데이터 제공원의 신뢰도를 고려한 확장 관계형 데이터 모델

정 철 용* · 이 석 균** · 서 용 무***

An Extended Relational Data Model for Database Uncertainty Using Data Source Reliability

Chul-Yong Jung* · Suk-Kyoon Lee** · Yongmoo Suh***

Abstract

We propose an extended relational data model which can represent the reliability of data. In this paper, the reliability of data is defined as the reliability of the source, from which the data originated. We represent the reliability of data at the level of attribute values, instead of tuples, then define the selection, product and join operators.

* 상명대학교 경영학과
** 단국대학교 전산통계학과
*** 고려대학교 경영학과

1. 서 론

90년대 이후 객체지향 데이터베이스, 멀티미디어 데이터베이스 등의 등장으로 더불어 데이터베이스의 기술은 급속도로 발전해왔으나 아직도 본질적으로 단순한 데이터의 저장과 검출(retrieval)이라는 기본적인 수준에 머무르고 있다. 인간생활에는 예측, 진단 등으로 인한 불확실한 데이터의 사용이 흔히 이루어지고 있으나, 불확실한 데이터를 위한 데이터베이스 시스템은 아직 상용화되어 있지 못하다. 이는 부분적으로는 데이터베이스의 데이터가 정확해야만 한다는 지금까지의 사용자의 의식의 한계에 기인하는 것으로 볼 수도 있으나, 인공지능의 전문가 시스템 분야에서는 불확실한 정보의 사용이 상당히 보편화되어있고 많은 상용 전문가시스템 쉘(shells)에서 여러 가지 불확실성 추론 기능을 제공하는 것을 볼 때, 필요에 의해 차차 해결될 수 있는 문제로 보인다. 점차 추론이 가능한 지능형 데이터베이스에 대한 요구가 증가하므로 가까운 미래에 데이터베이스 시스템에서 불확실한 정보의 처리는 필수 불가결할 것으로 보인다.

80년대, 90년대에 들어와 많은 연구가 있었으나, 여러 가지 측면에서 필요한 이론적 기반이 아직 이루어져 있지 못한 탓으로 보인다. 기존의 연구는 주로 불확실한 데이터의 표현형태에 관련된 것으로 퍼지이론, 확률이론, the Dempster-Shafer theory 등의 이론을 사용하여 확장된 관계형 데이터 모델의 제안에 관련 연구들이었다[Barbara et al. 1992; Buckles and Petry 1982; Cavallo and Pittarelli 1987; Lee 1992; Lee 1992; Morrissey 1990; Motro 1990; Pittarelli 1990; Prade and Testemale 1984; Raju and Majumdar 1988; Zemankova 1985]. 이들의 접근 방법은 단지 불확실한 데이터의 표현형태에 관심을 갖고 질의 평가의 정의에 그 연구가 집중되었다. 그러나, 이들의 연구는 불확실한 데이터의 표현 형태에 기인한 관계형 데이터 모델의 정의에만 관심을 갖고 관련 응용 분야와의 연계나 불확실한 정보를 다루는 전문가시스템과의 연동 등에는 연구는 거의 이루어지지 못했었다. 또한 불확실한 데이터의 표현 형태는 그 데이터의 도메인의 성격에 의해 결정될 수밖에 없기 때문에 어떤 접근 방법이 유일한 해결책이라기 보다는 상호 보완적이어야 할 것이고 오히려 이들의 접근 방법에 공통적인, 다시 말하면, 불확실한 정보를 다루는 데이터베이스의 공통적인 문제에 대한 연구가 이루어져야 할 것이다.

다른 형태의 연구 결과로 최근 Sadri는 불확실한 정보의 불확실성의 근원을 데이터를 제공하고 있는 제공원의 신뢰도로 보고 제공원 신뢰도벡터(source reliability vector)에 입각한 관계형 데이터 모델을 제시하였다[Sadri 1991]. 이 모델에서는 각 튜플(tuple)에 대하여 튜플의 내용을 제공한 제공원이 존재하고 튜플의 제공원의 신뢰도를 곧 튜플의 불확실성의 척도로 하여 질의 연산을 정의하였다. Sadri의 접근 방법은 처음으로 불확실한 정보의 제공원을 고려한 데이터 모델을 제기하였다는 데 그 의의가 있으나, 모든 튜플에 대한 제공원 정보를 유지해야 한다는 가정과 속성 차원의 정보의 제공원에 대한 고려가 없다는 문제점이 있다. 또한, Sadri의 데이터 모델은 지금까지 많은 연구가 이루어진 데이터의 불확실성 표현형태에 기초한 여러 연구 결과를 반영하지 못한다. 즉 정보의 도메인에 따라 확률 데이터, 퍼지 데이터, 불완전한(incomplete) 데이터 등 여러 형태의 데이터 표현형태가 필요할 뿐만 아니라, 신뢰도가 떨어지는 제공원에 의해 주어지는 정보의 경우, 그 불확실성 때문에 이러한 불확실한 데이터 표현형태가 필요하기 때문이다. 불확실성의 본질을 놓고 볼 때, 데이터의 표현형태와 그 제공원의 신뢰도의 두 가지 측면은 어느 하나만 떼어서 생각하기 어려운 경우가 종종 있다. 본 논문에서는 이러한 문제들을 해결하기 위해 제공원의 신뢰도를 고려한 확장 관계형 모델을 제시한다. 이 모델은 Sadri 모델의 한계 (튜플 중심의 제공원의 고려와 모든 제공원 정보의 유지)를 극복하여 속성 차원에서 선택된 속성들만의 신뢰도를 고려하여 보다 현실성을 높이며 이에 대한 질의 연산을 정의하였다. 이때, 각 제공원의 신뢰도 값이 주어지는 경우와 그렇지 않

은 경우에 대하여 셀렉션, 프로덕트, 조인 연산을 정의하고 있다.

