터가 있는 데이터 block을 한 번 액세스하기 때문에 $(1+H_{ii})N_{ii}$ 로 표현할 수 있다.

4. 수리 모형

4.1 가정

본 연구에서의 가정은 다음과 같다.

- 정규화된 설계가 주어져 있다.
- 기본키에 대한 인덱스가 존재한다.
- 인덱스 갱신 비용을 고려하지 않는다.
- CPU 시간, 주기억 장치의 한계는 고려하지 않는다.
- 결합, 정렬 비용은 고려하지 않는다.
- block에는 데이터만 저장되어 있고 정의된 길이보다 적은 속성도 정의된 길이만큼 고정길이로 저장된다.
- 모든 테이블의 튜플 크기는 1 block 보다 작다.
- 모든 갱신, 삭제 트렌젝션은 기본키를 액세스 경로로 하여 이루어지고 외부키에 의해 참조되는 테이블에 대한 삭제 트렌젝션은 없다.

4.2 기호 정의

본 연구에서 사용된 기호는 다음과 같다.

i, j, m : 테이블

k : 속성

R_i : 테이블 i의 튜플 개수

L_i : 정규화된 테이블 i의 튜플 크기

L_{ik}: 테이블 i의 속성 k의 크기

B : block 크기

W : blocking 계수×block 크기

H_i : 테이블 i에서 기본키 인덱스의 높이 (height)

H_{ti} : 트렌젝션 t에서 사용하는 테이블 i의 인덱스 높이

V_{ij} : 테이블 i의 한 개의 튜플에 대응되는 테이블 j의 평균 튜플 개수

F_t : tranaction t의 발생 빈도

z_{tji} 1 : 조회 트렌젝션 t에서 액세스하는 테이블 i의 모든 속성이 결합되는 테이블 j에 복사되어 있을 때

0 : 그렇지 않은 경우

y_{ij}, g_{ij} 1 : 테이블 i에 테이블 j의 속성이 적어도 한 개 복사되어 있을 때

0 : 그렇지 않은 경우

x_{ijk} 1 : 테이블 i에 테이블 j의 속성 k가 복사되어 있을 때

0 : 그렇지 않은 경우

U_{tij} 1 : 갱신 트렌젝션 t에서 j의 기본키를 참조하는 i의 외부키를 갱신할 때

0 : 그렇지 않은 경우

I_t : 삽입 트렌젝션 t에서 갱신하는 테이블 집합

U_t : 갱신 트렌젝션 t에서 갱신하는 테이블 집합

D_t : 삭제 트렌젝션 t에서 갱신하는 테이블 집합

P_t : 조회 트렌젝션 t에서 인덱스 액세스하는 테이블 집합

Q_t : 조회 트렌젝션 t에서 순차 액세스하는 테이블 집합

K_{ti} : 트렌젝션 t의 대상이 되는 테이블 i의 속성 집합

N_{ti} : 조회 트렌젝션 t에서 액세스하는 테이블 i의 튜플 개수

S : 가용한 여유 저정용량

4.3 수리 모형의 정식화

본 수리 모형은 삽입 비용(Ci), 갱신 비용(Cu), 삭제 비용(Cd), 조회 비용(Cq)의 합을 최소화하는 문제로서 다음과 같이 표현된다. 여기서 모든 비용은 각각의 트렌젝션에 대한 보조기억장치의 디스크 액세스 횟수를 나타낸다.

$$Min\ (Ci + Cu + Cd + Cq\ )$$

s.t.

$$\sum_t \sum_{j \neq i} \sum_k L_{jk} \cdot x_{ijk} \cdot R_i \leq S$$

$$\sum_k x_{ijk} - M \cdot y_{ij} \leq 0$$

$$\sum_{k \in K_{ti}} x_{jik} - M \cdot g_{ji} \leq 0$$

$$z_{tji} = \prod_{i \in K_{ti}} x_{jik}$$

$$z_{tji} + z_{tmj} \leq 1$$

본 수리 모형에서의 x_{ijk} 는 결정변수로서 테이블 i에 테이블 j의 속성 k를 복사하는 것을 의미한다. 만약, x_{ijk} = 1 이라고 결정되면 테이블 i에 새로운 속성을 복사한다는 것이고, 테이블 i에 새로운 속성이 한 개 추가된다. x_{ijk} 가 결정변수가 될 수 있는 조건은 다음의 세 가지가 동시에 만족할 경우이다.

