정규화 문제는 일반적 생각보다 더 기묘하다

필자는 뜻하지 않게 심도있는 정규화 이론을 둘러싼 다양한 문제들에 위해 본 칼럼의 많은 지면을 할애한 듯이 보인다. 결합 종속성과 최종 정규형(the final normal form, 5NF, 또는 PJ/NF)에 대하여 이야기하였다. 본 칼럼에서 필자는 일련의 추가적인 정규화 관련 문제들에 관해 논의하고자 한다. 적어도 몇몇 경우들에 있어서 이러한 문제들은 그들의 특성상 병적으로 간주될 수도 있지만, 여전히 우리의 궁극적 주제에서 외면할 수 없는 분야인 것이다.

BCNF, 4NF, 그리고 2진(BINARY) 관계변수들

필자의 이전 칼럼 '정규형이란 매우 흥미로운 것(2장)'에서 본인은 두가지 문제를 제기했는데, 그 둘은 모두 정규화와 관련된 것들이었다. 그 첫번째 문제는 다음과 같다.

■ *연관 모델이 지원하는 일반적 n진 관계변수들에 비해 2진 관계변수가 지닌 이점 중의 하나는 2진 관계변수가 항상 Boyce-Codd 정규형(BCNF)의 형태이기 때문이라는 사실은 종종 논쟁의 대상이 된다. 이 주장의 당위성을 증명하든지, 그렇지 않다면 반례(反例)를 들어 그 주장의 부당함을 증명하시오.*

상기의 주장은 - 거의 참에 가깝지만 - 거짓으로 밝혀진다. 2진 관계변수 R {A, B}를 생각해 보자. R이 BCNF의 형태라는 것을 증명하기 위해 우리는 먼저 R에 의해 충족되는 모든 비정규 기능종속은 R의 후보 키들에 의해 함축된다는 사실을 증명해야 한다. 그러므로 R이 틀림없이 BCNF의 형태라는 것을 증명하는 방법은 다음과 유사할 것이다:

1. 만약 R이 A→A와 같은 평범한 기능 종속을 제외한 어떠한 FD도 충족시키지 않는다면, R은 분명히 BCNF의 형태이다.(만약 FD가 전혀 존재하지 않는다면, BCNF이기 위한 요건은 무의미하다. 이 경우 R은 전체 키임을 주목하자.)

2. 만약 R이 FD A→B를 충족한다면, A는 당연히 하나의 후보 키이다.(K는 약분할 수 없는 경우에만 후보 키이며, FD K→C는 해당 관계변수의 전체 칼럼 C전반에 걸쳐 유지된다.) 만약 R이 FD B→A 또한 충족한다면, B 또한 당연히 후보 키이다. 이러한 후자의 조건이 유지되는지의 여부에 상관없이, 유일한 비정규 FD들은 후보 키들에 의해 함축되는 것들 뿐이며, 그러므로 R은 분명히 BCNF의 형태인 것이다.

3. 2번의 증명법은 분명히 A와 B에서 대칭을 이룬다. 그러므로 만약 R이 FD B→A를 충족한다면, R이 FD A→B를 또한 충족하는지의 여부에 상관없이 재차 이것은 BCNF의 형태이다. 따라서 R은 틀림없이 BCNF의 형태인 것이다.

상기의 증명법에서 잘못된 점이라면 이것이 또 하나의 가능성을 경시했다는 점이다. 다시 말해, R이 이것의 좌측에 공집합을 가진 하나의 FD를 충족하는 경우 등이다. 예를 들어, 'STATE'라고 해석되는 것은 'COUNTRY'의 구성원이며 COUNTRY는 모든 열(row)에서 미합중국인 관계변수 USA{COUNTRY, STATE}를 생각해 보자.

