퍼지 질의 처리를 위한 근접관계의 생성 방법
Generation Method of a Proximity Relation for Fuzzy Query Processing

김 창 석*, 김 대 수**, 이 상 조***
Chang Suk Kim*, Dae Su Kim**, Sang Jo Lee***

요 약

실용적인 퍼지 데이타베이스 시스템을 구축하는데 장애 요인중의 하나는 근접관계와 같은 의미 데이타를 습득하는 것이다. 근접관계란 어떤 도메인에서 데이타들간의 근사 혹은 유사한 정도를 정량적으로 표현한 것이다. 퍼지 데이타베이스 시스템은 부정확한 질의를 처리할때 이런 근접관계를 이용한다. 지금까지 근접도를 측정하는 체계적인 방법은 별로 알려진 것이 없고 대부분은 근접관계는 미리 주어진다는 가정하에 퍼지 데이타베이스를 연구하여 왔다. 본 논문에서는 퍼지 질의 처리에 필요한 근접관계 생성 방법을 제안한다. 제안된 방법은 퍼지 집합의 퍼지척도 측정 이론에 기반을 두었기 때문에 간단하고 체계적이며, 각 데이타에 특징값만 부여함으로써 해당 도메인내의 데이타들과의 근접도를 자동적으로 구할 수 있다. 특히 조정 변수를 이용하여 도메인내의 근접도 간격을 조절할 수 있어 실제 응용분야에 맞게 조절할 수가 있다. 퍼지 질의 처리를 위한 근접도 생성방법이 별로 발표되어 있지 않은 현 상황에서 본 논문에서 제시한 방법은 실용적인 퍼지 데이타베이스를 구현할때에 필요한 근접관계 관리 모듈에 사용될 수 있다.

ABSTRACT

One of the obstacles to building practical fuzzy database systems is to acquire semantic data such a proximity relation. The proximity relation is represented by the degree of 'closeness' or 'similarity' between data objects of a scalar domain. A fuzzy database system evaluates imprecise queries with the proximity relations. Only few of researchers have considered to systematically generate proximity relations up to now. Most of all the researchers assume that proximity relations are already given. In this paper, a generation method of proximity relation is proposed. The proposed method is simple and systematic since it is based on the well-known fuzzy set theory and it automatically generates degrees of proximity to assign feature values in each data objects. Also, the method is applicable to the real world applications with tuning parameter. The proposed generation method of proximity relations is essential to implement practical fuzzy relational database systems.

* 한국전자통신연구소 데이타베이스 연구실 선임연구원
** 한신대학교 전산학과 조교수
*** 경북대학교 컴퓨터공학과 교수

I. 서 론

Codd [1] 에 의해 제안된 관계 데이타 모델은 잘 정립된 이론적 배경과 데이타베이스 설계의 용이성 때문에 현재 대부분의 상용 데이타베이스 시스템들이 이 모델을 사용하고 있다. 이 관계 데이타 모델은 정확한(precise) 데이타와 질의들만을 처리할 수 있다. 그러나 자료 검색 분야나 의사 결정 등의 실세계 응용분야에서는 정확한 자료나 질의보다 부정확한(imprecise) 자료나 질의를 효과적으로 관리하고 처리해야 하는 경우가 많다. 컴퓨터의 사용을 기존에는 특수 분야의 전문가들만 사용하였으나 요즘에는 거의 모든 분야에서 비전문가들도 컴퓨터를 직접 사용하고 있다. 그러므로 이들 비전문가들이 원하는 자료를 검색하고자 할 때, 어떤 자료가 저장되어 있는지 혹은 자료 형태가 무엇인지 등을 자세히 숙지하고 있다고 볼 수 없다.

예를 들어, 사용자가 "관심분야가 데이타베이스인 회원을 검색하라"라는 정확한 질의를 했을 때 관심분야가 데이타베이스인 회원이 없으면 널(null) 값을 돌려 받는다. 이때 '데이타베이스'와 유사한 관심분야를 가진 회원을 검색하려면 위의 질의에서 데이타베이스 대신에 사용자가 컴퓨터에 저장되어 있음직한 항목을 대입하면서 질의를 반복해야 한다. 그러나 부정확한 질의와 자료를 취급하는 퍼지 데이타베이스에서는 '유사한(similar to)'이라는 기능이 있으므로 "데이타베이스와 유사한 관심분야를 가지는 회원을 검색하라"라는 질의를 할 수 있다. 이 질의를 수행했을 때 만약 '데이타베이스'를 관심분야로 가지는 회원이 없다면 '데이타베이스'와 유사한 의미를 갖는 '정보검색(information retrieval)' 이나 '자료처리(data processing)' 를 관심분야로 가지는 회원을 검색할 수 있다. 이 때 검색된 결과들은 0.9, 0.8 등의 정량적인 유사한 정도를 가지므로 순서화된 결과를 얻을 수 있다.

