Vague형 퍼지 데이터베이스에서의 튜플 소속척도와 질의에 대한 엔트로피 연구

(Tuple Membership Values & Entropy for a Vague Model of the Fuzzy Databases)

박 순 철*
(Soon C. Park)

요 약 이 논문에서는 vague형 퍼지 데이터베이스에서 질의에 대한 튜플의 소속척도 연산을 분석하여 사용자가 요구하는 결과를 보다 효과적으로 얻을 수 있는 방법을 제안한다. 또한 정보 이론을 이용하여 사용자의 질의에 대한 데이터베이스의 엔트로피를 계산함으로써 데이터베이스의 특성을 분석하고, 아울러 엔트로피를 제어할 수 있는 알고리즘을 제안한다.

Abstract In this paper, the methods which calculate the tuple membership values in a vague model of the fuzzy databases are analyzed. A method among them is proposed to offer the effective solutions to the users. Also an information theory is studied to calculate the entropy of the results of a fuzzy query and an algorithm is proposed to control the size of the entropy.

1. 서 론

최근에 정보통신 종사자뿐만 아니라 일반 사용자들도 WWW(World Wide Web)이나 디지털 도서관을 통하여 자료수집이나 검색을 하는 것이 일반화되었다. 따라서 데이터베이스에 저장된 정보 요소가 복잡해지고 방대해 지더라도 그러한 내용을 제대로 파악하지 못하고 있는 일반 사용자들도 데이터베이스를 지적 자원으로 활용할 수 있는 수단이 절실히 필요하다. 이러한 필요성을 충족시키기 위하여 정확한 키값을 요구하는 기존의 질의어 형식을 탈피한 퍼지 데이터베이스 시스템의 연구가 필수적이다.
퍼지 데이터베이스 시스템의 연구는 지난 십수년 전부터 관계형 데이터베이스를 기본으로 한 퍼지 관계형 데이터베이스 시스템이 Buckles와 Petry, Shenoi, Melton과 Fan, Motro 등에 의해서 활발히 진행되고 있다[3, 11, 13, 14].

퍼지 데이터베이스 시스템에는 크게 FRDB(Fuzzy Relational Database) 모델과 vague 모델이 있다[2, 11, 17]. FRDB 모델은 불확실한 데이터베이스와 질의어를 모두 수용하며 각 데이터의 불확실성을 처리할 수 있다. 반면에 vague 모델은 기존의 관계형 데이터베이스 시스템의 정규화 된 데이터를 그대로 사용하여 질의어 형태를 변경시켜 사용자의 애매한 질의를 처리할 수 있도록 지원한다. 이 논문에서는 기존의 정규화 된 관계형 데이터베이스를[5, 6, 10] 큰 수정 없이 그대로 사용할 수 있는 vague 모델을 이용한다.

이 논문의 목적은 첫째, 사용자의 질의에 효과적인 결과를 제공할 수 있는 튜플 소속척도의 계산하는 알고리즘을 분석하며, 둘째, 퍼지 질의어를 통해서 얻어진 결과의 정보 엔트로피를 제어하여 사용자가 원하는 정보를 좀더 정확하게 제공하는 것이다.

이 논문의 구성은 다음 2장에서는 엔트로피의 정의와 질의에 대해 효과적으로 결과 값을 제공할 수 있는 알고리즘에 대한 이론적 배경을 소개하고, 3장에서는 2장의 이론을 적용하여 얻어진 결과를 소개하겠다. 마지막으로 4장에서는 결론과 향후연구 방향에 대하여 서술하겠다.

† 이 논문은 1997년도 한국과학재단의 핵심전문연구지원사업 (과제번호:971-0902-0202) 지원에 의한 논문임.

* 전북대학교 정보통신공학과 조교수 및 전북대학교 부속정보통신연구소 연구원

2. 이론적 배경

2.1 근사관계

퍼지이론에서는 '젊다,' '어리다,' '늙다' 라는 판단 기준의 애매함을 소속척도로 나타낸다[9, 15, 16]. 각 표현 사이의 소속척도는 유사관계(Similarity Relation) 또는 근사관계(Proximity Relation)로 표현할 수 있다[4, 14]. 그러나 유사관계는 소속관계가 있는 데이터간의 이행관계(transitivity)를 너무 강조하여 사용자가 나타내고 싶은 소속 척도를 경우에 따라 불가능하게 한다. 그러므로, 이 논문에서는 유사관계의 단점을 보완한 근사관계를 이용한다.