- 테이블 i에는 테이블 j의 기본키를 참조하는 외부키 속성이 있어야 한다.
- 테이블 i와 테이블 j를 결합하여 조회하는 트렌젝션이 적어도 한 개 있어야 한다.
- 그 트렌젝션에서 테이블 j의 속성 k에 대한 조회 요구가 있어야 한다.

그림 1은 비정규화된 테이블 설계의 예로서 order 테이블의 customer_id는 customer 테이블의 customer_id를 참조하는 외부키이고, order 테이블의 customer_name은 customer 테이블의 customer_name을 복사한 것이다.

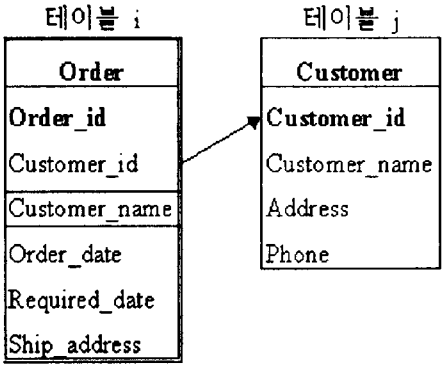

그림 1. 비정규화된 테이블 설계의 예

4.3.1 삽입 비용(Ci)

삽입 비용은 테이블 i에 새로운 튜플을 삽입할 때 발생하는 액세스 횟수와 복사된 속성의 값을 유지시키기 위한 테이블 j에 대한 액세스 횟수로서 다음과 같은 식으로 표현된다.

$$\sum_t F_t \sum_{i \in I_t} [1 + \sum_{j \neq i} y_{ij}(1 + H_j)]$$

여기서 테이블 j는 기본키에 의해 1개의 튜플을 액세스 한다. 테이블 j에 대한 액세스는 기본키에 대한 높이(H_j) 만큼 인덱스를 액세스하고 또, 데이터 블록을 한 번 액세스한다. yij는 추가된 결정변수로서 테이블 i에 테이블 j의 속성이 하나라도 복사하면 1이 되는 변수이다.

예를 들어, 그림 1에서 order에 새로운 튜플이 삽입되는 경우에 데이터 일관성을 유지하기 위해 customer_name을 customer 테이블에서 기본키에 의해 조회하는 비용이 추가된다.

4.3.2 갱신 비용(Cu)

갱신 비용은 복사한 테이블의 외부키가 인덱스 되어 있을 경우에 다음과 같이 표현된다. 갱신은 먼저 기본키에 대한 인덱스의 높이 만큼 액세스하고, 데이터 블록을 1 회 조회하고 1 회 갱신하는 액세스를 해야 한다. 또한 복사해 둔 속성이 있는 경우에 데이터 일관성을 유지하기 위한 액세스 비용이 다음과 같이 추가된다.

$$\sum_t F_t \sum_{i \in U_t} [(2+H_i+\sum_{j \neq i} y_{ij} \cdot U_{tij}\ (1+H_j + \sum_{j \neq i} V_{ij} \cdot g_{ji}\ (2+H_j)]$$

예를 들어, 그림 1에서 order의 customer_id가 갱신되는 경우에 데이터 일관성을 유지하기 위해 customer 테이블에서 1개의 튜플을 조회해야 한다. 또한 customer 테이블의 customer_name이 갱신되는 경우에 order에 복사한 customer_name을 갱신해야 한다. 복사한 테이블의 외부키가 인덱스 되어 있지 않을 경우는 다음과 같이 표현된다.

$$\sum_t F_t \sum_{i \in U_t} [(2+H_i)+\sum_{j \neq i} y_{ij} \cdot U_{tij}\ (1+H_j) + \sum_{j \neq i} g_{ji} \cdot (\frac{T_j \cdot R_j}{W} + V_{ij})]$$

여기서, $T_j = L_j + \sum_{m \neq j} \sum_k L_{mk} x_{jmk}$ 로서 테이블 j의 튜플 크기를 나타낸다.