위와 같은 부정확한 질의를 처리하려면 순수 데이타가 저장된 관계(relation)이외에도 데이타들 사이의 유사한 정도를 의미적인 관점에서 정량적으로 표현된 근접관계(proximity relation)가 반드시 필요하다. 지금까지 퍼지 데이타베이스에 관한 많은 연구가 이루어져 왔으나 대부분 이론적인 면에 치우쳐 실제 대용량의 퍼지 데이타베이스를 구축하고 처리시에 필요한 근접관계의 생성 방법에 관한 연구는 별로 없는 실정이다. 그러나 최근 퍼지 데이타베이스가 실용화 단계에 접어듬에 따라 이런 실제적인 문제점에 대한 연구가 필요하게 되었다. Lee[9]에 의해 근접도 구하는 식이 제안된 적이 있으나 조정변수와 특징값 할당 방법 등에 차이가 있다.

본 논문에서는 퍼지 집합의 퍼지척도 측정 (measures of fuzziness)방법에 기반을 둔 근접관계 생성 방법을 제안한다. 여기서 제안한 근접관계 생성 방법은 각각의 데이타에 몇 개의 특징값을 할당함으로써 다른 모든 데이타들간의 유사한 정도를 간단하고 쉽게 구할 수 있다. 특히 조정 변수를 이용하여 도메인내의 근접도 간격을 조절할 수 있어 실제 응용분야에 맞게 조절할 수가 있다.

본 논문의 구성은 다음과 같다. 2장에서 퍼지 질의 처리의 기본이 되는 퍼지 관계 모델과 근접관계의 이론적인 정립에 대하여 서술한다. 3장에서는 근접도 생성방법을 제안하고 4장에서는 제안된 방법을 예를 통하여 설명한다. 마지막으로 5장에서 결론과 향후 연구계획을 기술한다.

II. 퍼지 관계 모델과 근접관계

2.1 관계 모델의 확장

퍼지 질의를 처리하는 퍼지 관계 데이타 모델에 관한 많은 연구가 이루어져 왔다[2, 3, 4, 5, 6]. 이것들을 분석해 보면 크게 CDFQ (Crisp Data and Fuzzy Query)와 FDFQ (Fuzzy Data and Fuzzy Query) 두 부류로 분류할 수 있다. CDFQ 는 기존 관계 데이타베이스의 데이타를 가지고 퍼지 질의를 처리하는 방법이고, FDFQ 는 퍼지 개념을 가지는 데이타를 데이타베이스에 저장하고 질의도 퍼지 개념을 포함하는 형태이다.

퍼지 관계 데이타베이스는 기존의 관계형 데이타베이스와 호환성이 유지되어야 실용성이 있다. FDFQ 방법은 비록 데이타 표현 능력이 뛰어나지만 기존 데이타베이스와 호환성이 없기 때문에 실용성이 없다. 그러나 CDFQ 방법은 데이타 표현 능력에서는 FDFQ 보다 떨어지지만 기존 관계 데이타베이스보다는 월등하고, 호환성도 가지므로 실용적인 측면을 고려하면 CDFQ 방법이 더 유리하다고 할 수 있다. 본 논문에서는 CDFQ 접근 방법을 선택한다.

가장 간단한 퍼지 데이타 모델을 정의하는 방법은 기존의 관계 모델에 소속척도 (membership values) 를 추가하는 것이다. 그러면 퍼지 데이타베이스 D_f 는 퍼지 관계들의 집합으로 정의할 수 있다.

$$D_f = \{R_1, R_2, R_3, \cdots, R_n\}$$

여기서 R_i 는 아래와 같은 소속함수로 정의된 소속척도를 가진다.

$$\mu_{R_i} : U_{i1} \times U_{i2} \times \cdots \times U_{in} \rightarrow [0, 1]$$

여기서 U_{ij}는 관계 R_i 의 j번째 애트리뷰트 도메인을 의미한다.

그림 1은 퍼지 관계 데이타베이스의 예를 보인 것이다. 그림 1 (a) 의 μ 는 해당 투플이 퍼지 관계에 소속될 정도를 나타낸다. 기존의 관계 투플은 μ 가 모두 1 이다. 그림 1 (b) 는 과목 (Subject) 도메인에 속하는 데이타간의 근접도를 나타낸 것이다. "데이타베이스 (DB) 와 유사한 과목을 찾아라"라는 질의가 주어졌을때 기존 관계 데이타베이스는 그림 1 (a) 에 '데이타베이스'라는 과목이 존재하지 않기 때문에 널(null) 값을 돌려주고 끝낸다. 그러나 퍼지 데이타베이스에서는 그림 1 (a) 에 원하는 값이 없으면 그림 1 (b) 의 근접관계에서 데이타베이스와 의미적으로 가장 가까운 것을 찾아서 보여준다. 즉, 데이타베이스와 의미적으로 가까운 DP(Data Processing, 0.9)와 IR(Information Retrieval, 0.8) 을 찾는다. 그러므로 퍼지 관계 데이타베이스는 근접관계가 반드시 필요하고 다음과 같은 일반적인 관계 데이타베이스와 구별되는 특징을 가진다.