USA

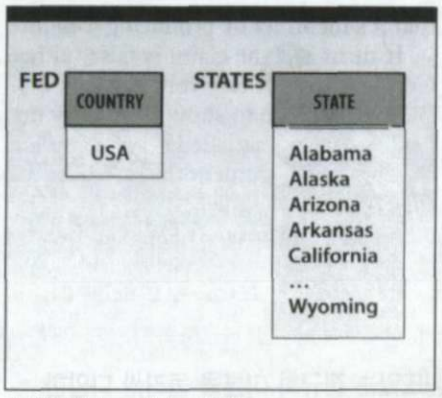
COUNTRY	STATE
USA	Alabama
USA	Alaska
USA	Arizona
USA	Arkansas
USA	California
...	...
USA	Wyoming

〈그림 1〉 Relvar USA (current relation value)

다음의 FD: Ø → COUNTRY가 관계변수 USA에서 유지되며, Ø 은 여전히 후보 키가 아니라는 사실을 눈여겨 보자. 그러므로 USA는 BCNF의 형태가 아니다.(필자가 이전의 기고에서 지적했듯이 어떤 칼럼 C에서 FD Ø →C를 충족하는 모든 관계변수는 전체 열들이 반드시 C와 동일한 값을 가지며, 상기의 예에서 보여지는 것이 바로 그러한 경우이다.)

예 USA를 좀더 살펴보겠다. USA는 BCNF의 형태가 아니기 때문에 정규화 이론은 이 관계변수가 프로젝션들로 비손실 재구성되어야 한다고 제안한다. 말하자면 〈그림 2〉에서 제시됐듯이 FED와 STATE 등이 그런 프로젝션들이다. 그러한 재구성에 있어서 STATE를 위한 유일한 후보 키는 물론 칼럼 STATE다. 더불어 FED를 위한 유일한 후보 키는 공집합이다.

필자가 이전의 칼럼에서 Ø 을 후보 키로 가진 모든 관계변수는 틀림없이 최대 하나의 열을 지니고 Ø 이외의 다른 후보 키를 소유하지 않는다는 점을 이야기 할 때 '공백 키'들의 존재 가능성에 대



〈그림 2〉 Normalizing USA

하여 같이 논의했었다. 이러한 특성들은 상기의 예에서도 유지된다.

물론, 관계변수는 아주 빈번히 갱신되는 것이 아니기 때문에 관계변수 USA가 BCNF의 형태가 아닌 사실이 큰 문제는 아니라고 반론할 수 있을 것이다. 반면에, 이러한 갱신 작업이 전무(全無)한 일이 아니고, 관계변수는 일련의 중복성을 수반하므로 몇몇 갱신 작업상의 이상 상태를 초래하는 주된 이유라는 사실을 부정하기는 어렵다. 그러나 필자는 상기의 예가 우리가 '궁극적' 경지(5NF)의 아래 단계인 정규화 작업 과정에 선택하고 싶은 방법임을 확신한다.

여기서 다른 사실 하나를 같이 지적하고자 한다. 지난호 칼럼 최종 정규형에서 필자는 정규화의 관점에서 볼 때 unary 프로젝션들은 동 프로젝션으로의 재구성이 보통 비손실적이 아니기 때문에 일반적으로 부적절하다고 이야기한 바 있다. 그럼에도 불구하고 우리는 상기의 예에서 분명히 unary 프로젝션으로의 비손실 재구성을 목격했다. 〈그림 1〉의 관계 USA는 〈그림 2〉의 관계들 FED와 STATES의 결합과 그 결과가 일치한다.(공유 칼럼이 존재하지 않을 경우, 결합은 카르테시안 곱(Cartesian product)으로 변질된다는 사실을 주지해야 한다.)

우연히도, 앞의 논의는 또 다른 유사한 문제를 제기한다. 즉, 그것은 BCNF의 형태이지만 4NF의 형태가 아닌 2진 관계변수가 존재하는지에 대한 질

문이다.

