논문 2005-42CI-3-2

데이터베이스 성능향상용 역정규화의 무용성

(Harmfulness of Denormalization Adopted for Database for Database Performance Enhancement)

이 혜 경*

(Rhee Hae Kyung)

요 약

정규화(normalization)는 데이터의 불필요한 중복의 정도를 최소화할 뿐만 아니라 데이터의 무결성을 높이는데 기여하기 때문에 데이터베이스를 효율적으로 설계하기 위해 수행하고 있다. 그러나 정규화를 깊숙이 수행한 데이터베이스인 경우 자료 검색 시 필요에 따라 테이블 간의 조인을 해야 하기 때문에 자료 처리 속도의 저하 현상이 발생될 수도 있다. 이러한 정규화의 부작용을 다소나마 해소하기 위한 수단으로 기업에서는 궁여지책으로 역정규화를 함으로써 어느 정도 완화시킬 수 있다고 보는 견해가 있다. 본 논문에서는 정규화와 역정규화와의 성능 평가를 위해 고객관련업무 시스템에 대해 두 가지 방법을 적용하여 데이터베이스 시스템을 구축하고 분석하여 비교하였다. 실험 결과 데이터베이스 크기에 따른 응답 시간은 전체적으로 역정규화 모델이 정규화 모델보다 더 길게 나왔다. 역정규화가 데이터의 중복을 발생시키기 때문에 시스템 성능 향상에 기여하는 바가 거의 없는 것으로 나타났다.

Abstract

For designing the database more efficiently, normailzation can be enforced to minimize the degree of unnecessary data redundancy and contribute to enhance data integrity. However, deep normalization tends to provoke multiple way of schema join, which could then induces response time degradation. To mitigate this sort of side effect that the normalization could brought, a number of field studies we observed adopted the idea of denormalization. To measure whether denormalization contributes to response time improvement, we in this paper developed two different data models about customer service system, one with perfect normalization and the other with denormalization, and evaluated their query response time behaviors. Performance results show that normalization case consistently outperforms denormalization case in terms of response time. This study show that the idea of denormalization, quite rarely contributes to that sort of improvement due ironically to the unnecessary data redundancy.

Keywords: Data Modeling, Normalization, Denormalization

I. 개 요

대다수의 기업에서 데이터베이스 시스템 구축을 해 놓은 것을 분석해 본 결과 개체나 관계가 아닌 속성들이 놀랍게도 많게는 65퍼센트 이상 중복되어 있다는 사실을 접할 수 있었다[1]. 만일 기업에서 관리하는 데이터의 양이 방대할 경우, 불필요하게 중복된 데이터를 처리하는 비용 또한 무시할 수 없는 수준일 것이다. 외래키로 인한 중복인 경우는 각 테이블 간의 조인을 위한 것이기 때문에 예외로 간주한다. 그러나 이와 같은 종류의 중복은 전체 데이터의 아주 작은 부분에 불과하다. 데이터 중복의 커다란 부분을 차지하고 있는 것은 데이터의 질에는 아무런 도움도 되지 않는 불필요한 중복이다[2].

불필요한 중복을 줄이기 위해 데이터베이스를 효율적이고 경제적으로 설계하는 일이 중요하다. 정규화(normalization)는 데이터의 불필요한 중복의 정도를 최소화할 뿐만 아니라 데이터의 무결성을 높이는데 기여한다. 데이터베이스의 분석 및 설계에 있어서 속성의 정규화는 보다 효율적인 데이터 관리를 위해 불필요한 속성 중복을 제거한다는 의미에서 중요하다. 속성 간의

* 정회원, 용인송담대학 컴퓨터게임정보과
(Yong-In Songdam College Dept. of Computer Game Information)
접수일자; 2004년11월4일, 수정완료일: 2005년5월2일

종속성을 분석하기 위한 방안으로 정규화 과정이 사용되는데, 정규화는 데이터베이스의 물리적 구조나 물리적 처리에 영향을 주는 것이 아니라 논리적 처리 및 품질에 큰 영향을 미친다.

정규화 과정을 수행하는 목적이 데이터의 중복을 최소화해야 할 뿐만 아니라 다른 한편으로는 질의응답 시간을 최소화하려는 이중적 목표를 추구해야 하는 관계로 설계 과정상의 복잡성이 불가피하게 유발된다. 정규화를 수행한 데이터베이스인 경우 자료 검색 시 조인을 추가적으로 해야 하기 때문에 자료 처리 속도의 저하 현상이 발생될 수 있다. 정규화 과정이 제1형에서 제5형으로 진행됨에 따라 테이블이 분해되어 수가 증가하고 조인이 증가함으로써 검색에 따른 추가 디스크 입출력이 발생함으로써 성능이 저하되는 부작용을 초래한다.