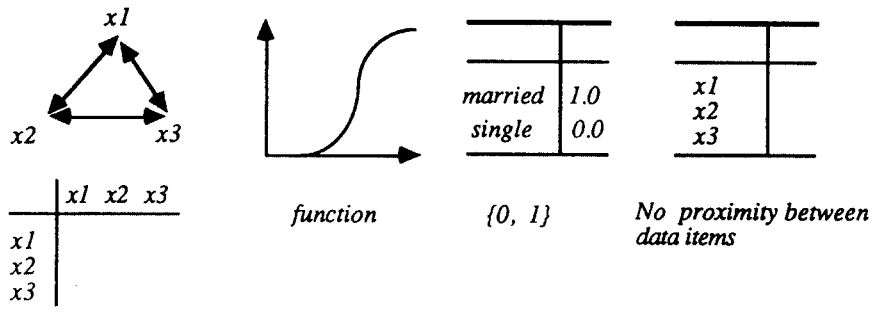
Subject	Teacher	μ
Data Processing	David	1.0
Artificial Intelligence	Larry	0.7
Information Retrieval	Jim	0.9
Operating System	Peterson	0.3

(a) 퍼지 관계의 예

	DB	AI	IR	OS	DP
DB	1.0	0.7	0.8	0.0	0.9
AI	0.7	1.0	0.8	0.0	0.3
IR	0.8	0.8	1.0	0.0	0.6
OS	0.0	0.0	0.0	1.0	0.0
DP	0.9	0.3	0.6	0.0	1.0

(b) 과목(Subject) 도메인의 근접관계

그림 1. 퍼지 관계 데이타베이스의 예



(a) 스칼라 형태 (b) 수치 형태 (c) 단위 간격 형태 (d) 일반 관계 형태

그림 2. 도메인의 형태

a) 원하는 자료가 존재하지 않으면 의미상 가장 가까운 것을 검색한다.
b) 결과로 출력되는 자료가 만족하는 정도를 수치로 가질 수 있으며, 순서화된 결과를 얻을 수 있다.

퍼지 관계 데이타베이스에서 도메인의 형태는 아래와 같이 세가지 형태로 나눌수 있다.

1) 스칼라 형태 : 각각의 데이타가 서로 연관관계가 없이 별개로 존재한다.
2) 수치 형태 : 데이타가 일차원적인 의미에서 서로 연결되어 있다.
3) 단위 간격 형태 : 데이타가 어떤 관점의 의미를 전부 혹은 전혀 만족하지 않는 형태이다.

그림 2는 도메인의 형태에 따른 소속척도의 표현 방법을 나타낸 것이다. 그림 2 (a) 는 '색깔 (color)' 이나 '일의 기능 (job function)' 또는 '과목 (subject)' 같은 스칼라 형태의 도메인을 나타낸 것이다. 이 형태에 대한 질의를 처리하기 위해서는 데이타간의 유사정도를 저장해 둔 근접관계가 필요하다. 그림 2 (b) 는 '나이'나 '경력' 혹은 '봉급' 등 데이타 값이 주어진 수치에 따라 순서적으로 표현되는 수치 형태를 나타낸 것이다. 그림 2 (c) 는 '성별' 이나 '결혼 유무' 등 소속정도 값이 1 또는 0 으로 나타낼 수 있는 단위 간격 형태이다. 그림 2 (d) 는 일반적인 관계 데이타베이스의 데이타 형태로 데이타간의 어떠한 연관관계도 가지지 않는다.

2.3 근접관계

본 논문의 CDFQ 모델에서 데이타베이스는 CDB (Crisp DataBase)와 SDB (Semantic DataBase) 로 나누어진다. CDB는 기존의 데이타베이스를 의미하고 SDB 는 근접관계나 데이타간의 관계를 함수로 표현된 데이타 등을 저장한다. 본 논문에서는 연구범위를 좁혀 근접관계에 관한 것만 다룬다. CDB 의 변경없이 주관적인 관점에 따라 SDB 를 얼마든지 다르게 정의할 수 있다. 또한 퍼지 질의를 효율적으로 처리하기 위해서 SDB 를 다른 형태로 컴퓨터내에 표현할 수도 있다.