2. 제공원의 신뢰도에 기초한 확장 관계형 모델

데이터의 불확실성의 원인은 여러 가지로 분석해 볼 수 있으나, 그 중의 중요한 것으로 데이터의 제공원의 신뢰도를 들 수 있다. 병원에서 환자를 진단할 때, 여러 가지 검사를 하고 그 검사 결과에 따른 진단을 할 수 있다. 이러한 경우, 수행한 검사의 신뢰도는 그 진단 결과에 영향을 끼친다. 대개 많은 의사의 진단은 일종의 불확실한 작업이라 할 수 있는 데, 이 경우 불확실성은 그 검사의 신뢰도에 기인한다고 할 수 있다.

2.1 Sadri의 확장 데이터 모델

데이터의 불확실성을 제공원의 신뢰도로 정의한 데이터 모델은 그다지 많지 않다. Sadri는 [11]에서 데이터베이스에 정보를 제공하는 제공원들의 리스트를 제공원 벡터로 정의하고 제공원 벡터에 입각한 확장된 관계형 모델을 제시하였다. 이는 그 전의 데이터에 있어서 불확실성 표현형태에만 관심을 기울였던 여러 연구 결과와 다른 접근 방법을 보였다. <표 1>을 통해 다음의 예를 참조해 보자.

〈표 1〉 SUPPLIER

S #	P #	Source Vector
s1	p1	1 0 0 0
s1	p2	1 0 0 0
s1	p3	1 0 0 0
s2	p2	0 1 0 0
s2	p4	1 0 0 0
s3	p3	0 1 0 0

S#와 P#는 테이블(릴레이션) Supplier의 속성이고 속성 Source Vector는 각 제공원들이 튜플의 존재에 어떤 영향을 주는 가를 나타내고 있다. 벡터의 길이가 4로 주어져 있는 것은 이 데이터베이스에는 4개의 제공원들이 있다는 내용이고 각 각 1, 0, -1 중의 한 값을 갖는다. j번째 제공원이 튜플의 내용을 확인하여 주었을 때, 제공원 벡터의 j번째 제공원의 값이 1로 나머지들은 0으로 정해진다. 대개 기저 테이블(base table)의 경우, 각 튜플의 제공원 벡터의 각각 제공원들은 1과 0으로 구성된다. 유도 릴레이션(derived relation)의 경우, 튜플의 제공원 벡터에 0, 1, -1의 값이 나타날 수 있다. j번째 구성원의 값이 -1 이라는 의미는 j번째 제공원이 틀릴 때 또는 j번째 제공원을 완전히 신뢰할 수 없을 때, 그 튜플이 내용이 옳다는 것이다. 어떤 질의의 결과로 <표 2>와 같은 테이블이 주어졌다고 하자.

〈표 2〉 Derived relation

S #	P #	Source Vector
s1	p1	1 0-1 0
s1	p2	1-1 0 1
s1	p2	0 0-1 1

이 테이블에서 첫 번째 튜플의 의미를 해석한다면, 제공원 1이 옳고, 제공원 3이 틀릴 때 (s1, p1)의 내용은 질의의 결과에 포함된다는 것으로 해석된다. Sadri는 제공원 벡터들간의 논리곱(conjunction), 논리합(disjunction), 부정(negation)에 대한 정의를 통하여 질의 연산자들(selection, projection, etc.)을 정의하였다. 상세한 설명을 위해서는 [Sadri 1991]을 참고하도록 하자.

위에서 설명한 Sadri의 접근 방법의 문제는 튜플 차원의 불확실성을 고려하고 이의 원인을 튜플의 내용을 제공한 제공원의 신뢰도로 삼고 있다는 점이다. 즉 속성 값의 불확실성(가령 확률 데이터)을 고려하고 있지 못하다는 점을 들 수 있다. 또한 데이터베이스에 저장되는 모든 데이터의 제공원을 고려한다는 것이 비현실적이다. 많은 경우, 불확실성이 고려되는 속성은 상대적으로 적은 수이기 때문에 이들만 특별히 고려하는 데이터 모델이 필요하다고 판단된다. 특히 데이터베이스에서 규칙(rule)의 사용이 점점 보편화되어 가는 것을 볼 때, 전문

가 시스템의 규칙들이 직접 데이터를 추론하여 그 결과가 저장되는 경우, 이러한 종류의 데이터들만 특별관리를 하는 것이 보다 더 현실적인 방법이다.

2.2. 제공원의 신뢰도에 기초한 확장 관계형 모델

이 절에서 제시하려고 하는 데이터 모델은 각 튜플의 모든 속성 값이 원자인 경우로 일부의 속성들은 완전히 신뢰할 수 없는 제공원들로부터 주어진 것으로 불확실한, 또는 신뢰할 수 없는 속성 값들을 가질 수 있는 경우이다. 가령 데이터베이스와 전문가시스템이 연동되어 전문가시스템의 규칙의 실행결과가 데이터베이스에 저장되는 경우를 생각할 수 있다. 또는 병의 진단의 경우와 같이 어떤 검사의 결과에 입각한 진단의 경우, 이는 검사의 신뢰도가 진단 결과의 신뢰도의 역할을 한다. 다음은 신뢰도를 고려한 확장 관계 스키마의 정의를 제안한다.