비록 정규화를 함으로써 데이터의 일관성을 향상시키는 효과는 볼 수 있지만, 일부에서는 질의 응답시간 측면에서 성능을 저하시킬 수도 있다는 것을 제기하고 있다[3]. 즉, 데이터베이스에서 좀더 정규화 된 구조가 테이블 수를 증가시키기 때문에 데이터 조작을 위해서는 보다 많은 조인을 수행해야만 한다는 것이다. 왜냐하면 조인은 응답시간을 지연시킬 수 있는 가장 중요한 요소 중 하나이기 때문이다. 보다 정규화 된 테이블은 응답시간 측면에서 볼 때 성능이 저하시킬 수도 있다. 이러한 이유로 비록 역정규화가 정규화 이론에는 반기를 들고 있지만, 성능을 향상시킬 수 있다[4].

그러나 역정규화 과정은 데이터의 중복을 허용하여 불필요하고 원하지 않는 데이터도 함께 삽입해야 하기 때문에 추가적인 데이터를 위한 저장 공간의 낭비를 초래할 수 있다는 함정이 도사리고 있다. 또한 데이터의 일관성을 해치고 동일한 데이터가 시스템 내의 여러 분산된 테이블에 중복 저장됨으로써 중복된 정보가 정확성을 유지할 수 있도록 하기 위해 갱신 작업이 부가적으로 요구된다. 또한 데이터베이스의 응용 프로그램과의 독립성을 위반하게 되며 데이터베이스를 공유하는 다른 응용 프로그램의 성능도 저하되며 시스템 유지 부담이 증가함으로써 추가적인 부담을 안게 된다. 역정규화의 목적은 디스크의 입출력 시간을 줄이는 효과를 얻을 수 있기 때문에 실제 업무에서 흔히 적용이 되지만 이러한 역정규화 작업이 설계자의 감각에 의존하여 아무런 기준 없이 이루어진다면 데이터베이스 시스템의 성능 또한 향상되리라 기대할 수가 없는 것이다. 따라서 시스템을 설계할 때 역정규화를 해서 데이터의 중복을 허용할 것인가, 아니면 데이터의 중복을 피할 것인가를 결정하기 위해 실제적이고 객관적인 비교실험을 통해 성능 평가를 할 필요가 있다.

예 1은 정규화 된 상태에서라면 두개의 테이블의 조인을 통해서만 얻을 수 있는 데이터를 역정규화를 통해 두개의 테이블을 하나의 테이블로 병합한 것이다. 이러한 방법은 데이터의 조회 시 조인 연산을 감소시켜 시스템의 성능향상은 도모할 수 있으나, 데이터의 삭제, 갱신이 발생했을 때 데이터의 변경 이상 현상이 발생할 수 있고 새로운 레코드 삽입 시 불필요한 자료 조회로 인한 불필요한 테이블의 추가적 참조 부담이 발생한다.

예 1(역정규화): 간단한 응용으로써, 다음과 같은 업무상의 규칙이 있다고 가정한다.

규칙1: "고객은 하나 이상의 계좌번호를 갖는다. 계좌번호는 오직 하나의 고객번호에만 소속된다."

규칙2: "고객은 담당 지점을 갖는다."

이 업무는 3개의 정규화된 테이블로 구성된다(그림 1).

그림 1에서와 같이 데이터베이스가 설계된 상태에서 " 고객명이 XXX인 사람의 계좌번호들을 조회하라"는 질의가 빈번히 발생된다고 가정을 하자. 이와 같은 질의를 실행하기 위해서는 고객과 계좌 테이블의 조인이 필요하다. 그림 2 에서는 역정규화에 의하여 고객과 계좌 테이블을 병합하여 계좌_고객 테이블을 만들었다.

그림 2에서의 질의응답 시간은 동일한 질의에 대해서 조인없이 오직 하나의 테이블에서만 질의를 수행할 수 있기 때문에 그림 1보다 짧아지리라 예상된다. 그러나 새로운 데이터의 삽입이나 기존 데이터의 갱신 작업을 할 경우 데이터의 일관성이 무너질 수 있기 때문에 데이터의 무결성을 유지하기 힘들다. 끝(예1).

고객

고객번호	고객명	전화번호	지점번호
10000	김은아	111-1111	L-A
10001	박찬숙	222-2222	L-B
10002	이명재	333-3333	L-C
10003	김태미	444-4444	L-D

지점

지점번호	지점장	지점위치	직원수
L-A	김철민	강남	20
L-B	박수동	강북	15
L-C	이남식	강동	16
L-D	김선경	강서	12
L-E	이정현	서초	18

계좌

계좌번호	개설지점	개설날짜	고객번호
000001	L-C	2004-12-24	10000
000002	L-C	2004-12-27	10000
000003	L-A	2004-12-27	10001
000004	L-E	2004-12-28	10002
000005	L-D	2004-12-29	10003

그림 1. 정규화 된 구조
Fig. 1. Normalized structure.