관계형 데이터베이스에서 ti를 임의의 튜플이라고 한다면 ti는 (di1, di2, · · ·, dim)으로 표현할 수 있다. 이 때 튜플의 구성요소인 dij는 영역집합(domain set) Dj의 영역값이다. 즉, dij ∈ Dj 이며, 근사관계의 소속척도 값은 다음과 같다.

$$s_i : D_j \times D_j \rightarrow [0,1] \tag{1}$$

식 (1)에서 [0, 1]은 0과 1사이의 실수를 말한다.

또한, x, y가 Dj의 원소라면(즉, x, y ∈ Dj), 근사관계는 다음과 같다.

$$\begin{aligned} s_j(x,x) &= 1 \\ s_j(x,y) &= s_j(x,y) \end{aligned} \tag{2}$$

이러한 근사관계는 유사관계의 Super Set이며 유사관계에 있는 모든 원소 값을 근사관계가 포함한다.

2.2 튜플 소속척도

사용자의 질의에 대한 튜플 소속척도는 각 영역 소속척도의 조합 연산으로 이루어진다[1, 4, 12]. 연산된 튜플 소속척도는 사용자에게 효과적인 결과를 제공하기 위해서 다음과 같은 세 가지 조건을 만족하여야 한다.

① 튜플의 영역 소속척도 값 중 하나가 0에 가까우면 전체 값도 0으로 수렴해야 한다. 예를 들면 한 결혼 적령기에 있는 남성이 원하는 질의어 "고등학교 정도의 학력을 갖춘 26세정도의 미혼 여성을 찾아라"에서 나이가 아주 많거나 적어서 질의에 대한 나이의 소속척도가 0에 가깝다면 질의에 대한 튜플 소속척도는 0에 수렴해야 한다.

② 특정 attribute에 weight를 줄 수 있어야 한다. ①항에서 제시한 질의에서 한 남성이 나이보다는 학력이 높은 것을 더 우선적으로 생각한다고 가정할 때 그것을 연산할 수 있어야 한다.

③ 보상연산이 가능해야 한다. ①항에서의 질의어에서 나이는 같은 여성이지만 학력이 사용자가 원하는 것과 유사할 경우 그 소속척도의 값은 그 반대편의 사람보다 더 높아야 한다.

퍼지 데이터베이스에서 질의어가 d1, d2, · · ·, dm-1, dm영역의 값의 조건을 포함하고 있고 각 영역에 대한 소속척도의 값이 각각 μ1, μ2, · · ·, μm-1, μm이라고 할 때 소속척도는 다음과 같은 식들로 구하여 질 수 있다 [1, 4, 12].

$$F_1 = \min(\mu_1, \mu_2, \cdots, \mu_{m-1}, \mu_m) \tag{3}$$

$$F_2 = (\mu_1 + \mu_2 + \cdots + \mu_{m-1} + \mu_m)/m \tag{4}$$

$$F_3 = \sqrt[m]{\mu_1 \times \mu_2 \times \cdots \times \mu_{m-1} \times \mu_m} \tag{5}$$

$$F_4 = \sqrt[m]{\mu_1^{\alpha_1} \times \mu_2^{\alpha_2} \times \cdots \times \mu_{m-1}^{\alpha_{m-1}} \times \mu_m^{\alpha_m}} \tag{6}$$

식 (3)의 경우 영역의 소속척도 중 가장 작은 값을 제공한다. 이러한 경우 논리적으로 간단하다는 장점이 있지만 위에서 정의한 ②항과 ③항을 만족시키지 못한다.

식 (4)의 경우는 ③항만을 만족시킨다.

식 (5)는 ①과 ③을 만족시키기는 하지만 어떤 애트리뷰트를 강조하고 싶을 때 ②항을 만족시킬 수 없다.