Shenoi 와 Melton[8]이 제안한 근접관계는 Buckles 와 Petry[7] 가 제안한 유사관계 (similarity relation) 를 포함하므로 여기서는 근접관계를 이용한다. 근접관계를 정의하면 다음과 같다.

[정의 1] 근접관계 (proximity relation) 는 $s_j : D_j \times D_j \rightarrow [0,1]$와 같은 매핑관계를 가진다. 이때 $x, y \in D_j$이다.
(i) $s_j(x, x) = 1$ (반사성),
(ii) $s_j(x, y) = s_j(y, x)$ (대칭성).

Ⅲ. 근접도의 생성

근접도 (proximity degree) 란 주어진 관점에 대한 데이타 개체간의 유사한 정도를 정량적으로 나타낸 것이다. 어떤 도메인에 속하는 여러 데이타 개체간의 근접도를 행렬 (matrix) 로 표현한 것을 근접관계 (proximity relation) 라 한다. 근접관계는 대칭성 (symmetricity), 반사성 (reflecivity) 을 만족한다[8, 9].

본 장에서는 크리스프(crisp) 하고 원자적인(atomic) 성질을 갖는 데이타가 의미를 가지도록 어떤 특징 값을 부여하여 퍼지화 (fuzzification) 하고, 퍼지 집합의 퍼지척도 측정(measures of fuzziness) 기법을 응용한 데이타간의 근접도의 생성 식을 제안한다.

3.1 퍼지척도 측정 방법

근접도를 구하는 식을 유도하기 전에 퍼지집합의 퍼지척도(불확실한 정도) 구하는 식을 살펴본다. 왜냐하면 퍼지집합의 퍼지척도를 구하는 식으로 부터 근접도 구하는 식을 유도할 수 있기 때문이다. 퍼지척도를 측정하는 방법으로는 샤논 (shannon)의 엔트로피 (entropy)에 기초를 둔 방법과 거리측정 (metric distance)에 기초한 방법이 있다. 각각의 방법에 대하여 간략히 정리하고 의미를 서술한다[10].

· 샤논의 엔트로피에 기초한 방법

샤논의 엔트로피는 불확실성이나 정보의 양을 측정하는 척도로 많이 이용되고 있으며, 정보이론 분야의 가장 기본이 되는 이론이기도 하다. 이것은 확률이론에 기반을 두고 있는데 다음과 같은 함수로 표현된다.

$$H(p(x) \mid x \in X) = -\sum_{x \in X} p(x) \log p(x)$$

여기서 $(P(x) \mid x \in X)$ 는 집합 X내의 확률분포를 나타내고, $P(x)$의 합은 반드시 1 이다.

$$\sum_{x \in X} P(x) = 1$$

퍼지집합에서는 위의 확률의 합 대신 전체집합 X 내의 소속척도 $\mu_A(x)$ 에 의해 표현되는데 그것의 합이 반드시 1 로 제한되지 않는다.

$$\sum_{x \in X} \mu(x) = 1 \qquad \text{(불필요한 조건임)}$$

이상을 바탕으로 퍼지집합 A의 퍼지정도를 측정하는 척도 f(A)를 정의하면 다음과 같다[10].

$$f(A) = -\sum_{x \in X} (\mu_A(x) \log_2 \mu_A(x) + [1-\mu_A(x)] \log_2 [1-\mu_A(x)])$$

f(A) 를 정규화 (normalization) 하면 정규화된 척도 $f(A)$ 를 얻을 수 있다.

$$f(A) = \frac{f(A)}{|X|}$$

여기서 |X| 는 전체집합 X 의 원소 갯수이다. 정규화된 척도는 다음과 같은 범위를 가진다.

$$0 \le f(A) \le 1,$$

즉, x 의 소속척도 값이 0.5 에 가까운것이 많을수록 불확실성이 높다.

· 거리측정에 기초한 방법

또 다른 퍼지척도를 구하는 방법으로 거리측정의 개념에 바탕을 두는 것이 있다. 두 집합사이에 거리를 측정하는 척도로 해밍거리(Hamming distance)와 유크리드거리(Euclidean distance)가 있다. 퍼지집합 A의 거리를 측정하기 위해서는 먼저 집합 A에 대응되는 보통집합 C를 다음과 같이 정의해야 한다.

$$\mu_C(x) = 0 \qquad \text{if } \mu_A(x) \le 0.5$$
$$\mu_C(x) = 1 \qquad \text{if } \mu_A(x) > 0.5$$

위와 같이 퍼지집합 A에 대응하는 보통집합 C 가 정의되면 퍼지집합 A와 보통집합 C 사이에 해밍거리와 유크리드거리를 적용한 것이 퍼지척도가 된다. 이때 해밍거리와 유크리드거리를 적용한 퍼지척도는 다음과 같다.

$$f(A) = \sum_{x \in X} |\mu_A(x) - \mu_C(x)| \qquad \text{(해밍거리를 적용한 퍼지척도)}$$

$$f(A) = (\sum_{x \in X} |\mu_A(x) - \mu_C(x)|^2)^{1/2} \qquad \text{(유크리드거리를 적용한 퍼지척도)}$$

$\mu_A(x)$의 값이 0.5에 가까울수록 f(A)의 값이 커져서 퍼지척도가 크다.