계좌_고객

계좌번호	계설지점	계설날짜	고객번호	고객명	전화번호	지점번호
000001	L-C	2004-12-24	10000	김은아	111-1111	L-A
000002	L-C	2004-12-27	10000	박찬숙	222-2222	L-B
000003	L-A	2004-12-27	10001	이명재	333-3333	L-C
000004	L-E	2004-12-28	10002	김태미	444-4444	L-D
000005	L-D	2004-12-29	10003	안남일	555-5555	L-E

지점

지점번호	지점장	지점위치	직원수
L-A	김철민	강남	20
L-B	박수동	강북	15
L-C	이남식	강동	16
L-D	김선경	강서	12
L-E	이정현	서초	18

그림 2. 두개의 테이블을 병합한 역정규화된 구조
Fig. 2. Normalized structure by composing two tables.

본 논문에서는 정규화와 역정규화와의 성능 평가를 위해 서비스 업체의 고객관련업무 시스템에 대해 설명하고 이를 바탕으로 두 가지 방법을 적용하여 데이터베이스 시스템을 구축하고 분석하여 비교한다. 고객정보처리시스템은 서비스 업체에서 가장 중심이 되는 업무로서 데이터의 정확성, 응답시간이 가장 중요한 요소이므로 이를 성능 분석의 표준으로 결정하였다. 업무분석을 통해 도출된 정규화된 테이블과 이를 바탕으로 역정규화를 실행한 테이블 들 간의 시스템 거래 발생 시 나타나는 성능 평가를 응답시간 측면에서 비교하고자 한다. 이 성능 평가를 통해 역정규화를 통한 시스템 구축이 기업의 정보시스템의 성능에 기여하고 있는지 여부를 파악하고자 한다.

본 논문의 구성은 다음과 같다. 제 II장 관련 연구에서는 정규화 과정과 역정규화 과정의 간략한 개념소개와 그 부작용에 대해 소개하며, 제 III장에서는 성능 평가를 위한 실제 표준 업무를 설정하여 정규화 테이블을 구성하고, 또한 성능 비교를 위해 역정규화된 테이블 목록을 작성한다. 작성된 테이블을 기반으로 실제 데이터베이스를 구성하고 그 결과를 분석, 비교한다. 제 IV장에서는 본 연구의 결론과 한계점을 다루었다.

II. 관련 연구

1. 정규화

정규화란 자료의 손실이나 불필요한 정보의 삽입 없이 데이터의 일관성을 유지하며 최소한의 데이터 중복을 허용하여 최대의 데이터 안정성 확보를 위한 안정적 자료구조로 변환하는 기법이다. 정규화 과정은 일반적으로 3차 정규화나 BCNF형까지 진행하는 경우가 대부분이며 정규화 과정을 거치는 동안 데이터 모형은 종속성 및 활용성이 높아진다. 정규화 과정이 제1차에서 제5차로 진행됨에 따라 테이블이 분해 되어 수가 증가하고 조인이 증가함으로써 검색에 따른 추가 디스크 입출력이 발생함으로써 성능이 저하되는 부작용을 초래할 수 있다.

고급 정규형으로 제 4 정규형, 제 5 정규형은 데이터의 정확성, 무결성, 일치성은 확고히 지킬 수 있으나 테이블 수의 증가로 인한 추가적 조인이 많이 발생하기 때문에 디스크 입출력으로 인한 시스템 성능 저하와 같은 정규화의 부작용을 크게 증가시킨다.

2. 역정규화

정규화를 통해 데이터의 정확성, 무결성, 일치성을 유지하는 면에서 유리하기는 하나 테이블 수의 증가와 추가 조인으로 인한 디스크 입출력 발생을 유발한다. 데이터의 정확성과 비중복성은 일부 저하되더라도 역정규화를 통해 테이블의의 수와 디스크 입출력 횟수를 줄여 시스템의 성능을 높이자는 주장들[5][6]이 나오고 있다. 테이블의 수와 추가 디스크 입출력 횟수를 줄이기 위해 사용되는 역정규화 방법에는 중복칼럼의 추가, 테이블 분할, 항목 추가를 통한 방법이 주로 사용된다.

세부 테이블의 컬럼을 중복시키는 방법은 가장 일반적으로 널리 사용되는 방법이다. 해당 테이블과 함께 빈번히 요구되는 컬럼 인 경우에 추가로 중복하여 사용한다. 그러나 이러한 역정규화는 복사된 컬럼과 원래의 컬럼 간의 정보가 항상 일치되어야만 하기 때문에 과도한 갱신 부담을 요구한다.