정의 1 : 확장 관계 스키마(extended relation schema)

확장 관계 스키마 R은 다음과 같이 정의된다. $R = \{A_1, A_2, ..., A_m, UA_1, UA_2, ..., UA_n, VA)$ for m, n ≥ 0. A_i 는 일반 속성이고, UA_j 는 불확실 속성들의 리스트로 각각 $UA_1 = (sA_1, dA_{11}, ..., dA_{1p})$, ..., $UA_i = (sA_i, dA_{i1}, ..., dA_{iq})$ 등으로 정의될 수 있다[1]. sA_i 는 dA_{i1}, ..., dA_{ip} 등의 속성 값들의 제공원을 나타내는 속성이고, dA_{i1}, ..., dA_{ip}는 sA_i로부터 주어진 속성으로 그 값을 완전히 신뢰할 수 없다. sA_i 를 dA_{i1}, ..., dA_{ip}의 제공원 속성이라고 하고 dA_{i1}, ..., dA_{ip} 들은 sA_i 의 데이터 속성이라고 한다. VA는 논리식을 위한 특별한 속성으로 그 의미(semantic)는 각 튜플의 적합도(validity)를 나타낸다.

$UA_i = (sA_i, dA_{i1}, ..., dA_{iq})$의 의미는 조건부 확률 개념과 비슷하다. sA_i 의 값이 완전히 신뢰할 수 있는 것일 경우, 각 dA_{i1}, ..., dA_{iq}의 값들은 옳은 것(valid)으로 간주되나, 그렇지 않은 경우는 dA_{i1}, ..., dA_{iq} 의 값의 적합도는 sA_i 값의 신뢰도를 반영한다. 뿐만 아니라 sA_i 값은 논리 변수로 해석되어 VA의 논리식을 구성하게 된다. 즉 sA_i 의 값는 다음 두 가지 역할을 담당한다[2] : (1) 관련된 불확실 속성 값들에 대한 제공원 값의 역할; (2) 관련된 불확실 속성 값들의 확실성을 나타내는 논리변수의 역할. 그러므로 질의의 결과 (또는 유도 릴레이션(derived relation))의 경우, 각 튜플이 결과에 포함되는 지의 적합도의 여부는 제공원 속성들의 값과 논리 접속사(∧, ∨, ¬)로 구성된 논리식, 즉 VA의 값으로 결정된다. 기저 테이블의 튜플의 경우, 튜플이 주어진 것으로 간주되거나, 또는 조건이 없음으로 인하여 조건식을 만족하는 것으로 간주되어 VA의 값은 디폴트로 참의 값을 갖게 된다. 그러나 질의의 결과에 포함되는 튜플은 조건식에 데이터 속성이 관계될 때 관련된 제공원 속성의 신뢰도가 반영되므로 결과에의 포함 정도(조건에 대한 튜플의 적합도)를 VA의 값 즉 논리식으로 나타내게 된다.

<표 3>은 정의 1의 확장 관계 스키마를 반영한 가상의 예로 여러 경제연구기관으로부터의 각종 금리예측을 보여 준다.

<표 3>의 예는 "예측기관명"이라는 제공원 속성의 값으로부터 "예측이율"이라는 데이터 속성의 값들이 주어진 경우를 보인 것이다. Forecast는 기저 테이블이므로 VA의 값은 참으로 주어져 있다. 다음은 위에서 정의한 데이터 모델에 대하여 셀렉션 연

1) 스키마의 정의에서 $UA_i = (sA_i, dA_{i1}, ..., dA_{iq})$은 확장 스키마의 (중첩 릴레이션(nested relation)의 서브릴레이션에 대응하는) 서브스키마의 개념이 아니라 제공원 sA_i가 제공하는 데이터 속성 dA_{i1}, ..., dA_{iq} 들을 논리적 단위로 그룹핑하는 역할만을 수행한다. 다시 말하면, 확장 스키마 $R = \{A_1, A_2, ..., sA_1, dA_{11}, ..., dA_{1q}, ..., sA_i, dA_{i1}, ..., dA_{ir}, ...VA\}$로 정의하여 직접 풀어 적어도 관계없으나 설명의 편의상 $UA_i = (sA_i, dA_{i1}, ..., dA_{iq})$로 묶어서 정의한 것이다.

2) sA_i의 도메인이 정확하게 기술되어 있지 않은 것은 [Sadri 1991]에서의 방법을 그대로 이용한 것으로 불필요한 정의와 용어사용을 줄이기 위함이다. 이를 정확히 구분하기 위해서는 제공원의 식별자으로 문자열 타입이 필요할 것이고 이에 대응하는 논리변수의 타입과 이를 사상시켜주는 함수의 사용이 필요할 것이다.

산을 정의하고자 한다. 관계 T와 연산조건 C에 대한 셀렉션 연산자는 ***select T where (C)***로 나타낸다.

〈표 3〉 Rate_Forecast

예측항목	예측기관명	예측이율	VA
회사채유통수익율	D연구소	12%	true
회사채유통수익율	K연구원	11.1%	true
CD유통수익률	K연구원	11.8%	true
콜금리	K연구원	11.3%	true
CD유통수익율	D연구소	12.5%	true

정의 2 : 셀렉션 연산

임의의 튜플 t (∈ T)와 조건 C에 대하여 A(t)는 튜플 t에 관계된 모든 속성들의 집합을 의미한다. 이때 셀렉션 연산자는 다음과 같이 정의된다.

$$\text{select T where (C)} = \{t \mid \exists t' \in T \wedge A(t) = A(t') \wedge (\forall a)_{a \in A(t)-\{VA\}}[t.a = t'.a] \wedge t.VA = \varphi(t', C)\}$$

위에서 $(\forall a)_{a \in A(t)-\{VA\}}[t.a = t'.a]$의 의미는 VA를 제외한 모든 속성에 있어서 튜플변수 t와 튜플변수 t'의 값이 같음을 의미하며, φ는 질의에 대해 결과되어지는 튜플들의 적합도를 나타내는 논리식을 계산하는 함수로 이의 정의는 다음과 같다.