위와같이 퍼지집합의 퍼지한 정도를 구하는 방법을 살펴보았다. 두 방법 모두 퍼지집합을 구성하는 원소들의 소속척도 값이 0.5에 가까운 것이 많을수록 불확실함이 큰 것을 알 수 있다. 바꾸어 말하면 가장 불확실성이 큰 소속척도 값인 0.5를 기준으로 주어진 퍼지집합의 원소값들의 거리를 구한다. 우리는 위와 같은 원리를 데이타간의 근접도를 구하는데 응용한다.

3.2 근접도 생성 식의 도출

크리스프(crisp) 하고 원자적인(atomic) 성질을 갖는 관계형 데이타베이스에서 데이타들간에 의미적인 유사한 정도를 측정한다는 것이 불가능하며, 단지 그 값의 ASCII 코드를 비교하여 같거나 혹은 대소를 판정할 수 있다. 데이타간의 근접도를 구하기 위해서는 먼저 크리스프한 데이타를 퍼지하게 변환하여야 그 데이타가 문자적인 의미외에 내용적인 의미를 가진다. 데이타가 내용적인 의미를 갖는다는 것은 다른 데이타와 구별되는 특징을 갖는다고도 할 수 있다. 근접도 생성 식을 도출하는데 필요한 두 가지 용어를 정의한다.

[**정의 2**] 한 도메인은 어떤 주어진 관점에 대한 서브 속성(sub-property)으로 분해할 수 있다. 이 서브 속성을 범주(category)라 하고 다음과 같이 표현한다.

x_i (여기서, $1 \le i \le l$, l은 유한 정수).

[**정의 3**] 각 범주에 할당하는 값을 특징값(feature value)라 하고 다음과 같이 표현한다.

$\mu(x_i) \in [0, 1]$.

데이타에 내용적인 의미를 부여하기 위해서는 먼저 그 데이타가 속한 도메인을 어떤 관점에서 세부 범주(category) 로 나누어야 한다. 각 데이타는 이 범주에 속하는 특징값을 가지며, 이 특징값은 다른 데이타와 구별되는 기준이다. 이것은 마치 어떤 퍼지 집합이 유한개의 개체들과 그들의 소속척도 집합으로 구성되어 있는 것과 유사하다. 이 특징값은 [0, 1]의 값을 가지며 도메인에 속한 임의의 한 데이타는 아래와 같이 표현된다.

$$\theta = \sum_{i=1}^{l} \mu(x_i)/x_i$$

여기서 x_i 는 범주의 이름이고, $\mu(x_i)$는 특징값을 의미한다.

예를들어 '야구팀포지션'이라는 도메인 (Θ) 이 존재하고, 그 도메인의 데이타 (θ_i) 로 '투수, 포수, 일루수, 이루수, 삼루수, 유격수, 우익수, 중견수, 좌익수'가 존재한다고 가정하자. 도메인 Θ는 그림 3과 같이 5개의 범주로 구성하면 각 데이타는 5개의 특징값을 가진다.

Θ = {투수, 포수, 일루수, 이루수, 삼루수, 유격수, 우익수, 중견수, 좌익수}

x_i = {볼콘트롤, 볼받는 능력, 순발력, 볼던지는 힘, 주력}

θ_1(투수) = $\mu_1/x_1 + \mu_2/x_2 + \mu_3/x_3 + \mu_4/x_4 + \mu_5/x_5$

이때 특징값을 할당하는 방법에는 퍼지집합에서 개체들의 소속척도값을 할당하는 방법인 직접등급 (direect rating), 역등급 (reverse rating), 투표(polling), 집합값 통계 (set valued statistics) 와 같은 방법을 이용한다 [9, 11]. 특징값을 할당하는 방법도 중요한 연구 분야이나 본 논문의 논제를 벗어나므로 여기서는 더 이상 언급하지 않는다.