테이블 분할을 통한 역정규화를 하는 방법에는 행의 크기가 큰 경우에 사용하는 수직 분할과 행의 개수가 많은 경우 사용하는 수평분할로 구분된다. 수직 분할을 함으로써 저장 공간 면에서나 검색 속도 면에서 성능이 우수하다고 할 수는 있으나 주키(Primary Key)의 유일성 유지관리가 어렵고, 주키의 중복으로 인한 디스크 낭비를 가져올 수 있다. 또한 분할되어 있는 데이터를 동시에 요구할 경우에는 다중 테이블을 조인해야 하므로 SQL문이 복잡해지고 추가 조인으로 인한 부담을 초래한다. 실제 기업의 정보 시스템에서는 업무의 특성상 특정 테이블의 행의 수가 많아질 수 있다. 이 때 사용하는 역정규화 방법이 테이블 수평 분할인데, 수평분할의 경우 테이블 검색 범위 축소로 인한 액세스 량 감소, Backup 및 Recovery와 같은 테이블 관리가 용이

인터넷 서비스 업체의 영업 정보 시스템은 가입된 고객에게 인터넷 접속 및 사용을 위한 모든 서비스를 제공하는 업무를 주로 하고 있다. 고객에 관련된 업무와 인터넷 서비스 제공에 관련된 업무로 분류되며, 고객 관련 업무는 고객 유치를 위한 캠페인에서부터 고객을 위한 콜 센터 업무 등을 포함하고, 인터넷 서비스 제공과 관련된 업무는 출장에서부터 장비에 관련된 업무 등을 포함한다.

그림 3. 인터넷 서비스 영업 정보 시스템의 표준 업무 내용

Fig. 3. Standard description of Internet Service Business Information Systems.

한 이점이 있다. 그러나 주키의 유일성 유지 관리가 어렵고, 주키의 중복으로 인한 디스크 낭비를 가져올 수 있다.

수평분할은 행의 수가 지나치게 많은 테이블인 경우 테이블을 수평적으로 분할함으로써 역정규화를 수행한다. 일반적으로 시스템에는 몇 년 동안의 자료가 들어 있지만 검색하는 것은 그날그날 필요한 최근 자료뿐이다. 예를 들어 주문 정보 시스템인 경우 외부 주문에 대한 자료는 계속적으로 하나의 테이블에 보존되지만 일상적인 보고에 필요한 내용은 해당 년도의 자료에 대해 이루어지는 것이 일반적이다. 그 이전의 자료는 데이터베이스의 일상적 작업에는 부정적인 영향을 미친다고 볼 수 있다. 현재의 자료를 검색하고자 할 때 그 이전의 자료까지 색인의 일부로 읽어야 하기 때문이다. 그러나 주키의 분산으로 인한 유일성 유지관리가 어렵고, 또한 주키의 중복으로 인한 디스크 낭비를 초래할 수 있다. 만일 분할되어 있는 데이터를 동시에 요구할 경우 다중 테이블을 조인해야 하므로 SQL문이 복잡해지고 추가 조인으로 인한 부담이 발생한다.

자주 요청되는 데이터를 정규화 된 테이블에 추가시킴으로써 시스템의 성능을 향상시키기도 한다. 정규화된 상태에서라면 두개의 테이블의 조인을 통해서만 얻을 수 있는 데이터를 파생데이터를 삽입한 테이블 하나로 얻을 수 있는 방법이다. 이러한 방법은 질의를 빠르게 실행시킬 수 있다는 장점을 가지고 있으나, 파생되고 반복되는 데이터를 제대로 갱신하지 않으면 데이터 무결성을 손상시킬 수 있는 위험이 있다.

역정규화 방법에는 위에서 살펴본 방법 이외에 키 칼럼의 추가, 테이블 병합과 같은 방법 등이 있으나 모두 데이터 검색 시 시스템 성능 향상에 초점을 둔다는 공통점이 있다. 역정규화의 장점을 살리기 위해서는 반드시 정규화 과정을 끝낸 후에 수행해야 하나 특정한 기준이 없이 오직 경험만으로 정규화 과정을 끝내기도 전에 이루어지고 있는 실정이다.

Ⅲ. 표준 업무 설정 및 성능 평가

기업에서 가장 활용도가 높은 정보 시스템은 온라인 거래 등이 빈번하여 데이터의 정확성 유지와 응답 시간이 중요한 요소인 영업 업무 시스템이다. 보편적으로, 역정규화의 필요성을 비교하기에 충분하고, 특히 각 개체들 간의 행위 관계가 명확하여 단위 업무에 대해서 비교 분석이 가능하다. 따라서 본 논문에서는 정규화와 역정규화의 효용성을 비교 분석하기 위해 시스템 설계 및 구축의 대상이 되는 업무로 서비스 영업 시스템을 설정하였다. 영업 정보 시스템의 업무는 그림 3과 같은 체계를 가지고 있다.

데이터베이스를 구축하기 위해 적용하는 정규화 과정은 데이터베이스 설계 시 데이터가 불필요하게 중복되는 것을 막고, 질의에 대한 응답시간을 줄이기 위한 수단이다. 그러나 응답시간을 줄이기 위해 데이터의 중복을 허용하는 역정규화를 수행하는 경우가 있다. 역정규화는 시스템 성능 측면에서도 질의의 종류나 상황에 따라 반드시 우수하지 않을 수 있기 때문에 보다 객관적인 성능 판단을 위해 정규화 된 테이블과 역정규화 된 테이블의 비교 실험이 필요하다.