정의 3 : 튜플의 적합도 함수 φ

임의의 튜플 t와 조건 C에 대하여, $\varphi(t, C)$는 논리식을 리턴하는 함수로 조건 C에 따라 다음과 같이 정의된다. θ는 {=, ≠, ≤, ≥, <, >} 가운데 하나의 비교연산자로 한다.

① simple condition case

when a condition C is satisfied,

$\varphi(t, A_i\ \theta\ c) = true \quad \varphi(t, A_i\ \theta\ A_j) = true$

$\varphi(t, dA_{ij}\ \theta\ c) = t.sA_i \quad \varphi(t, dA_{ij}\ \theta\ A_k) = t.sA_i$

$\varphi(t, dA_{ij}\ \theta\ dA_{ik}) = t.sA_i$

$\varphi(t, dA_{ij}\ \theta\ dA_{kl}) = t.sA_i \wedge t.sA_k$ where $i \neq k$

when a condition C is NOT satisfied,

$\varphi(t, A_i\ \theta\ c) = false \quad \varphi(t, A_i\ \theta\ A_j) = false$

$\varphi(t, dA_{ij}\ \theta\ c) = \neg t.sA_i \quad \varphi(t, dA_{ij}\ \theta\ A_k) = \neg t.sA_i$

$\varphi(t, dA_{ij}\ \theta\ dA_{ik}) = \neg t.sA_i$

$\varphi(t, dA_{ij}\ \theta\ dA_{kl}) = \neg t.sA_i \vee \neg t.sA_k$ where $i \neq k$

주의 : 비교 연산에 sA_i가 관련된 경우는 일반 속성 A_i로 처리한다.

② compound condition case

$\varphi(t, C_1 \vee C_2) = \varphi(t, C_1) \vee \varphi(t, C_2)$

$\varphi(t, C_1 \wedge C_2) = \varphi(t, C_1) \wedge \varphi(t, C_2)$

$\varphi(t, \neg C) = \neg \varphi(t, C)$

셀렉션 연산과 적합도 함수가 각각 정의 2와 정의 3과 같이 정의되었을 때, select Rate_Forecast where (예측항목 = 회사채유통수익률) ∧ (예측기관명 = D연구소)의 질의의 결과는 <표 4>와 같다. 주목할 것은 조건에 관련된 속성들이 보통 속성과 제공원 속성이라는 점과 튜플들의 VA의 값이 참/거짓으로 주어져 있다는 점이다. 참인 경우는 해당 튜플이 조건 (예측항목 = 회사채유통수익률) ∧ (예측기관명 = D연구소)을 만족한다는 의미이고 거짓인 경우는 만족하지 못한다는 의미로 주어진 질의에 대한 각 튜플의 적합도를 나타낸다.

〈표 4〉 select Rate_Forecast where (예측항목 = 회사채유통수익률)∧(예측기관명=D연구소)

예측항목	예측기관명	예측이율	VA
회사채유통수익율	D연구소	12%	true
회사채유통수익율	K연구원	11.1%	false
CD유통수익률	K연구원	11.8%	false
콜금리	K연구원	11.3%	false
CD유통수익율	D연구소	12.5%	false

질의의 조건에 데이터 속성이 관계되어 있는 경우, 관련 제공원 속성이 튜플의 적합도를 나타내는 논리변수로 이용된다. select Rate_Forecast where (예측이율 > 11.5%) 질의의 결과는 정의 2, 3에 기초하여 <표 5>에 주어져 있다.

〈표 5〉 select Rate_Forecast where (예측이율 〉 11.5%)

예측항목	예측기관명	예측이율	VA
회사채유통수익율	D연구소	12%	D연구소
회사채유통수익율	K연구원	11.1%	¬ K연구원
CD유통수익률	K연구원	11.8%	K연구원
콜금리	K연구원	11.3%	¬ K연구원
CD유통수익율	D연구소	12.5%	D연구소

<표 5>는 Forecast 관계 중에서 (예측이율 > 11.5%)를 만족하는 튜플들의 집합이다. 각 튜플의 적합도는 "예측이율" 데이터의 제공원인 "예측기관명" 속성값의 신뢰도에 달려 있다. 예를 들어 첫 번째 튜플의 적합도는 "예측이율 = 12%"이라는 데이터의 제공원 값인 "D연구소"의 신뢰도에 달려 있다는 의미이다. 두 번째 튜플은 (예측이율 > 11.5%)의 조건에 대해 "예측이율 = 11.1%"의 제공원 값 "K연구원"이 부정확하다면, 질의의 결과에 포함된다는 의미이다. 즉 튜플의 적합도가 제공원 값의 신뢰도를 반영한다. NOT이 포함된 조건의 셀렉션 연산을 보도록 하자. 질의 select Rate_Forecast where ¬ (예측이율 =< 11.5%)의 경우는 사실상 select Rate_Forecast where (예측이율 > 11.5%)와 같은 질의이다. 정의 2, 3에 기초하여 <표 6>을 참조한다.