그 질문의 해답은 실상 어떤 종류의 '카르테시안 곱' 관계변수도 그런 조건을 만족시킨다는 것이다. 예를 들어, 공급자-부품 데이터베이스와 S{S#} Times P{P#}라고 정의된 관계변수 하나를 생각해 보자. 이 관계변수는 분명히 BCNF의 형태지만, 아래와 같은 다중 값을 가진 종속(multivalued dependencies, 또는 MVDs)을 충족시키기 때문에 4NF의 형태는 아니다.

Ø →→ S#
Ø →→ P#

이러한 MVD들은 아마도 설명을 필요로 할 것이다. 일반적으로 A, B, C가 칼럼들의 집합인 관계변수 R{A, B, C}는, R의 어떤 한 합법적 값 내에서 일정한 A의 값에 대응하는 B 값들의 집합이 그 A 값에 의존할 경우에만 아래와 같은 MVD를 만족시킨다고 일컫어진다(C의 값들은 이 경우 부적절하다.)

A →→ B

다시 말해서 만약 열(a,b1,c1)과 (a,b2,c2) 양쪽이 다 존재한다면, 열(a,b1,c2)와 (a,b2,c1) 또한 둘 다 존재하는 것이다. 만약 MVD A→→B가 충족된다면, MVD A→→C 도 마찬가지로 충족된다는 사실을 주지하기 바란다. MVD들은 항상 이러한 방식으로 쌍으로 함께 성립된다.

이제 A를 칼럼들의 공집합이라고 가정하자. 그러면 MVD들은 아래와 같이 변할 것이다.

Ø →→ B
Ø →→ C

그리고 이러한 MVD들이 의미하는 것은 만약 열(b1,c1)와 (b2,c2)가 둘 다 존재한다면, 열(b1,c2)와 (b2,c1) 또한 둘 다 존재한다는 사실이다. 다시 말해서 관계변수 R은 "카르테시안 곱" 구조를 가지고 있다. 이러한 두 MVD들은 일반적이지 않다는 사실에 유념하기 바란다. (이들은 분명히 모든 2진 관계변수들 R {B, C}에 의해 충족되지 않는다.)

그러므로 이전의 예 S{S#} TIMES P{P#}로 다시 돌아가서 이야기하면, 필자는 해당 관계변수는 4NF의 형태가 아니라는 것을 보여준 것이다. 해당 관계변수가 4NF의 형태가 아닌 이유는, 이것이 비일반적이며 후보 키들에게 함축되는 FD를 포함해서 어떠한 FD들도 아닌 MVD들을 충족시키기 때문이다.

상기의 관계변수가 이러한 두 MVD들을 충족시킨다는 말은 이 관계변수가 결합 종속(join dependency, JD) *{S#,P#}를 충족시킨다는 뜻과 같다. 왜냐하면 이 관계변수가 S#와 P# 상에서 unary 프로젝션으로 비손실-재구성 될 수 있기 때문이다. 그런데, 상기의 관계변수는 다음과 같은 MVD들 또한 충족시킨다.

S# →→ P#
S# →→

and

P# →→ S#
P# →→ Ø

그러나, 이러한 MVD들은 일반적이다. 왜냐하면 그들은 칼럼 S#과 P#를 지닌 모든 2진 관계변수들에 의해 충족되기 때문이다.

BCNF, 4NF, 그리고 '전체 키' (ALL KEY) RELVAR들

두번째 문제는 다음과 같다.

■ *만약 BCNF 관계변수가 "all key"(다시 말해 모든 칼럼이 해당 관계변수의 독점적 후보 키일 때)*

일 경우, 정규화의 최고 단계(5NF)에 위치하는데 실패한다는 주장은 종종 논쟁의 대상이 된다. 이러한 주장의 정당성을 증명하거나, 반례를 들어 이 주장의 부당성을 입증하라.

다음은 최근의 데이터 베이스 디자인 관련 책자에서 발췌한 이러한 주장의 예이다.