1. 실험 모델 및 가정

모의실험을 위해 본 논문에서 사용된 질의 처리 모형은 DB, 운영체계, DBMS, 사용자 단말기와 같은 구성요소로 이루어진다. 사용자 단말기는 4대로 구성되며 각 단말기들은 거래를 생성하여 각 연산 단위의 질의를 DBMS로 전송한다. DBMS는 사용자의 질의를 받아 운영체계의 파일 관리자에게 질의에 대한 검색을 요청하고, 파일 관리자는 DBMS가 원하는 레코드가 어느 페이지에 저장되어 있는지 조사해서 디스크 관리자에게 그 페이지의 검색을 요청한다. 디스크 관리자는 파일 관리자가 원하는 페이지의 실제 저장 위치를 알아내어 검색에 필요한 디스크 입출력 명령을 내림으로써 데이터베이스에 저장된 데이터가 검색된다. 사용자는 다시 디스크 관리자, 파일 관리자를 거쳐서 DBMS를 통해 검색 결과를 통보 받는다. 본 실험에서는 사용자가 질의를 작성하여 전송하고 DBMS로부터 결과를 받기까

지의 응답시간을 주요 변수로 측정한다. 본 실험에 사용된 서버 시스템으로는 TG 삼보(Pentium 2.4GHz)를 사용하였고, 운영체계는 MS Windows 2000 Server, DBMS는 MS SQL Server 2000 Enterprise Edition을 사용하였다.

성능 평가의 신뢰성을 높이기 위해 인덱스 공간의 배제, 질의의 보편성, 질의 영향의 균일성 등 몇 가지 가정이 필요하다.

가정 1(인덱스 공간의 배제): 실험에 사용되는 데이터베이스 저장 공간의 크기는 인덱스가 차지하는 공간을 배제한 순수 데이터 영역만으로 구성되어 있다고 가정한다. 원래 데이터베이스에서 테이블의 크기는 인덱스가 저장되는 부분과 실제 데이터가 차지하는 부분으로 구성되어 있으나, 인덱스가 차지하는 비중은 미미하기 때문에 인덱스 공간을 배제하는 것이 모의실험에서 각 모델간의 비교에 있어서 본 논문에서 연구하고자 하는 성능 평가 목적에 부합된다.

가정 2(질의의 보편성): 모의실험에 도입되는 질의는 특정 시간에 편중되어 있지 않고 동일한 시간 간격으로 수행됨을 가정한다. 또한 질의의 발생 횟수의 빈도에 따른 시스템 변화를 보다 효과적으로 측정하기 위해 동일한 시간 간격으로 발생됨을 가정한다.

가정 3(질의 영향의 균일성): 모의실험에 사용되는 질의들은 시스템에 같은 크기의 영향을 미친다고 가정한다. 임의의 질의들은 질의의 종류가 동일한 경우라도 접근 시간이나 대기 지연시간, 검색 데이터의 이질성으로 인하여 연산 수행 시간이 다를 수 있다. 그러나 보다 질의 비교의 명확성을 강조하기 위해서 임의의 질의에 대해서는 질의 처리를 위한 시스템 영향이 균일한 것으로 가정한다.

2. 시스템 성능 평가 방법

시스템의 성능은 시스템을 이용하는 환경과 업무의 특성에 따라 그 결과가 매우 다르게 나올 수 있기 때문에 실제 기업 환경과 유사한 데이터베이스를 설계하여 구축하는 것이 매우 중요하다. 그러므로 실험 결과의 정확성과 객관성을 유지하기 위해 모의실험의 기반이 되는 데이터베이스를 기반으로 실제 데이터베이스를 구축하여 시스템 성능 실험을 한다.

실험에 도입되는 요소로서 데이터베이스의 크기, 조건결합의 개수, 질의의 종류와 같은 변수들을 다양하게 변화시킴으로써 데이터베이스의 응답시간을 측정한다. 실험에서 나타난 응답시간을 시스템의 성능 평가의 가장 중요한 요소로서 간주하고 두 과정 간의 성능 차이 발생 원인으로 이용한다.

역정규화와 정규화된 테이블과의 확연히 구별되는 차이가 전체 시스템의 20퍼센트 미만의 수준으로 분석되기 때문에 데이터베이스 시스템 전체를 대상으로 실험을 수행하는 것보다 축소된 데이터베이스를 바탕으로 성능 평가를 하는 것이 더 명확한 결과를 얻을 수 있다. 그리고 축소된 데이터베이스에 대해서 실험에 사용되는 데이터의 규모와 질의의 수를 충분히 크게 하여 실험을 함으로써, 실험에 필요한 다양한 조건의 결합 수, 질의 종류, 질의 빈도의 발생이 가능함으로 전체 데이터베이