4.3.3 삭제 비용(Cd)

삭제 비용은 다음과 같은 액세스 횟수로 표현되며, 본 연구에서는 속성을 복사해 두는 비정규화 방법을 사용하기 때문에 삭제 이변이 발생하지 않는다.

$$\sum_t F_t \sum_{i \in D_t} (2+H_i)$$

4.3.4 조회 비용(Cq)

조회 비용은 다음과 같이 순차 액세스의 경우와 인덱스 액세스 의 경우의 두 가지로 다음과 같이 표현된다.

$$\sum_t F_t\ [\sum_{i \in Q_t} (1-z_{tji})\ \frac{T_i \cdot R_i}{W} + \sum_{i \in P_t} (1-z_{tji}) \cdot (1+H_{ti}) N_{ti}]$$

여기서 z_{tji}는 추가된 결정 변수로서 결합 조회에서 필요한 테이블 i의 속성이 모두 테이블 j 에 복사되어 있으면 1이 되어 테이블 i를 액세스하지 않아도 결합 조회의 결과를 얻을 수 있다.

4.3.5 제약 조건

제약 조건은 다음과 같이 표현된다.

$$\sum_i \sum_{j \neq i} \sum_k L_{jk} \cdot x_{ijk} \cdot R_i \leq S$$

위의 제약식은 저장용량에 관한 제약으로서, 속성을 복사함으로써 증가되는 저장용량은 가용한 여유저장용량보다 적도록 한다.

$$\sum_k x_{ijk} - M \cdot y_{ij} \leq 0$$

$$\sum_{k \in K_a} x_{jik} - M \cdot g_{ji} \leq 0$$

여기서 M은 상수로서 크기가 큰 수(big M)를 나타내며, 추가된 결정변수 y_{ij} 및 g_{ij}에 대한 제약조건이다. 만약, x_{ijk}가 하나라도 1이면 y_{ij} 및 g_{ij}는 1이 되도록 제약한다.

$z_{tji} = 0$ $i, j \in Q_t \cup P_t$, t는 단일 테이블 조회 트렌젝션

$z_{tji} = \prod_{i \in K_a} x_{jik}$ $i, j \in Q_t \cup P_t$, j는 i를 참조하는 외부키 테이블, t는 결합 조회 트렌젝션

이 제약조건은 j->i 의 관계에 있는 테이블의 결합조회 트렌젝션을 처리하기 위한 제약으로서 만약, 결합조회에서 필요한 테이블 i의 모든 속성이 j에 복사되어 있으면 z_{tji}는 1로 제약되어 그 조회를 처리하기 위한 테이블 i의 액세스는 0이 될 수 있도록 한다.

$$z_{tji} + z_{tmj} \geq 1$$

이 제약조건은 세 개의 테이블 이상을 결합하는 조회 트렌젝션에 대한 제약이다. 만약, 조회 트렌젝션에서 필요한 테이블 i의 속성들이 테이블 j에 복사되어 있고 또, 테이블 j의 속성들이 테이블 m에 복사되어 있다고 가정하면, 이 트렌젝션을 처리하기 위해서는 테이블 i와 테이블 j 중에서 적어도 한 개는 액세스해야만 하고 또, 테이블 j와 테이블 m 중에 적어도 한 개는 액세스해야만 한다.

본 연구에서의 결정변수 x_{ijk}, y_{ij}, g_{ij}, z_{tji} 는 모두 1 또는 0 의 값을 갖는 이진 정수이다.