식 (3)~(6)중 식 (6)만이 ①, ②, ③을 모두 만족시킨다. 그러므로 이 논문에서는 식 (6)을 질의에 대한 튜플 소속척도를 구하는 알고리즘으로 강조한다.

2.3 퍼지 질의에 대한 데이터베이스의 엔트로피

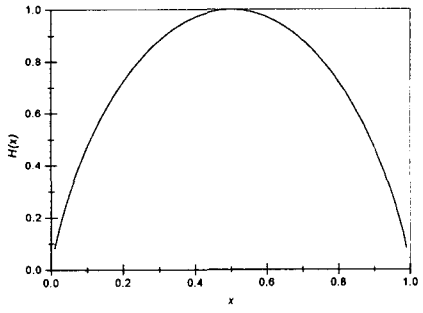
샤론의 엔트로피는 식 (7)과 같이 정의된다[7]. 여기서 x는 확률을 의미하지만 우리는 이것을 fuzziness로 대신할 수 있다[4].

$$H(x) = -x \cdot \log_2(x) - (1-x) \cdot \log_2(1-x) \tag{7}$$

<그림 1>은 엔트로피 함수의 개략적인 모양을 나타낸 것이다. x값이 0과 1일 때 엔트로피 H(x)가 0값임을 보이며 x값이 0.5일 때 최대 값인 1값을 보인다.

튜플 n개를 포함한 릴레이션 r에 관해서 질의 Q에 대한 데이터베이스의 엔트로피는 식 (8)과 같이 정의될 수 있다[4].

$$H_S(r|Q) = \sum_{i=1}^{n} -\mu_Q(t_i)\cdot\log_2(\mu_Q(t_i)) - (1-\mu_Q(t_i))\cdot\log_2(1-\mu_Q(t_i)) \quad (8)$$



<그림 1> 엔트로피 함수 $H(x)$

식 (8)에서 모든 i에 대하여 $\mu_Q(ti) = 0$ 또는 $\mu_Q(ti) = 1$일 때 HS(r|Q) = 0 이다. 그 이외의 경우 HS(r|Q)는 0보다 크게된다.

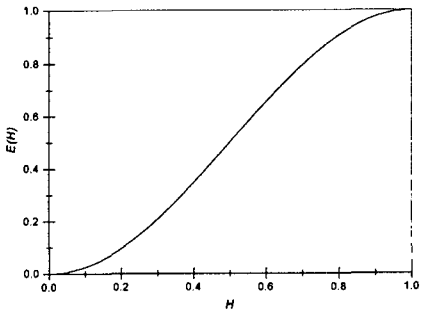
식 (8)은 사용자의 질의에 대한 데이터베이스의 특성을 나타낸다. 즉 HS가 크면 질의에 대한 데이터베이스의 값의 분포가 다양하여 불확실성이 증가하며 HS가 작으면 질의에 대한 데이터베이스의 대부분의 값이 분명한 모습을 나타낸다.

2.4 엔트로피 감소 연산

2.3 절에서 언급하였듯이 엔트로피의 증가는 사용자의 질의에 대한 데이터베이스의 불확실성이 증가하는 것을 의미한다. 이러한 엔트로피를 임의적으로 감소하기 위하여 식 (9)와 같은 연산을 도입할 수 있다.

$$E(H) = (1-\cos(\pi\cdot H))/2 \quad (9)$$

E(H)는 H에 대하여 .5를 기준으로 H가 0에 가까우면 더욱더 0에 가까운 값을 갖도록 하고 H가 1에 가까우면 더욱더 1에 가까운 값을 갖도록 설계되었다. 즉 H가 .8인 경우 E(.8)은 .905의 값을 갖고, H가 .2인 경우 E(.2)는 .095의 값을 갖는다. 엔트로피는 0과 1인 경우 최소의 값을 갖기 때문에 0.5를 제외한 나머지 값들은 <그림 2>에서 보이는 것과 같이 0과 1에 가깝게 함으로써 전체의 엔트로피를 줄일 수 있다. 질의에 대한 데이터베이스의 엔트로피가 작다는 말은 불분명한 정보가 적다는 말과 일치한다.