<표 6> select Rate_Forecast where ¬ (예측이율 =< 11.5%)

예측항목	예측기관명	예측이율	VA
회사채유통수익율	D연구소	12%	¬ ¬ D연구소
회사채유통수익율	K연구원	11.1%	¬ K연구원
CD유통수익률	K연구원	11.8%	¬ ¬ K연구원
콜금리	K연구원	11.3%	¬ K연구원
CD유통수익율	D연구소	12.5%	¬ ¬ D연구소

<표 4>, <표 5>, <표 6>의 질의의 결과에는 긍정적인 정보 뿐만 아니라 부정적인 정보까지도 포함하고 있다. <표 4>의 경우, 조건을 만족하지 못하는 튜플의 경우 VA의 값에 false가 주어져 있다. <표 5>의 질의에서는 조건을 만족하지 못하는 경우, 조건을 만족시키지 못하는 데이터 속성의 제공원 값의 신뢰도를 부정함으로 질의의 결과에 포함하고 있다. 전통적인 관계형 모델이 폐쇄세계 가정(closed world assumption) 하에 긍정적인 데이터만을 허용하는 것과 같은 이유로 여기서 제시되는 확장 모델에서도 긍정적인 데이터만을 허용하고자 한다. 이는 새로운 모델의 목적이 조건을 만족하는 튜플들 중에서 관계된 제공원 값의 신뢰도를 고려하자는 것이지, 조건을 만족하지 못하는 경우는 굳이 고려할 필요가 없기 때문이다. 앞서 주어진 셀렉션 연산의 정의와 적합도 함수 φ의 정의를 다음과 같이 변환하고자 한다.

정의 4 : 폐쇄세계 가정 하의 셀렉션 연산

임의의 튜플 t (∈ T)와 조건 C에 대하여 C(t)는 튜플 t의 만족여부를 평가하는 술어(predicate)의 역할을 하고, A(t)는 튜플 t에 관계된 모든 속성들의 집합을 의미한다. 이때 셀렉션(selection) 연산자는 다음과 같이 정의된다.

$$\text{select } T \text{ where } (C) = \{t \mid \exists t' \in T \wedge A(t) = A(t') \wedge C(t') \wedge (\forall a)_{a \in A(t)-\{VA\}}[t.a = t'.a] \wedge t.VA = \varphi(t', C)\}$$

정의 2와의 차이는 C(t)가 술어의 역할을 함에 있다.

정의 5 : 폐쇄세계 가정 하의 튜플의 적합도 함수 φ

임의의 튜플 t와 조건 C에 대하여, $\varphi(t, C)$는 논리식을 리턴하는 함수로 조건 C에 따라 다음과 같이 정의된다. θ는 {=, ≠, ≤, ≥, <, >} 가운데 하나의 비교연산자로 한다.

① simple condition case (튜플 t가 다음 조건을 만족할 때)

$\varphi(t, A_i \ \theta \ c) = true \quad \varphi(t, A_i \ \theta \ A_j) = true$

$\varphi(t, dA_{ij} \ \theta \ c) = t.sA_i \quad \varphi(t, dA_{ij} \ \theta \ A_k) = t.sA_i$

$\varphi(t, dA_{ij} \ \theta \ dA_{ik}) = t.sA_i$

$\varphi(t, dA_{ij} \ \theta \ dA_{kl}) = t.sA_i \wedge t.sA_k$ where $i \neq k$

주의 : 비교 연산에 sA_i가 관련된 경우는 보통 속성 A_i로 처리한다.

② compound condition case

$\varphi(t, C_1 \vee C_2) = \varphi(t, C_1) \vee \varphi(t, C_2)$

$\varphi(t, C_1 \wedge C_2) = \varphi(t, C_1) \wedge \varphi(t, C_2)$

$\varphi(t, \neg C) = \varphi(t, C')$

where C′ is an equivalent formula produced by propagating NOT to the inside of formula C by using De Morgan's law and changing primitive comparison operators.

새로운 정의에서는 셀렉션 연산의 C(t′)을 포함

시켜 조건을 만족하는 튜플들에 대하여 적합도를 계산하게 된다. 따라서 적합도 함수의 정의에 VA의 값에 NOT (¬)을 허용할 필요가 없고, 혼합조건(compound condition)의 경우 부정에서도 NOT이 없는 동등한 논리식으로 변환하여 처리한다. 따라서 질의 select Forecast where ¬(예측이율 =< 11.5%)의 결과는 정의 4, 5에 기초하여 <표 7>에 새로이 주어져 있다. 주목할 것은 조건을 만족하지 못하는 튜플은 더 이상 나타나지 않는 것이다.

〈표 7〉 select Rate_Forecast where ¬(예측이율 =〈11.5%)

예측항목	예측기관명	예측이율	VA
회사채유통수익율	D연구소	12%	D연구소
CD유통수익률	K연구원	11.8%	K연구원
CD유통수익율	D연구소	12.5%	D연구소

같은 제공원 속성 값을 갖는 데이터 속성 값간의 비교(dA_{ij} θ dA_{ik})는 같은 제공원 속성 값의 신뢰도의 영향을 받고, 서로 다른 제공원 속성 값들로부터의 데이터 속성들간의 비교는 서로 다른 제공원 속성 값들의 신뢰도들의 영향을 받는다.

지금까지는 VA의 값, 즉 논리식을 통하여 각 튜플들의 적합도의 여부를 설명하였다. 그러나, 각 제공원 속성 값에 대한 수량화된 신뢰도 값이 주어질 수 있다면, 이를 이용할 수 있다. 제공원 속성 값의 신뢰도가 주어져 있다는 가정 하에 이를 계산하여 주는 신뢰도 함수를 다음과 같이 정의한다. 다음에서 dom은 각 속성의 도메인을 나타내는 함수이며, 예를 들어 테이블 Forecast에서 dom(예측값)는 실수 도메인, dom(예측기관명)은 문자열의 도메인을 의미한다.

정의 6 : 신뢰도 함수 cr

함수 cr : $\cup_i$ dom(sA_i) → [0, 1] 은 신뢰도 함수로 부르는데, 이는 제공원 속성 값을 인수로 받아 cr(s)는 s의 신뢰도 값을 리턴한다. 인수가 true/false의 경우, 다음과 같이 정의된다.

cr(true) = 1, cr(false) = 0.