5. 분지한계법을 이용한 알고리듬

5.1 분지한계법의 기본 개념

분지한계법은 기본적으로 열거법(enumeration)의 일종으로서 그 중에서 부분열거법에 속한다[1]. 분지한계법은 각 단계에서 최적해가 포함되어 있을 가능성이 가장 커보이는 부분집합을 선정하여, 이 부분집합에 계산량을 줄일 수 있는 어떤 전략을 사용하여 두 개 이상의 새로운 부분집합으로 분할한다. 이를 분지(branch)라 한다.

각 분지된 문제에서 그 부분집합에서의 해의 한계를 계산한다. 만약 해의 한계를 구하는 중에 최적해를 구했으면 알고리듬을 끝낸다. 그렇지 않

으면 후보문제들의 한계를 사용하여 새로 분지할 부분집합을 선정하고 위의 과정을 반복한다.

5.2 본 연구의 분지한계법

본 연구에서의 분지는 기본적으로 복사할 대상이 되는 속성을 복사한다(x_{ijk} = 1)와 복사하지 않는다(x_{ijk} = 0)로 하면 된다. 본 연구에서는 x_{ijk} = 1, x_{ijk} = 0로 분지하지 않고 x_{ijk} = 1, x_{ijk} = 0 중에서 한 개로만 분지(부분 분지)할 수 있는 경우가 있다.

그림 2는 결합조회 트렌젝션 t^*에서 테이블 1과 테이블 2를 결합하는 경우 테이블 1에 테이블 2의 속성 3과 4를 복사하는 것을 분지하는 예이다. 본 연구의 수리모형은 속성을 복사하여 결합조회 트렌젝션에서 액세스하는 테이블의 수를 줄여서 총 비용을 최소화하는 것이므로, 그 수를 줄일려면 결합조회 트렌젝션 t^*에서 필요한 모든 속성을 복사하여야 한다. 그런데 $t \neq t^*$인 결합조회 트렌젝션 t가 테이블 1과 테이블 2의 속성 4를 결합하지 않는다면 후보노드와 다른 값을 갖는 분지는 같은 값을 갖는 분지보다 항상 비용이 크게 된다. 왜냐하면, 만약 속성을 일부만 복사하면 삽입, 갱신 비용은 속성을 복사하지 않는 경우보다 증가하게 되고, 결합조회 비용은 감소하지 않기 때문이다.

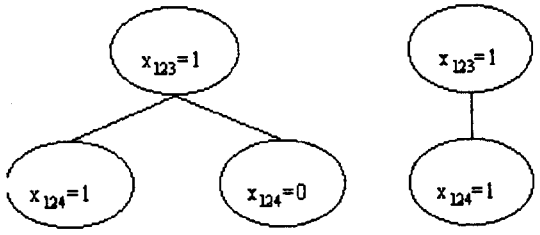

그림 2. 분지 전략

본 연구에서 개발한 분지한계법의 절차는 다음과 같다.

단계 0 : [초기화]
후보노드를 비어 있음으로 표지하고 단계 2로 간다.

단계 1 : [후보노드 결정]
아직 분지하지 않은 후보노드 중에서 최저한계비용이 가장 적은 노드를 찾는다. 만약 그 노드가 현재까지의 최적해 노드의 최저한계비용보다 크거나 같으면 단계 1을 반복한다. 만약 분지하지 않은 후보노드가 없으면 끝낸다.

단계 2: [분지 노드 결정]
단계 1에서 후보노드를 x_{ijk} 라고 하면 x_{ijk}를 사용하는 결합 조회 트렌젝션 중에서 아직까지 분지하지 않은 속성(k_c)을 분지노드로 한다. 만약 그 k_c가 부분 분지 기준에 부합되면 분지는 후보노드 x_{ijk}의 결정변수 값과 동일하게 분지한다. 그러한 k_c가 없으면 결합을 사용하는 조회 트렌젝션 중에서 분지하지 않은 새로운 테이블 i와 테이블 j를 찾아서 테이블 j의 임의의 속성 k를 분지노드로 한다.

단계 3 : [분지노드의 최저한계 계산]
4장의 수식을 사용하여 아직 분지하지 않은 노드의 결정변수는 조회 트렌젝션에서는 1로 하여 계산하고 삽입, 갱신, 삭제 트렌젝션은 0으로 하여 계산하여 가능성이 있는 최저한계를 계산한다.