"일단 데이터의 구조가 BCNF의 형태라면, 나머지의 모든 정규화 작업상의 문제들은 모든 칼럼들이 특정 키의 일부인 관계변수를 운용할 때 발생한다."

실상 이러한 주장은 옳지 않다. 특정 과목(C)을 어떤 교사(T)가 가르칠 수 있고, 어떤 교재(X)를 사용하는지를 보여주는 관계변수 CTX {C,T,X}를 생각해 보자. 위에서 규정된 것들의 보다 자세한 의미는 아래와 같다.

- 첫째, 약간 비현실적이지만 교사와 교재가 상호간에 독립적이라고 가정한다. 무슨 말이냐 하면, 어느 특정 과목을 가르칠 때는 교사가 누구인지에 관계없이 동일한 교재가 사용된다는 뜻이다.
- 둘째, 특정 교사/교재의 조합이 최대 한 과목에 관련해서 발생한다고 가정한다.
- 마지막으로 어느 한 교사나 교재가 과목 수에 관계없이 관련될 수 있다고 가정한다.

〈그림 3〉은 이러한 가정들에 충실한 예인 CTX 관계 값을 보여준다. 이 경우 관계변수 CTX에 사용할 수 있는 유일한 비정규 종속은 다음과 같다.

1. $C \twoheadrightarrow T$
   $C \twoheadrightarrow X$

이러한 두 MVD들은 교사들과 교재들이 상호간에 독립적이기에 유지된다.

2. $\{T,X\} \rightarrow C$

상기의 FD는 교사/교재의 조합이 최대 한 과목에만 관련되어 발생하기 때문에 유지된다. (다시 말해, T와 X는 함께 하나의 후보 키를 형성하며, 실상 이것은 관계변수 CTX의 유일한 후보 키이다.)

그러므로 우리는 다음과 같은 사실들을 알 수 있다:

- 상기의 1로부터 우리는 CTX가 4NF가 아니라는 사실을 알 수 있다. 왜냐하면, 이것이 후보 키들에 내포되어 있는 FD들을 포함해서 어떠한 종류의 FD도 아닌 비정규 MVD들을 충족시키기 때문이다. 그리고 이것이 4NF의 형태가 아니므로 당연히 5NF의 형태도 아닌 것이다.
- 상기의 2로부터 우리는 CTX가 BCNF의 형태라는 사실을 알 수 있다. 그 이유는 이것이 충족시키는 유일한 비정규 FD가 후보 키들에 내포된 바로 그 FD이기 때문이다. 그러므로 관계변수 CTX는 BCNF의 형태지만, 4NF 또는 5NF의 형태가 아니다. 또한 이것은 여전히 'all key'가 아니다. 이 경우 독점적 후보 키는 조합 {T,X}이다.

그러므로 정규화 이론은 관계변수 CTX가 〈그림 4〉에서 제시된 방식으로 비손실 재구성되기를 제안한다. 〈그림 4〉에 제시된 프로젝션 CT, CX는 둘 다 "all key"이며, 실상 그 둘은 모두 4NF의 형태인 동시에 5NF의 형태이다. 그러나 이러한 재구성하에서는, CTX에 적용되는 후보 키의 제약 조건이

CTX

C	T	X
Physics	Prof. Green	Basic Mechanics
Physics	Prof. Green	Principles of Optics
Physics	Prof. Brown	Basic Mechanics
Physics	Prof. Brown	Principles of Optics
Math	Prof. Green	Statics and Dynamics
Math	Prof. Green	Vector Analysis
Math	Prof. Green	Trigonometry

〈그림 3〉 Sample CTX relation value

CT

C	T
Physics	Prof. Green
Physics	Prof. Brown
Math	Prof. Green

CX

C	X
Physics	Basic Mechanics
Physics	Principles of Optics
Math	Statics and Dynamics
Math	Vector Analysis
Math	Trigonometry

〈그림 4〉 Normalizing CTX

두가지 관계변수를 이어주는 제약 조건으로 변한다. 이 점에 대해서는 다음에 다시 논의하겠다.