신뢰도 함수 cr은 제공원 속성 값에 의해 결정되는 것으로 주어진 것으로 가정한다[3]. 이에 기초하여 확장 신뢰도 함수 Cr이 정의되는데 이는 VA의 값을 인수로 받아, 그 논리식의 신뢰도를 계산하는 함수이다.

정의 7 : 확장 신뢰도 함수 Cr

함수 Cr : dom(VA) → [0, 1] 는 확장 신뢰도 함수로 부르고 튜플의 적합도를 나타내는 논리식 f (또는 f_1, f_2)를 인수로 받아 그 튜플의 신뢰도를 리턴한다[4].

$Cr(f) = cr(f)$ when $f \in \cup_i dom(sA_i)$

$Cr(\neg f) = 1 - Cr(f)$

$Cr(f_1 \vee f_2) = 1 - (1 - Cr(f_1)) \times (1 - Cr(f_2))$

$Cr(f_1 \wedge f_2) = Cr(f_1) \times Cr(f_2)$

정의 8 : 수정된 확장 관계 스키마(extended relation schema)

확장 관계 스키마 R은 정의 1에 CR을 추가함으로 정의된다. $R = \{A_1, A_2, ..., A_m, UA_1, UA_2, ..., UA_n, VA, CR)$ for $m, n \geq 0$. CR은 각 튜플의 수량화된 적합도를 나타내는 속성으로 VA의 값을 Cr함수로 계량화한 값이다. 기저릴레이션의 튜플인 경우 1의 값을 갖는다.

위의 수정된 확장 관계 스키마와 확장 신뢰도 함수를 사용하여 Rate_Forecast 테이블을 다시 정의하면 <표 8>과 같다.

〈표 8〉 신뢰도 함수를 사용한 Rate_Forecast 테이블

예측항목	예측기관명	예측이율	VA	CR
회사채유통수익율	D연구소	12%	true	1
회사채유통수익율	K연구원	11.1%	true	1
CD유통수익률	K연구원	11.8%	true	1
콜금리	K연구원	11.3%	true	1
CD유통수익율	D연구소	12.5%	true	1

3) 제공원들의 신뢰도 값들의 저장 문제는 구현단계의 문제로 특정 테이블에 관련 제공원들의 신뢰도를 저장함으로 표현할 수 있고 cr함수는 이 테이블의 참조(lookup)으로 계산될 수 있을 것이다.

4) 확장 신뢰도 함수의 정의는 신뢰도 함수 cr이 각 제공원 속성의 신뢰도를 확률로 표현한다는 가정하에 이루어진다. 사건(Event) A, B가 독립일 때, Prob(A^c), Prob(A∪B), Prob(A∩B)의 정의와 동일함을 알 수 있다.

동일한 방법으로 신뢰도 함수를 사용할 경우의 셀렉션 연산을 다시 정의하면 다음과 같다.

정의 9 : 셀렉션 연산 - 신뢰도 함수 사용의 경우

$$\text{select } T \text{ where } (C) = \{t \mid \exists t' \in T \wedge A(t) = A(t') \wedge C(t') \wedge (\forall a)_{a \in A(t)-\{VA,CR\}}[t.a = t'.a] \wedge t.VA = \varphi(t', C) \wedge t.CR = Cr(t.VA)\}$$

예를 들어 제공원 값의 신뢰도 함수의 값이 다음과 같이 주어졌다고 하자: cr(D연구소) = 0.8, cr(K연구원) = 0.85. 이 경우 select Rate_Forecast where ¬ (예측이율 =< 11.5%)의 결과는 <표 9> (정의 6, 7, 8, 9에 기초)와 같이 나타난다.

<표 9> select Rate_Forecast where ¬ (예측값 =< 11.5%)

예측항목	예측기관명	예측이율	VA	CR
회사채유통수익율	D연구소	12%	D연구소	0.85
CD유통수익률	K연구원	11.8%	K연구원	0.8
CD유통수익율	D연구소	12.5%	D연구소	0.85

위에서 정의한 데이터 모델에 기초하여 프로덕트와 조인 연산은 다음과 같이 정의된다.

정의 10 : 프로덕트, 조인 연산

임의의 튜플 $t_1(\in T_1)$, $t_2(\in T_2)$에 대하여 $A_i(t_i)$는 튜플 t_i에 관계된 모든 속성들의 집합을 의미한다. 이때 프로덕트(Cartesian product) 연산자와 조인(join) 연산자는 다음과 같이 정의된다.

$$\text{product } T1, T2 = \{t^{(n+m)} \mid \exists t_1^{(n)} \in T_1 \wedge \exists t_2^{(m)} \in T_2 \wedge (\forall a)_{a \in A(t1)-\{VA\}}[t.a = t_1.a] \wedge (\forall b)_{b \in A(t2)-\{VA\}}[t.b = t_2.b] \wedge t.VA = (t_1.VA \wedge t_2.VA) \wedge t.CR = Cr(t.VA)\}$$

join T1, T2 where (C) = select (product T1, T2) where (C)

즉 프로덕트에 의해 생성된 튜플 t의 VA 값은 각 튜플 t_1, t_2의 VA 값의 논리식으로 표현된다.

<표 10>은 금융상품에 대한 향후의 수요를 예측한 것으로 계획시나리오(제공원 속성)에 따라 분기별 예측잔액(데이터 속성)의 값에 대한 신뢰도가 결정된다. select Volume_Forecast where (금융상품 = CD(1년만기)) ∧ (예측잔액 >= 100)의 결과는 <표 11>과 같으며, <표 11>과 <표 9>의 프로덕트는 <표 12>와 같다.