단계 4 : [가능해 여부 결정]
분지가 잎새노드(leaf node)까지 되었고 가용한 저장용량보다 적으면 가능해로, 아니면 불능해로 표지한다.

단계 5 : [후보노드 집합에 분지한 노드 삽입]
분지한 노드를 후보노드에 삽입하고 단계 1로 간다.

6. 수치 예제

본 수치 예제는 그림 3과 표 1과 같이 정규화된 테이블 설계가 주어져 있을 때 비정규화하여 최적의 설계를 구하는 문제이다. 표 1에서 화살표의 시작점은 외부키를 표시한 것이고, 화살표의 끝점은 외부키가 참조하는 기본키를 표시한 것이며, 또 진한 글씨로 되어 있는 것은 기본키를 표시한 것이다. 표 2는 각 테이블에 대한 삽입 트렌젝션, 표 3은 갱신 트렌젝션, 표 4는 삭제 트렌젝션, 그리고 표 5는 조회 트렌젝션이다. block 크기는 1024, blocking 계수는 4로 하고, V_{13}=1, V_{24}=5, V_{34}=5로 주어진 것으로 가정한다. 또, 가용한 여유 저장용량은 1,500 Kbyte로 가정한다. 표 1에서 속성이름이 진하게 되어 있는 것은 외부키를 표시한 것이다. 본 예제에서는 모든 외부키는 인덱스되어 있다고 가정한다.

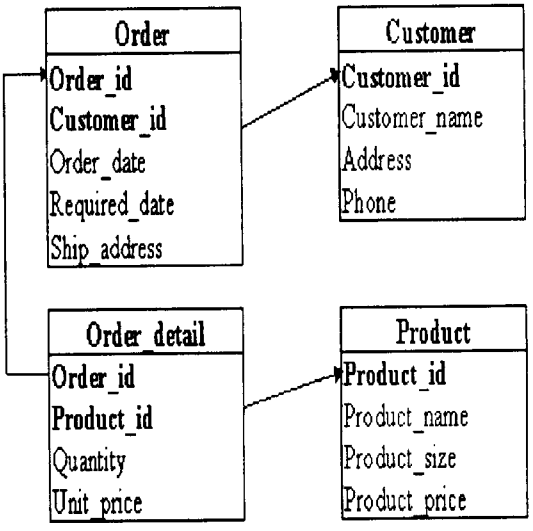

그림 3. 정규화된 테이블 설계의 예

표 1. 예제 테이블

테이블 번호	테이블 이름	튜플 개수	기본키 높이	속성 번호	속성 이름	길이
1	customer	5,000	2	1	customer_id	5
				2	customer_name	20
				3	address	100
				4	phone	15
2	product	5,000	2	1	product_id	5
				2	product_name	30
				3	product_size	5
				4	standard_price	10
3	order	5,000	2	1	order_id	5
				2	**customer_id**	5
				3	order_date	7
				4	required_date	7
				5	ship_address	100
4	order_detail	25,000	3	1	**order_id**	5
				2	**product_id**	5
				3	quantity	5
				4	unit_price	10