<그림 2> 엔트로피 감소 함수

3. 응용

이 장에서는 2장에서 제시한 이론적 배경을 실제 관계형 데이터베이스에 응용하여 설명한다.

<표 1>는 관계형 데이터베이스에서의 '개인 신상 관계' 를 나타낸 것이다. 여기서 사용자의 질의가 "고등학교정도의 학력이 있는 26세정도의 젊은 미혼 여성을 찾는다" 일 경우를 살펴보기로 하자.

질의에 대한 퍼지 데이터베이스의 결과를 구하기 위해서는 <표 1>의 각 애트리뷰트에 대한 영역 소속척도를 구하여야한다.

영역의 값이 나이처럼 연속적인 집합인 경우에는 소속척도의 값을 식(10)과 같은 정규 분포 함수를 이용하여 결정할 수 있다[1, 8]. 이 연구에서 고려한 나이는 26세정도임으로 평균값(a)에 26를 대입하고 표준편차(α)에 3(3은 임의의 수임)을 대입하면 각 나이 x에 대한 소속척도의 값을 구할 수 있다.

$$\mu(x,a) = e^{-\frac{1}{2}\left[\frac{(x-a)}{\alpha}\right]^2} \quad (10)$$

<표 1> 개인 신상 관계

이름	나이	최종학력	성별	결혼유무
박순임	30	대졸	여	유
김영자	26	대중퇴	여	무
이상호	19	고졸	남	무
최은희	18	고졸	여	무
정숙희	28	고졸	여	무
신완선	45	대졸	남	유
채자실	55	대중퇴	여	무
천경숙	27	고졸	남	유
마광희	24	고졸	여	유
이향심	28	중졸	여	무

<표 2>은 식 (10)으로 계산된 26세에 대한 각 나이의 소속척도를 보인다.

<표 2> 26세에 대한 나이 소속척도

나이	21	22	23	24	25	26	27	28	29	30	31
소속척도	.2	.4	.6	.8	.9	1	.9	.8	.6	.4	.2

영역의 값이 구분되어 있는 경우(discrete한 경우) 사용자에 의해서 임의의 값을 이용할 수 있다. 이러한 경우 대부분 표로써 소속척도를 표현한다. 즉 학력의 소속척도를 나타내는 <표 3>의 경우 학력 사이의 소속척도를 사용자에 의해서 임으로 정의했다. 질의에서는 고등학교정도의 사람을 요구하므로 <표 3>에서 실제로 사용되는 값은 고졸과 다른 학력간의 소속척도 값들이다.

<표 2>와 <표 3>을 이용하고 성별과 결혼유무에 대한 분명한 관계를 고려할 경우 질의어 "고등학교정도의 학력이 있는 26세정도의 젊은 미혼 여성을 찾는다" 에 대한 <표 1>의 개인 신상 관계의 각 영역값 간의 소속척도는 <표 4>와 같이 나타낼 수 있다.

<표 4>의 영역 소속척도의 값을 질의에 대한 튜플 소속척도의 값으로 변환 할 경우 식(3) ~ (6)의 알고리즘을 이용하여 비교하면 <표 5>와 같은 결과 값을 얻을 수 있다. 2장에서 언급하였듯이 F1의 방법은 튜플에 있는 각 영역의 소속척도의 값 중 가장 작은 값을 취하기 때문에 김영자, 정숙희, 이향심이 똑 같은 .60의 튜플 소속척도 값을 갖는다. 그러나 정숙희와 이향심의 경우 모두 나이가 28세이므로 .60의 튜플 소속척도의 값을 가졌지만 정숙희의 학력이 이향심의 학력보다 사용자가 원하는 수준에 가까우므로 정숙희의 소속척도를 이향심의 경우보다 높게 연산하여야 할 필요성이 있다. 그것을 보상하기 위하여 F2, F3, F4의 방법이 고려되었다. 그러나 F2의 경우 채자실은 나이가 55세이므로 실제로 사용자의 질의의 범위에서 벗어남에도 불구하고 튜플 소속척도의 값이 .65의 높은 값을 취하고 있다. 이러한 단점을 보완할 수 있는 것이 F3과 F4이다. 특히 F4의 경우 사용자가 학력에 강조한 질의를 했을 경우 거기에 대한 보상 결과 값까지 제공할 수 있다. 예를 들면 김영자와 정숙희가 F1, F2, F3의 경우 똑같은 소속척도의 값을 갖지만 F4의 경우 학력이 사용자가 원하는 것과 가장 가까운 정숙희가 가장 높은 소속척도의 값을 얻게된다.