<표 10> Volume_Forecast

금융상품	기준금리	스프레드	계획시나리오	예측잔액	VA	CR
실세예금	회사채유통수익률	1.5%	보수적	30	True	1
CD(1년만기)	CD유통수익률	2.0%	낙관적	110	True	1
CD(1년만기)	CD유통수익률	2.0%	보수적	100	True	1

<표 11> select Volume_Forecast where (금융상품=CD(1년만기)) ∧ (예측잔액>=100)

금융상품	기준금리	스프레드	계획시나리오	예측잔액	VA	CR
CD(1년만기)	CD유통수익률	2.0%	낙관적	110	낙관적	0.7
CD(1년만기)	CD유통수익률	2.0%	보수적	100	보수적	0.9

<표 12> Product 예

금융상품	기준금리	스프레드	계획 시나리오	예측잔액	예측항목	예측기관	예측이율	VA	CR
CD(1년만기)	CD유통 수익률	2.0%	낙관적	110	회사채유통 수익률	D연구소	12.0%	낙관적∧D연구소	0.595
CD(1년만기)	CD유통 수익률	2.0%	낙관적	110	CD유통 수익률	K연구원	11.8%	낙관적∧K연구원	0.56
CD(1년만기)	CD유통 수익률	2.0%	낙관적	110	CD유통 수익률	D연구소	12.5%	낙관적∧D연구소	0.595
CD(1년만기)	CD유통 수익률	2.0%	보수적	100	회사채유통 수익률	D연구소	12.0%	보수적∧D연구소	0.765
CD(1년만기)	CD유통 수익률	2.0%	보수적	100	CD유통 수익률	K연구원	11.8%	보수적∧K연구원	0.72
CD(1년만기)	CD유통 수익률	2.0%	보수적	100	CD유통 수익률	D연구소	12.5%	보수적∧D연구소	0.765

데이터 제공원의 신뢰도를 고려한 확장 관계형 모형에서의 투사, 합과 차 연산은 다음과 같이 정의된다.

정의 11 : 투사, 합과 차 연산

투사(projection), 합(union)과 차(set difference) 연산자는 다음과 같이 정의된다.

$$\text{project } a_1, ..., a_k\ T = \{t^{(k+2)} \mid \exists t' \in T \wedge A(t) = \{a_1, ..., a_k, VA, CR\} \wedge t.a_1 = t'.a_1 \wedge ... \wedge t.a_k = t'.a_k \wedge t.VA = t'.VA \wedge t.CR = Cr(t.VA)\}$$

$$\text{union } T1, T2 = \{t \mid \exists t \in T_1 \vee T_2\}$$

$$\text{difference } T1, T2 = \{t \mid \exists t \in T_1 \wedge \neg (\exists t \in T_2)\}$$

2.3 응용의 예

금융기관들은 매 분기 마다 외부 금융시장의 여건과 내부 마케팅 전략에 따라 여러 시나리오로 나누어 각 금융상품에 대한 수요를 예측하고, 외부 전문기관으로부터의 금리변화 예측을 참고하여 추정 대차대조표 및 손익계산서를 작성하여 봄으로써 전략수립에 기초적인 정보를 제공하고 있다. 예를 들어 앞에서의 <표 9>는 외부 경제연구기관으로부터 제공된 기준금리에 대한 예측자료를 나타내는 테이블로서 어떤 경제연구기관에서 예측하였는가에 따라 예측이율 값을 사용할 경우 그 값에 대한 신뢰도 정도가 차이나게 된다. <표 11>은 금융기관 내부에서의 금융상품 수요에 대한 예측자료를 나타내는 테이블로서 어떤 관점(보수적 혹은 낙관적 관점)에서 계획 시나리오를 작성하였는가에 따라 예측잔액 값을 사용할 경우 그 값에 대한 신뢰도 정도가 차이나게 된다.

추정이자수입은 예측잔액 * (예측이율 + 스프레드)로 정의되고 Volume_Forecast 테이블과 Rate_Forecast 테이블의 join 조건은 금융상품에 적용되는 기준금리가 예측항목과 동일하여야 하므로 앞에서 정의된 연산자를 사용하여 금융상품 수요와 금리변화에 대한 예측 데이터를 이용하여 이자수입을 추정하여 본다면, 다음과 같은 질의로 표현될 수 있다.

```
project 금융상품, 예측시나리오, 예측잔액, 예측기관, 예측이율, (예측잔액 * (예측이율 + 스프레드)) AS [추정이자]
(select (product Volume_Forecast, Rate_Forecast)
where 기준금리 = 예측항목)
```

이 질의의 결과는 <표 13>에 나타나 있는데 이는 계산된 추정이자 값들에 대한 신뢰 정도가 데이터 제공원, 즉 예측시나리오와 예측기관들의 신뢰도에 따라 서로 상이함을 보여주고 있다. 예를 들어 1년만기 CD의 경우 예측된 추정이자와 그 예측값에 대한 신뢰도는 (15.95, 0.595), (15.18, 0.56), (14.5, 0.765), (13.8, 0.72)로 상이하게 나타남을 보여주고 있다.