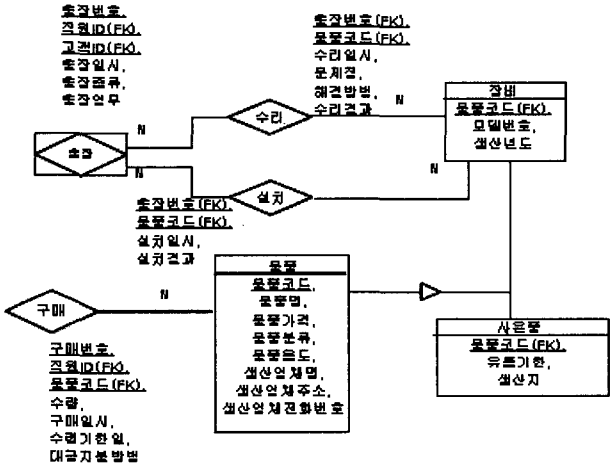

그림 4. 정규화 과정을 통한 ERD 일부
Fig. 4. Part of ERD by normalization.

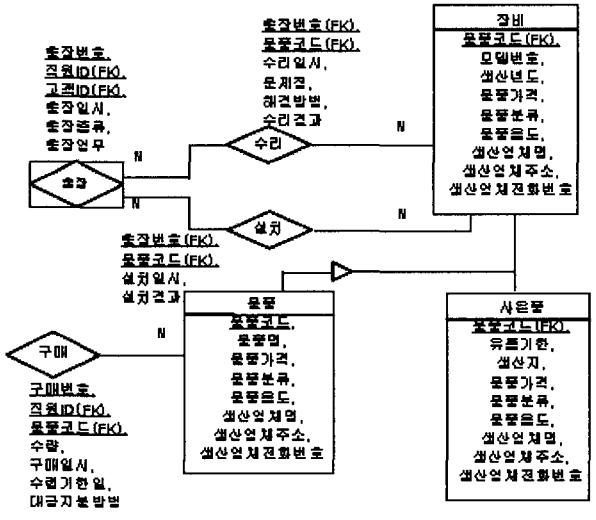

그림 5. 역정규화 과정을 통한 ERD 일부
Fig. 5. Part of ERD by denormalization.

표 1. 성능 평가를 위한 지수
Table 1. Parameters setting for simulation.

입력변수	종속변수	값의 범위
데이터베이스의 크기 (레코드의 개수)	저장공간의 크기 (KB), 응답시간(초)	1,000 - 10,000개
조건결합의 개수	평균 응답시간(초)	1 - 5개
질의의 종류(검색질의: 갱신비율)	응답 시간(초)	100개(0 - 100%)

스에 대한 평가대신 축소된 데이터베이스를 대상으로 한 성능평가가 보다 심층 분석을 하기위해 효율적이라 생각된다.

그림 4와 그림 5에서 나타난 업무는 각각 두 가지 방법을 기반으로 설계한 데이터베이스 중 실험에 사용된 부분으로 그림 5는 역정규화 기법 중 일부인 컬럼 중복을 사용하여 설계된 역정규화 모델의 특징을 보여 준다.

가. 성능 평가 지수

시스템 성능 평가에 사용되는 지수로는 우선 데이터베이스의 크기, 질의의 종류, 질의를 위한 조건 결합의 개수, 질의의 빈도이다. 표 1에서와 같이 각 지수에 대한 종속 변수, 값의 범위를 정하였다.

3. 성능 평가 결과 및 분석

본 절에서는 데이터베이스의 크기 변화에 따른 응답시간의 측정, 저장 공간의 크기 변화, 검색 조건 결합의 수를 기준으로 성능 실험을 한 결과를 분석한다. 데이터베이스의 크기에 따른 성능 변화를 위해 레코드의 개수를 1000개에서 10000개까지 1000개 단위로 늘려가면서 실험을 한다. 또한 역정규화에서 중요하게 고려되는 요소가 검색 시 조건 결합을 줄여서 성능을 높이자는 것인데 역정규화의 효용성을 검사하기 위해 조건 결합의 수을 테이블 1개에서 5개까지 늘려가면서 성능을 측정한다.

가. 데이터베이스 크기에 따른 응답 시간 평가

정규화, 역정규화 방법으로 설계된 데이터베이스를 바탕으로 각 방법의 효율성을 측정하기 위하여 데이터베이스의 크기에 따른 응답시간의 변화를 검사한다. 응답 시간을 측정하기 위해 사용되는 질의의 종류는 검색질의와 갱신질의를 동률로 하여 총 질의 개수를 100개로 고정시킨 가운데 데이터베이스의 크기만을 변화시킨다. 각 방법에 따라 설계된 데이터베이스가 레코드 수

표 2. 데이터베이스 크기의 변화에 따른 응답 시간 변화 (단위: 초)
Table 2. Response time by changing the size of database. (unit: seconds)