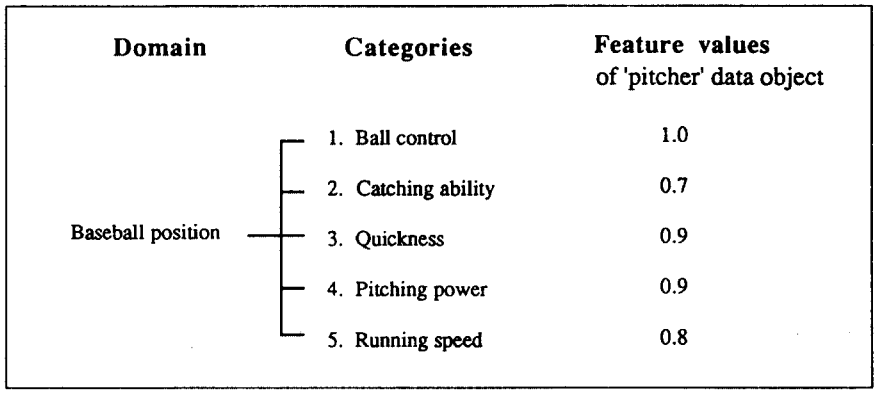


그림 3. 도메인 '야구팀 포지션' 범주와 특징값 (투수)의 예

근접도는 두 데이타간에 나타나는 유사도이기 때문에 Θ를 2차원 관계인 $\Theta \times \Theta$으로 확장한다.

$$p : \Theta \times \Theta \rightarrow [0, 1] \tag{1}$$

두 데이타간의 근접도가 크다는 것은 두 데이타안에 존재하는 특징값들간의 거리차가 적다고도 할 수 있다. 이것을 식으로 표현하면 다음과 같다.

$$\sum_{x \in X} [\mu_{\theta i}(x) - \mu_{\theta j}(x)] \geq \sum_{x \in X} [\mu_{\theta i}(x) - \mu_{\theta k}(x)] \tag{2}$$

if and only if $p(\theta_i, \theta_j) \leq p(\theta_i, \theta_k)$

그래서 두 특징점간의 거리를 계산하여 근접도를 구하는 식을 유도해 본다. 데이타 θ_i 와 데이타 θ_j의 한 개의 특징값간의 거리를 아래와 같이 표현한다.

$$| \mu_{\theta i}(x) - \mu_{\theta j}(x) | = \delta_{\theta i, j}(x) \tag{3}$$

이것을 그 데이타내에 속하는 모든 범주의 특징값 대하여 구하면 다음과 같다.

$$dis(\theta_i, \theta_j) = \sum \delta_{\theta_{i,j}}(x) \text{ (here, } x \in X) \tag{4}$$

식 (4)를 좀 더 일반화하기 위하여 특징값들간의 거리차가 가장 큰 임의의 두 데이타 (θ_H, θ_L)를 도입한다. θ_H는 특징값이 모두 1 인 데이타이고, θ_L는 특징값이 모두 0인 데이타이다. (θ_H, θ_L)과의 거리에서 주어진 두 데이타간의 거리를 빼면 두 데이타간의 근접도가 되며, 아래와 같이 표현된다.

$$P(\theta_i, \theta_j) = dis(\theta_H, \theta_L) - dis(\theta_i, \theta_j) \tag{5}$$

여기서 $P(\theta_i, \theta_j)$의 범위는 다음과 같다.

$$0 \le P(\theta_i, \theta_j) \le dis(\theta_H, \theta_L) \tag{6}$$

위의 식 (5)와 (6)의 범위가 [0, 1]이 되도록 정규화하여 근접도 p를 구하면 식 (7)과 (8) 같이 표현된다.

$$p(\theta_i, \theta_j) = 1 - \frac{dis(\theta_i, \theta_j)}{dis(\theta_H, \theta_L)} \tag{7}$$

$$0 \le p(\theta_i, \theta_j) \le 1 \tag{8}$$

$dis(\theta_H, \theta_L)$은 |X|와 같으므로 식 (4)와 |X|를 식 (7)에 대입하면 다음과 같다.

$$p(\theta_i, \theta_j) = 1 - \frac{\sum_{x \in X} \delta_{\theta_{ij}}(x)}{|X|} \tag{9}$$

식 (3)을 식 (9)에 대입하면 식 (10)를 얻을 수 있다.

$$p(\theta_i, \theta_j) = 1 - \frac{\sum_{x \in X} |\mu_{\theta i}(x) - \mu_{\theta j}(x)|}{|X|} \tag{10}$$

식 (10)을 좀 더 일반적인 식으로 표현하면 다음과 같이 표현할 수 있다.

$$p_w(\theta_i, \theta_j) = 1 - \frac{[\sum_{x \in X} |\mu_{\theta i}(x) - \mu_{\theta j}(x)|^w]^{1/w}}{|X|} \tag{11}$$

여기서 $w \in [1, \infty]$를 조정변수(tuning parameter)라 한다. μ가 증가할수록 근접도의 격차는 줄어들고, w가 감소할수록 근접도의 격차는 늘어난다. 응용분야에 따라 w를 조정하여 적용할 수가 있다. 실제 응용분야에서는 소속척도 μ값들 끼리 여러번 집계하거나 α-절단을 사용하여 결과를 도출하는 경우가 많다. T-norm, T-conorm, averaging operator 등 수 많은 퍼지 집합 집계 연산자들이 이 경우에 사용된다. w 값의 간격 조정 여하에 따라 근접도가 해당 응용분야에서 반영되는 비중이 달라질 수 있어 중요한 매개변수 역할을 한다.