3. 결 론

기존의 많은 접근 방법들은 인공지능의 불확실성 연산(uncertainty calculus)에 기초하여 데이터의 불확실성 표현형태와 질의 연산자들을 정의한 확장된 관계형 데이터 모델(퍼지 데이터 모델, 확

〈표 13〉 Join 연산의 예

금융상품	예측 시나리오	예측잔액	예측기관	예측이율	추정이자	VA	CR
CD(1년만기)	낙관적	110	K연구원	11.8%	15.18	낙관적∧K연구원	0.56
CD(1년만기)	낙관적	110	D연구소	12.5%	15.95	낙관적∧D연구소	0.595
실세예금	보수적	30	D연구소	12%	4.05	보수적∧D연구소	0.765
실세예금	보수적	30	K연구원	11.1%	3.78	보수적∧K연구원	0.72
CD(1년만기)	보수적	100	K연구원	11.8%	13.8	보수적∧K연구원	0.72
CD(1년만기)	보수적	100	D연구소	12.5%	14.5	보수적∧D연구소	0.765

률 데이터 모델, 등)들을 제시하였다. 그러나, 이들의 접근 방법은 주로 데이터의 불확실성 표현형태에만 관심을 둔, 불확실성의 원인에 대한 고려나 응용분야와의 연계가 고려되지 않은 순수 이론적인 모델들이었다. 최근에 Sadri는 제공원의 신뢰도를 고려한 데이터 모델을 제시하였으나 제공원 벡터에 의해 제공원의 신뢰도만을 고려하였다. 이 두 가지 접근 방법은 순수 이론적인 모델이라는 측면에서 그 의미는 있지만, 현실적인 모델이 되기엔 부족하다.

본 논문에서는 이러한 기존의 데이터 모델들이 가지고 있는 여러 가지 단점을 지양하여 데이터의 제공원 신뢰도를 고려한 모델을 제시하였다. Sadri의 모델은 튜플 차원의 제공원 신뢰도만을 고려하여 각 개별의 속성 차원의 제공원 신뢰도는 나타낼 수 없다는 한계를 지니고 있으나, 본 논문에서 제안한 확장 관계형 모델은 필요한 속성들에 대해서만 신뢰도를 고려하여 Sadri의 모델보다 현실성 있는 모델이다. 이는 불확실성의 표현형태가 필요하지 않거나 혹은 표현이 불가능한 데이터에 대해 신뢰도만을 표현해야 할 경우에 사용될 수 있다.

이 연구의 결과는 아직 초기 단계에 불과한 것으로 다중 제공원의 데이터의 문제, 논리적 일치성(consistency)의 문제 등은 향후 연구 과제로 두고자 한다.

참 고 문 헌

[1] Barbara, D., Garcia-Molina, H., and Porter, D. "The Management of Probabilistic Data," IEEE Trans. on Knowledge and Data Engineering Vol.4, No.6, Oct. 1992, pp.487-502.

[2] Buckles, B. P. and F. E. Petry, "A Fuzzy Representation of Data for Relational Databases," Fuzzy Sets Syst. Vol.7, 1982, pp. 213-226.

[3] Cavallo, R. and M. Pittarelli, "The Theory of Probabilistic Databases," In Proc. 13th Int. Conf. on Very Large Databases(VLDB), 1987, pp.71-81.

[4] Lee, S.K. "Imprecise and Uncertain Information in Databases : An Evidential Approach," In Proc. 8th Int. Conf. Data Engineering, 1992, pp.614-621.

[5] Lee, S. K. "An Extended Relational Database Model For Uncertain and Imprecise Information," 18th Int. Conf. on Very Large Data Bases, 1992, pp.211-220.

[6] Morrissey, J. M. "Imprecise Information and Uncertainty in Information Systems," ACM Trans. Inf. Syst., Vol.8, No.2 April, 1990, pp. 159-180.

[7] Motro, A. "Imprecision and incompleteness in relational databases : survey," Information and software technology, Vol.32, No.9, Nov. 1990.

[8] Pittarelli, M., "Probabilistic Databases for Decision Analysis," International Journal of Intelligent Systems, Vol.5, 1990, pp.209-236.

[9] Prade, H. and C. Testemale, "Generalizing database relational algebra for the treatment of incomplete/uncertain information and vague queries," Information Sciences Vol.34, 1984, pp.115-143.

[10] Raju, K V S V N and A Majumdar, "A Fuzzy functional dependencies and lossless join decomposition of fuzzy relational database systems," ACM Trans. Database Syst. Vol.13, No.2, June 1988, pp.129-166.

[11] Sadri, Fereidoon "Reliability of Answers to Queries in Relational Databases," IEEE Trans. on Knowledge and Data Engineering, Vol.3, No.2, June 1991, pp.245-251.

[12] Zemankova, M. "Implementing Imprecision in Information Systems," Information Science, Vol.37, 1985, pp.107-141.

■ 저자소개

정 철 용

서울대학교 경제학과를 졸업하고 University of Washington에서 경영학 석사, University of Texas (at Austin)에서 경영정보학 박사학위를 취득하였다. 한국금융연구원(KIF) 부연구위원을 역임하였으며, 현재는 상명대학교 경영학과에 재직하고 있다. 주요 관심분야는 전자상거래, 소프트웨어 엔지니어링, 데이터웨어하우스, 데이터마이닝, 금융정보시스템 등이다.

이 석 균

서울대학교 경제학과를 졸업하고 University of Iowa에서 전산과학 석사와 박사학위를 취득하였다. 세종대학교 정보처리학과에 재직하였으며, 현재는 단국대학교 전산통계학과에 재직하고 있다. 1992년 IEEE 8th International Conference on Data Engineering에서 최우수 논문상을 수상한 바 있다. 주요 관심분야는 데이터 모델링, 데이터베이스에서 불완전 정보관리, 객체지향 데이터베이스 시스템, 데이터베이스 질의어, 시각질의어, 다중처리기에서의 실시간 스케줄링, 데이터웨어하우스 등이다.

서 용 무

서울대학교 사범대학 수학과, 한국과학원 전산학과를 졸업하고, 한국과학기술 연구소 전산센터에서 연구원으로 재직시 도미하여, University of Texas (at Austin)에서 전산학석사, 경영정보학박사를 취득한 후, 현재 고려대학교 경영대학에 재직하고 있다. 주요 관심분야는 web-based organizational computing, data warehouse, data mining 등이다.