상기의 CTX 예와 관련해서 한가지 사실을 더 지적하고자 한다. 우리가 보았듯이 CTX는 4NF가 아닌 BCNF의 형태이며 대략 아래와 같은 형태를 취한다.

R {A,S,C}
CANDIDATE KEY {A,B}

그리고 CTX는 또 MVD A→→B와 A→→C를 충족시킨다. 관련 서적들에서 종종 발견되는 내용 중에, 4NF의 형태가 아니며, 'all key'가 아닌 BCNF 관계변수는 반드시 이러한 형식을 따른다라는 주장이 있다. 하지만 이러한 주장 역시 옳지 않다. 다음과 같은 관계변수를 생각해 보라.

R {A,B,C,D}
CANDIDATE KEY {A,C}
CANDIDATE KEY {A,D}
CANDIDATE KEY {B,C}
CANDIDATE KEY {B,D}

R에 의해 충족되는 유일한 비정규 FD들은 상기의 네 가지 후보 키들에 내포된 것들 뿐이라고 가정하자. 그렇다면 T(＊＊＊역자주 : R의 오자로 사료됨)는 BCNF의 형태인 것이다. 이제 R이 아래와 같은 비정규 MVD들도 충족시킨다고 가정해 보자.

Ø →→ AB
Ø →→ CD

상기의 두 MVD들은 후보 키들에 내포된 경우를 포함한 어떤 종류의 FD도 아니다. 그러므로 R은 4NF의 형태가 아닌, BCNF 형태의 관계변수인 것이다. 그리고 R은 'all key'가 아니고, 본 논의의 초반부에서 제시된 일반적 형식도 아니다. 이러한 모든 제약 조건들과 모순되지 않는 R의 표본 값은

R

A	B	C	D
a1	b1	c1	d1
a2	b2	c2	d2
a1	b1	c2	d2
a2	b2	c1	d1

〈그림 5〉 Relvar R (sample relation value)

〈그림 5〉에 나와있다.

만일 상기한 예들이 난해하게 느껴진다면, 다음과 같은 예로 독자들의 이해를 돕고자 한다. 첫째, MON {A,B}를 하나의 관계변수라 하고, 이 관계변수의 각 열은 월요일동안 측정된 시간 간격을 나타낸다고 가정하자(A와 B는 둘 다 timestamp들이다).

또, 어떤 임의의 시간 간격들도 동일한 시작 시간 A 또는 종료 시간 B를 가지지 않는다고 가정하자.(그러므로 A와 B는 둘 다 후보 키이다) 마지막으로, 시작 시간은 항상 종료 시간보다 적다고 가정하자.

둘째, TUE {C,D}를 상기와 동일한 값을 화요일 동안 측정한 관계변수라고 가정하자. 그러므로, 모든 A 값들과 B 값들은 항상 모든 C 와 D의 값들보다 적다는 점에 유의하자. 끝으로, 관계변수 R은 MON TIMES TUE로 정의된다고 가정하자.

참고:R은 4NF의 형태가 아니기 때문에, 정규화 이론은 R을 AB와 CD의 두 binary 프로젝션으로 해체시키도록 제안할 것이다. 하지만 그렇다면 먼저 논의된 CTX의 예와 마찬가지로 R에 적용되는 후보 키의 제약 조건은 해당 재구성 내에서 두 가지 관계변수들을 이어주는 제약 조건으로 변한다. 필자는 다음 칼럼에서 위의 내용에 관해 다시 논의할 것이다.

▶ C. J. Date : 필자는 관계형 데이터베이스 시스템의 전문가이며 컨설턴트로서 초청강사와 자유기고가로서 활동하고 있다.