Ⅳ. 근접관계 생성 예

근접도를 구하는 절차와 실제 예를 들어 근접도를 구해본다. 또한 조정변수를 변화 시켜가면서 근접도를 구하여 보고 근접도 간격 변화를 살펴본다. 먼저, 근접도를 구하는 절차를 도시하면 그림 4와 같다. 임의의 도메인 Θ가 주어지고 이것의 구성요소가 θ_i $(1 \le i \le n)$일때 근접도를 구하는 절차는 먼저 도메인 Θ를 몇개의 범주로 나누어야 한다. 각 데이타 θ_i를 범주에 따라 특징값을 할당한다. 그리고 식(11)를 사용하여 데이타 (θ_i, θ_j)의 근접도를 구한다.

[예 1] 세 데이타 θ_1(투수)와 θ_2(유격수) 및 θ_3(포수)가 다음과 같은 특징값을 가질 때 투수와 유격수, 투수와 포수 사이의 유사한 정도, 즉 근접도를 구하라. 이때 조정변수 $w=1$ 라고 가정한다.

$\theta_1 = \{\mu_1 = 1.0, \mu_2 = 0.7, \mu_3 = 0.9, \mu_4 = 0.9, \mu_5 = 0.8\}$,
$\theta_2 = \{\mu_1 = 0.8, \mu_2 = 1.0, \mu_3 = 1.0, \mu_4 = 0.9, \mu_5 = 0.9\}$,
$\theta_3 = \{\mu_1 = 0.6, \mu_2 = 1.0, \mu_3 = 0.4, \mu_4 = 1.0, \mu_5 = 0.5\}$.

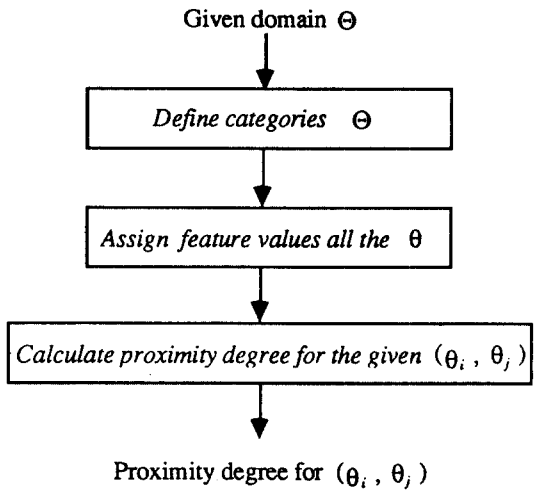


그림 4. 근접관계 생성 절차

[풀이] 식(11)을 사용하여 데이타 θ_1과 θ_2, θ_1과 θ_3의 근접도를 각각 구하면 다음과 같다.

$$p(\theta_1, \theta_2) = 1 - \frac{\sum_{x \in X} |\mu_{\theta 1}(x) - \mu_{\theta 2}(x)|}{|X|}$$

$$= 1 - \frac{|1.0-0.8| + |0.7-1.0| + |0.9-1.0| + |0.9-0.9| + |0.8-0.9|}{5}$$

$$= 0.86$$

$$p(\theta_1, \theta_3) = 1 - \frac{\sum_{x \in X} |\mu_{\theta 1}(x) - \mu_{\theta 3}(x)|}{|X|}$$

$$= 1 - \frac{|1.0-0.6| + |0.7-1.0| + |0.9-0.4| + |0.9-1.0| + |0.8-0.5|}{5}$$

$$= 0.68$$

$p(\theta_1, \theta_2) = 0.86$이므로 데이타 θ_1(투수)와 θ_2(유격수)와는 비교적 유사정도가 높다고 할 수 있다. $p(\theta_1, \theta_2) = 0.68$이므로 데이타 θ_1(투수)와 θ_2(포수)와는 유사정도가 낮다고 할 수 있다.

[예 2] 조정변수 w의 효과를 보일려고 한다. $w=2$, $w=3$ 그리고 $w=4$ 라고 가정하고 [예 1] 을 수행하라.