레코드 수 / 모델	1000	2000	3000	4000	5000	6000	7000	8000	9000	10000
정규화	1.483	2.046	1.127	1.607	1.873	1.937	2.094	2.233	2.407	2.186
역정규화	2.467	3.106	3.156	3.517	3.750	3.767	4.190	4.237	5.457	5.767

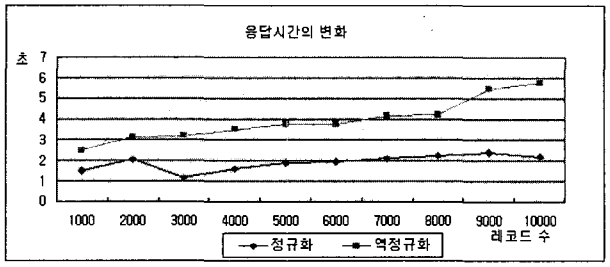

그림 6. 데이터베이스 크기의 변화에 따른 응답 시간의 변화
Fig. 6. Table 2. Response time by changing the size of database.

가 변화함에 따라 반응하는 응답 시간은 표 2와 같다.

그림 6에 의하면 정규화 모델과 역정규화 모델 모두 레코드 수의 크기가 증가함에 따라 응답 시간도 선형적으로 증가하고 있다. 하지만 역정규화된 데이터베이스인 경우 레코드 수가 8000에서 9000으로 증가하는 부분에서 급격히 응답시간이 증가한 것을 볼 수 있다. 이러한 역정규화 모델의 급격한 응답 시간의 증가는 본 모의 실험에서 사용한 데이터 무결성 유지를 위한 실시간 동기화를 시도한 것과 관련이 있다. 데이터 무결성을 위해 실시간으로 갱신된 데이터를 테이블 트리거를 사용하였는데, 검색인 경우는 영향이 없으나, 갱신 질의인 경우 무결성을 유지하기 위해 중복되어 있는 테이블마다 데이터를 갱신하는 작업을 수행해야만 하기 때문에 응답시간이 정규화 모델보다 더 걸릴 수밖에 없는 것이다. 시스템의 레코드 수가 증가하면 할수록 이러한 트리거에 의해 갱신되는 레코드 증가하므로 전체 시스템 응답 시간에 영향을 미치게 된다. 정규화 모델인 경우 레코드의 수가 2000에서 3000으로 증가하는 부분에서 응답시간이 급격히 감소한 이유는 정규화된 데이터베이스에서는 조인 연산이 일어나는 필드의 많은 부분이 주기억장치에 상주할 수 있기 때문에 응답시간이 계속적으로 증가하는 것이 아니라 어느 시점을 기준으로 응답시간은 줄어들게 되는 것 같다.

표 3. 검색 조건 결합수에 따른 응답 시간의 변화
(단위: 초)

Table 3. Response time by the number of joins.
(unit: seconds)

테이블수 / 모델	1	2	3	4	5
정규화	0.936	1.000	1.687	1.920	2.278
역정규화	0.954	0.954	0.954	0.954	0.954

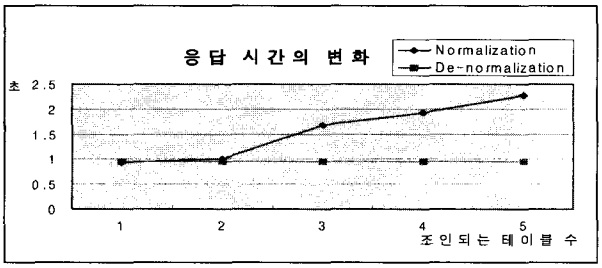

그림 7. 검색 조건 결합수에 따른 응답 시간의 변화
Fig. 7. Response time by the number of joins.

나. 검색 조건 결합수에 따른 성능 평가

역정규화를 하는 주요 목적은 데이터 검색 시 되도록 조인을 줄여서 시스템의 성능을 높이고자 하는 것이다. 그러므로 검색 시 사용되는 조인이 역정규화의 효용성에 얼마나 영향을 미치는지 측정하기 위해 검색 시 조인되는 테이블의 수를 변화시켜가면서 응답 시간을 조사하였다. 특히 역정규화 모델의 효용성을 살펴보기 위하여 실험에 사용된 SQL문은 장비, 사은품 테이블을 기준으로 조인을 각 20회 씩 수행하여 표 3과 같은 결과를 얻었다.

그림 7에서 역정규화 모델인 경우 질의에 필요한 데이터가 하나의 테이블에 존재하기 때문에 추가적인 조인을 할 필요가 없기 때문에 응답시간의 변화가 일어나지 않는다. 정규화 모델인 경우 조인되는 테이블 수가 증가할수록 응답 시간이 증가한다. 그러나 테이블이 하나씩 증가 할수록 증가되는 평균 응답 시간이 0.3355초이므로 절대 시간상으로 아주 미미할 정도이고, 대부분 소요된 시간이 조인에 따른 물리적 디스크 접근으로 인한 것으로 판단이 되므로 실제 업무에서 레코드의 수가 크게 증가된다 하더라도 응답 시간의 증가는 미미할 것으로 생각된다.