<표 3> 학력 소속척도

	국졸	중졸	고졸	대중퇴	대졸	대졸 이상
국졸	1	.6	.2	.1	0	0
중졸	.6	1	.6	.2	.1	0
고졸	.2	.6	1	.6	.2	.1
대중퇴	.1	.2	.6	1	.6	.4
대졸	0	.1	.2	.6	1	.8
대졸 이상	0	0	.1	.4	.8	1

<표 4> 개인 신상 관계의 영역 소속척도

이름	나이	최종학력	성별	결혼유무
박순임	.2	.2	1	0
김영자	1	.6	1	1
이상호	.1	.6	0	1
최은희	0	1	1	1
정숙희	.6	1	1	1
신완선	0	.2	0	0
채자실	0	.6	1	1
천경숙	.8	1	0	0
마광희	.9	1	1	0
이향심	.6	.6	1	1

퍼지 데이터베이스 시스템에서 질의에 대한 시스템의 엔트로피 계산은 데이터베이스의 성질을 판단하는 데 도움이 된다. 즉 엔트로피가 높을 경우 데이터베이스의 내용이 질의에 대하여 분명한 값을 많이 포함한 경우이며 그 역도 성립한다.

<표 5> 질의에 대한 튜플 소속척도

이름	F1	F2	F3	F4
박순임	0	.35	0	0
김영자	.60	.90	.88	.77
이상호	0	.43	0	0
최은희	0	.75	0	0
정숙희	.60	.90	.88	.88
신완선	0	.05	0	0
채자실	0	.65	0	0
천경숙	0	.45	0	0
마광희	0	.73	0	0
이향심	.60	.80	.77	.68

<표 6>은 튜플 소속척도를 계산한 F1, F2, F3, F4의 엔트로피를 계산한 값이다. F2의 경우 가장 높은 값을 보인다.

<표 6> 질의에 대한 데이터베이스 엔트로피 비교

	F_1	F_2	F_3	F_4
엔트로피	2.91	7.45	1.83	2.20

데이터베이스의 엔트로피를 감소하여 질의에 대한 데이터베이스의 튜플 소속척도를 0과 1로 양분화 할 필요성이 있을 경우 감소함수를 적용할 수 있다.

2장 3절에서 언급한 감소함수를 적용했을 때 <표 4>에 나타난 튜플 소속척도는 <표 7>로 변환된다. 즉 <표 7>의 결과는 <표 5>의 튜플 소속척도를 0혹은 1과 가까운 값으로 변환을 시켰다. <표 7>에 대하여 각 방법에 대한 데이터베이스의 엔트로피를 계산해 보면 <표 8>과 같다. 질의에 대한 데이터베이스의 엔트로피가 <표 6>에 비하여 현저히 줄어든 것을 알 수 있다.

<표 7> 1차 감소 연산 후의 튜플 소속척도

이름	F_1	F_2	F_3	F_4
박순임	0	.27	0	0
김영자	.65	.98	.96	.88
이상호	0	.38	0	0
최은희	0	.85	0	0
정숙희	.65	.98	.96	.96
신완선	0	.01	0	0
채자실	0	.73	0	0
천경숙	0	.42	0	0
마광희	0	.82	0	0
이향심	.65	.90	.88	.77

<표 8> 1차 감소 연산 후의 엔트로피 비교

	F_1	F_2	F_3	F_4
총엔트로피	2.79	5.74	.97	1.53

<표 9>와 <표 10>은 2차 감소 연산 후의 튜플 소속척도와 데이터베이스의 엔트로피 결과 값을 보인다.