표 2. 예제 삽입 트랜잭션

트랜잭션 번호	테이블 번호	발생 횟수
1	1	1
2	2	1
3	3	10
4	4	50

표 3. 예제 갱신 트랜잭션

트랜잭션 번호	테이블 번호	속성 번호	발생 횟수
1	2	3	1
2	3	4	1
3	4	3,4	2

표 4. 예제 삭제 트랜잭션

트랜잭션 번호	테이블 번호	발생 횟수
1	4	2

표 5. 예제 조회 트랜잭션

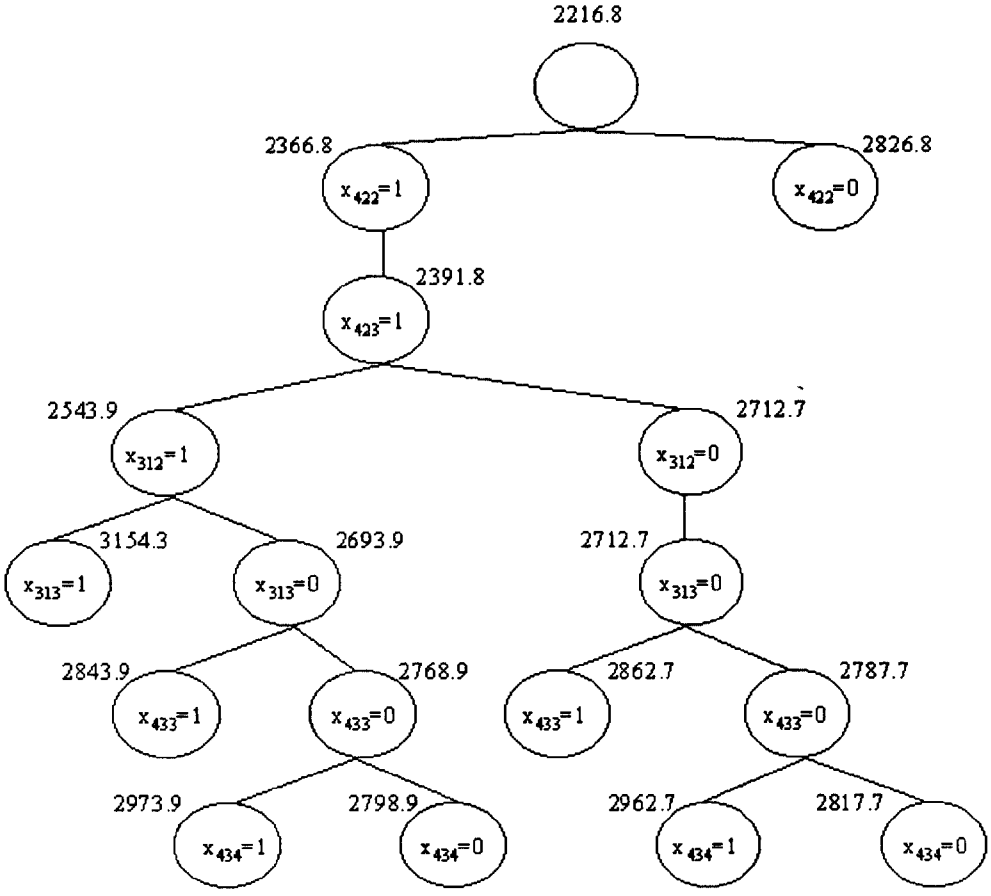

그림 4. 분지한계법을 이용한 해법

트랜잭션 번호	테이블 번호	속성 번호	액세스 방법	인덱스 높이	액세스 튜플 개수	발생 횟수
1	2	3	인덱스	2	1	100
2	3	3	인덱스	2	1	50
	1	2,3	인덱스	2	1	
3	4	3	인덱스	3	10	10
	2	2,3	순차	-	5,000	
4	4	3	인덱스	3	1	5
	3	3	인덱스	2	5	
5	1	2,3,4	인덱스	2	1	100
6	3	1,2,3,4,5	순차	-	5,000	5
7	4	2,3	인덱스	3	50	1
	3	4	인덱스	2	10	
	1	2	순차	-	5,000	

본 수치 예제에서 주어진 정규화 설계에서의 총 액세스 횟수는 3252.7 이고 본 연구에서의 비정규화한 설계의 최적해는 그림 4에서와 같이 2798.9 이고 총 액세스 횟수는 453.8 이 감소하였다. 최적해는 order_detail에 product의 product_name, product_size, 또 order에 customer의 customer_name을 복사해 두는 비정규화된 테이블 설계이다.

7. 결론

본 연구에서는 가용한 저장용량에 제한이 있는 경우에 데이터베이스의 수행도를 향상시키기 위한 방법 중에서 정규화된 데이터베이스 설계를 비정규화된 데이터베이스 설계로 조율하는 방법을 수리적으로 모델링하고 분지한계법을 사용하여 최적해를 구하였다.