Ⅳ. 결 론

시스템의 성능을 높이기 위해 실제 기업에서 시행되고 있는 데이터의 중복을 허용한 역정규화 방법의 효율성을 측정하기 위해 정규화의 장점인 데이터 무결성, 비중복성, 정확성 측면을 배제한 온전한 시스템의 성능 측면에서 두 방법의 효용성을 비교 분석하였다. 정규화와 역정규화와의 성능 평가를 위해 고객관련 업무를 바탕으로 두 가지 방법을 적용하여 데이터베이스 시스템을 구축하고 분석하여 비교하였다. 업무분석을 통해 도출된 정규화 된 테이블과 이를 바탕으로 역정규화를 실행한 테이블 들 간의 시스템 거래 발생 시 나타나는 성능 평가를 응답시간 측면에서 비교하였다. 이 성능 평가를 통해 역정규화를 통한 시스템 구축이 기업의 정보시스템의 성능에 기여하고 있는지 여부를 파악하고자 하였다.

실무에서 데이터베이스를 설계하는 시스템 설계자들이 검색 시간의 효율성을 높이기 위해 간혹 역정규화를 사용하고 있지만 단계적인 정규화 과정을 거치지도 않은 상태에서 역정규화를 시행함으로써 실제 업무에서 성능 향상을 보장할 수 없는 형편이다. 비교 실험에 의하면 검색 질의인 경우 역정규화 과정을 수행한 데이터베이스가 응답 시간이 적게 걸리지만 갱신 질의인 경우에는 정규화 과정을 거친 데이터베이스보다 응답 시간이 더 걸리는 것을 알 수 있었다. 검색 질의와 갱신 질의가 함께 빈번히 발생하는 일반 업무에서는 역정규화의 성능이 오히려 더 떨어지고, 사용자가 많은 경우 데이터 점유를 위한 로킹이 빈번히 발생하므로 시스템의 성능저하가 우려된다. 그러므로 영업업무와 같은 기업 환경에서는 정규화 과정을 거친 데이터베이스가 더 성능이 우월함을 알 수 있었다.

본 논문에서는 대상업무가 실제 인터넷 서비스업체의 시스템을 대상으로 하였으나, 시스템을 보다 간략화하여 영업 정보시스템을 실험 대상으로 하였다. 실험에 사용된 데이터도 가상으로 생성한 것으로 실제 업무 상황과는 다소 다를 수 있다. 질의 검색인 경우는 큰 시스템에서 발생할 수 있는 복잡한 조인검색에 대해서는 성능 결과를 제시하지 못하였다.

앞으로 수행되어야 할 연구는 보다 현실감 있는 성능 평가를 위해서 대상 시스템을 축소할 것이 아니라 전체 시스템을 구축한 상태에서 다수의 사용자들이 자유롭게 접근할 수 있는 환경을 조성하여 시스템 부하에 따른 응답시간의 변화를 정확히 측정하는 것이다. 또한 역정규화 기법 들 중 한 가지 방법만을 선택하여 비교 실험을 하였는데, 앞으로는 더 나아가 다른 역정규화 기법과의 비교 실험으로 확장되어야 할 것이다.

참 고 문 헌

[1] S. Moon, "Unclassified data is merely garbage: data modeling is more crucial than programming," Hitech Information, vol. 14, pp. 50-51, 2003.
[2] R. Y. Wang, V. C. Storey and C. P. Firth, "A framework for analysis of data quality research," IEEE transactions on Knowledge and Data Engineering, vol. 7, no. 7, pp. 623-640, 1995.
[3] G. L. Sanders and S. Shin, "Denormalization Effects on Performance of RDBMS," Proceedings of the 34th International Conference on System Sciences, Hawaii, pp. 1-9, 2001.
[4] D. B. Bock and J. F. Schrage, "Denormalization guidelines for base and transaction tables," ACM Special Interest Group on Computer Science Education, vol. 34, no. 4, pp. 1, 2002.
[5] M. Hanus, "To normalize or denormalize, that is the question," in Computer measurement Group (CMG) Proceedings, No. 1, Chicago, IL, USA, 1994.
[6] U. Rodgers, "Denormalization: Why, What, and How?," in database Programming & Design, Dec., 1989.

저 자 소 개

이 혜 경(정회원)
1979년 2월 숭실대학교 전자계산학과 졸업
1985년 4월 University of Illinois(Urbana-Champaign) 전산학과 석사
2000년 2월 성균관대학교 정보공학과 박사
1988년 3월~1989년 2월 국립천안공업전문대학 전자계산과 전임강사
1992년 3월~2001년 8월 경인여자대학 멀티미디어정보 전산학부 조교수
2001년 9월~현재 용인송담대학 컴퓨터게임정보과 조교수
<주관심분야 : 데이터 모델링, 모바일 컴퓨팅, 동시성제어>