[풀이] 식(11)을 사용하여 $w=2$, $w=3$, $w=4$ 일때 데이타 θ_1과 θ_2, θ_1과 θ_3의 근접도를 각각 구하면 다음과 같다.

i) case $w=2$:

$p(\theta_1, \theta_2) = 0.92$ $\quad$ $p(\theta_1, \theta_3) = 0.85$

ii) case $w=3$:

$p(\theta_1, \theta_2)=0.93$ $p(\theta_1, \theta_3)=0.88$

iii) case $w=4$:

$p(\theta_1, \theta_2)=0.94$ $p(\theta_1, \theta_3)=0.89$

조정변수 w가 증가할수록 근접도 $p(\theta_1, \theta_2)$과 $p(\theta_1, \theta_3)$의 간격은 줄어듬을 알 수 있다.

본 장에서는 퍼지척도 구하는 식을 기반으로 근접도 구하는 식을 제안했고, 예를 들어 설명했다. 이 제안된식은 크리스프한 데이타를 퍼지하게 처리하기 위해 범주와 특징값 개념을 새롭게 도입하여 그 특징값 사이의 거리를 측정하여 두 데이타간의 근접도, 즉 유사한 정도를 측정했다. 또한 조정변수 w를 사용하여 응용분야에 따라 근접도 간격을 조정할 수 있게하여 일반화된 특징도 가짐을 보았다.

V. 결 론

본 논문에서는 퍼지 질의 처리에 필요한 근접도를 생성하는 방법을 제안하고 예를들어 설명하였다. 퍼지척도 측정 방법을 이용한 근접관계 생성 식은 각 데이타에 대한 특징값만 부여함으로서 해당 도메인내의 다른 데이타들과의 근접도를 자동적으로 구할 수 있음을 보았다. 또한 이 방법은 적용되는 응용분야에 따라 근접도 간격을 조절할 수 있어 일반성을 가짐을 알 수 있었다. 지금까지 근접도를 측정하는 뚜렷한 방법이 제시되어 있지 않고 단지 근접관계가 미리 주어진다는 가정하에서 퍼지 데이타베이스가 연구되어 왔다. 그래서 대용량의 퍼지 데이타베이스를 실용화하는데 문제가 되어왔다. 이런 상황에서 본 논문에서 제시한 방법은 실용적인 대용량의 퍼지 데이타베이스를 구현하는 측면에서 의의가 있다고 본다.

그러나 위에서 제시한 근접도 생성 방법은 거리 측정의 기본 요소가 되는 특징값을 할당하는 체계적인 방법이 정립하지 못한 문제점도 가지고 있다. 이것은 향후 연구 과제로 남기기로 한다. 앞으로의 계획은 위의 방법을 채용한 근접관계 관리 모듈을 구현하여 퍼지 데이타베이스 관리 시스템의 한 부분으로 사용할 예정이다.

참 고 문 헌

1. E. Codd, "A relational model for large shared data banks," Comm. of the ACM, Vol. 13, No. 6, pp. 377-387, 1970.
2. S. C. Park, C. S. Kim and D. S. Kim, "Fuzzy querying in relational databases," Fifth IFSA World Congress, pp. 533-536, 1993.
3. D. Lee and M. Kim, "A fuzzy relational data model and extended semantics of relational operations," Proceedings InfoScience 93, pp. 275-281, 1993.
4. M. Umano and S. Miyamoto, "Recent development of fuzzy database systems and applications," Fifth IFSA World Congress, pp. 537-540, 1993.
5. T. Ichikawa and M. Hirakawa, "ARES : A relational database with the capability of performing flexible interpretation of queries," IEEE Trans. on Software Engineering, Vol. SE-12, No. 5, pp. 624-634, 1986.
6. A. Motro, "VAGUE : A user interface to relational databases that permits vague queries," ACM Trans. on Office Information Systems, Vol. 6, No. 3, pp. 187-214, 1988.
7. B. Buckles and F. Petry, "A fuzzy representation of data for relational databases," Fuzzy Sets and Systems, Vol. 7, pp. 213-226, 1982.
8. S. Shenoi and A. Melton, "Proximity relations in the fuzzy relational database model," Fuzzy Sets and Systems, Vol. 7, pp. 285-296, 1989.

9. D. Lee and M. Kim, "Elicitation of semantic knowledge for fuzzy database systems," Conf. on Korea Information and Science Scieoty, pp. 113-116, Oct. 1993.
10. G. Klir and T. Folger, Fuzzy Sets, Uncertainty, and Information, Prentice-Hall International Editions, 1988.
11. I. Turksen, "Measurement of membership functions and their acquisition," Fuzzy Sets and Systems, Vol. 40, pp. 5-38, Oct. 1991.