<표 9> 2차 감소 연산 후의 튜플 소속척도

이름	F_1	F_2	F_3	F_4
박순임	0	.17	0	0
김영자	.73	1.00	1.00	.96
이상호	0	.32	0	0
최은희	0	.95	0	0
정숙희	.73	1.00	1.00	1.00
신완선	0	0	0	0
채자실	0	.83	0	0
천경숙	0	.38	0	0
마광희	0	.93	0	0
이향심	.73	.98	.96	.88

<표 10> 2차 감소 연산 후의 엔트로피 비교

	F_1	F_2	F_3	F_4
총엔트로피	2.51	4.05	.28	.79

4. 결론

이 논문에서는 vague형 퍼지 데이터베이스을 이용하여 퍼지 질의에 대한 튜플 소속척도를 효과적으로 계산할 수 있는 알고리즘을 제안함으로써 사용자에게 보다 정확한 결과를 제공할 수 있도록 하였다. 또한 질의에 대한 각 튜플의 불확실성(샤론의 엔트로피)을 계산함으로써 질의 결과에 대한 전체 시스템의 특성을 분석할 수 있었다. 아울러 엔트로피를 제어하는 새로운 개념의 엔트로피 감소 연산을 도입함으로써 질의 결과를 제어할 수 있는 기틀을 마련하였다.

향후 연구계획은 현재 한국과학재단 과제로 진행되고 있는 Fuzzy DBMS에 이 논문의 연구 결과를 접목하여 실제로 구현하는 것이다. 또한 이 논문의 연구결과는 미래의 데이터 검색 시스템이 필수적으로 요구되는 자연어 검색과 음성을 통한 데이터베이스 접근에 도움이 될 것으로 기대된다.

참 고 문 헌

[1] 박순철, "관계형 데이터베이스에서 퍼지 질의어 처리," 전북대학교 논문집, 제38집, 자연과학편, 1994, 123-129.

[2] 오길록, 이광형, 『퍼지 이론 및 응용』, 제2권, 홍릉과학출판사, 1991

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

[4] B. P. Buckles and F. E. Petry, "Information-Theoretical Characterization of Fuzzy Relational Databases," IEEE Trans. on Systems, Man, and Cybernetics, Vol. SMC-13, No. 1, pp. 74-77, January/February 1983.

[5] E. F. Codd, "A Relational model of data for large shared data banks," Comm. ACM, 13, 377-387, 1970

[6] C. J. Date, An Introduction to Database Systems, Vol. 1, Addison Wesley, 1985.

[7] R. W. Hamming, Coding and Information Theory, 2nd., Prentice-Hall, Englewood Cliffs, NJ, 1986.

[8] G. Keller, B. Warrack and H. Bartel, Statistics for Management and Economics, Wadsworth, California, 1988.

[9] G. J. Klir and T. A. Folger, Fuzzy Sets, Uncertainty, and Information, Prentice-Hall International (UK) Limited, London, 1988.

[10] H. F. Korth and A. Silberschatz, Database System Concepts, 2nd., McGRAW-HILL, N.Y., 1991.

[11] A. Motro, "VAGUE: A User Interface to Relational Databases that Permit Vague Queries," ACM trans. on Office Information System, Vol. 6, No. 3, PP., 187-214, 1988.

[12] T. Nomura, T. Odaka, N. Ohki, T. Yokoyama and Y. Matsushita, "Generating Ambiguous Attributes for Fuzzy Queries," IEEE Internatioanl Conference on Fuzzy Systems, San Diego, California, March 1992.

[13] S. Shenoi and A. Melton, "Proximity Relations in the Fuzzy Relational Database Model," Fuzzy Sets and Systems, Vol. 31, pp. 285-296, 1989.

[14] S. Shenoi, A. Melton, and L. T. Fan, "An Equivalence Classes Model of Fuzzy Relational Databases," Fuzzy Sets and Systems, Vol. 38, pp. 153-170, 1990.

[15] L. A. Zadeh, "Fuzzy Sets," Information Contr., Vol. 8, PP. 338-353, 1965.

[16] L. A. Zadeh, "Outline of a New Approach to the Analysis of Complex Systems and Decision Process," IEEE Trans. on Systems, Man, and Cybernetics, Vol. SMC-3, N0.1, January, 1973.

[17] M. Zemankova and A. Kandel, "Implementing Imprecision in Information Systems," Information Science, Vol. 37, PP. 107-141, 1985.