본 연구의 비정규화는 조회뿐만 아니라 삽입, 갱신, 삭제 비용을 고려하였고 데이터베이스의 필수 조건인 데이터 일관성을 유지하는 데 드는 비용도 고려하였다.

데이터베이스 설계는 빈번하게 하는 것이 아니고 그 설계 주기가 매우 크므로 최적해를 구하는 시간이 다소 많이 걸리더라도 최적해를 구하는 것이 바람직하다.

본 연구의 수리모형에서의 최악의 complexity는 결정변수의 수를 n이라고 하면 $O(2^n)$으로서 지수형이지만 4장에서의 조건에 맞는 결정변수의 수는 많지 않으며, 또한 중요한 트렌젝션에 대해서만 데이터를 수집하여 분석하게 되므로 본 연구의 최적해법은 매우 유용한 해법이라고 할 수 있다.

참 고 문 헌

1. 강맹규, *네트워크와 알고리듬*, 박영사, 서울, 1991.
2. 장영관, 강맹규, "관계형 데이터베이스에서 저장용량에 제약이 없는 경우 비정규화를 고려한 데이터베이스 설계," *공업경영학회지*, 제 19 권, 37 집, 1996.
3. Armstrong, E., S. Bobrowski, J. Frazzini, B. Linden, and M. Pratt, *ORACLE7 Server Application Developer's Guide*, Oracle Corporation, Redwood, CA, 1992.
4. Batini, C., S. Ceri, and S. B. Navathe, *Conceptual Database Design*, The Benjamin/Cummings Publishing Company, Inc., Redwood, CA, 1994.
5. Bobrowski, S., E. Armstrong, C. Closkey, and B. Linden, *ORACLE7 Server Administors Guide*, Oracle Corporation, Redwood, CA, 1992.
6. Bobrowski, S., E. Armstrong, C. Closkey, B. Linden, and M. Pratt, *ORACLE7 Server Concepts Manual*, Oracle Corporation, Redwood, CA, 1992.
7. Chu, W. W. and Ion T. I, "A Transaction-Based Approach to Vertical Partitioning for Relational Database Systems," *IEEE Transactions on Software Engineering*, Vol. 19, No. 8, pp. 804-812, 1993.
8. Corrigan, P. and M. Gurry, *ORACLE Performance Tuning*, O'Reilly & Associates, Inc., Sebastopol, CA, 1993.
9. Elmasri, R. and S. B. Navathe, *Fundamentals of Database System*, The Benjamin/Cummings Publishing Company, Inc., Redwood, CA, 1994.
10. Hawryszkiewycz, I. T., *Relational Database Design*, Prentice Hall, Englewood Cliffs, NJ, 1990.
11. Hoffer, J. A., "An Integer Programming Formulation of Computer Data Base Design Problems," *Information Sciences*, Vol. 11, pp. 29-48, 1976.
12. Lee, H., "Justifying Database Normalization: A Cost/Benefit Model, *Information Processing & Management*," Vol. 31, No. 1, pp. 59-67, 1995.
13. Rajiv, M. D. and B. Gavish, "Models for the Combined Logical and Physical Design of Databases," *IEEE Transactions on Computers*, Vol. 38, No. 7, pp. 955-967, 1989.
14. Rodgers, U., "Denormalization: Why, What, and How?," *Database Programming & Design*, Vol. 2, No. 12, pp. 46-53, 1989.
15. Rumbaugh, J., M. Blaha, W. Premerlani, F. Eddy, and W. Lorensen, *Object-Oriented Modeling and Design*, Prentice-Hall International Inc., Cliffs, NJ, 1991.
16. Teorey, T. J., *Database Modeling & Design: The Fundamental Principles*, Morgan Kaufmann Publishers, Inc., San Francisco, CA, 1994.
17. Westland, J. C. "Economic Incentives for Database Normalization," *Information Processing & Management*, Vol. 28, No. 5, pp. 647-662, 1992.