퍼지 속성 집합을 이용한 데이터 분석 모델

이진호, 이전영
포항공과대학교 전자계산학과

(790-784) 경상북도 포항시 남구 효자동 산 31 번지
포항공과대학교 정보통신연구소 지능정보시스템연구실

TEL: 0562-279-{5662, 2242} FAX: 0562-279-5699
Email: {zino, jeon}@postech.ac.kr

Data Analysis Model using the Fuzzy Property Set

Zino Lee, Jeonyoung Lee
IIS Lab./PIRL/CSE Dept./POSTECH

San 31, HyoJa Dong, NamGu, Pohang, 790-784, Korea Republic of

Abstract

In this paper, we will propose the methodology of data analysis using the fuzzy property set model. In real world, the data can be represented with the *object*, θ, and the *property*, π, and its *has-property relation*, P. Then, the *conceptual space* can be defined with the chosen properties. Each object has a unique location in the conceptual space.

In Fuzzy model, the fuzzy property, and fuzzy conceptual space can be redefined. To analyze data using the fuzzy property set model, the rough set need to be defined in the fuzzy conceptual space.

서론

데이터베이스 시스템(DB system)은 다량의 데이터를 무결성(consistency)을 유지하면서, 사용자의 요구에 따라 데이터의 입력과 삭제는 물론, 적절한 질의(user query)에 따라, 요구 조건에 맞는 데이터를 빠짐없이 추려내는 컴퓨터 시스템을 지칭한다. 이러한 데이터베이스 시스템을 구현하기 위해서는 우선 실세계의 정보를 자료 형식을 빌어 기술하는 개념적인 도구인 데이터 모델(data model)을 중심으로 이론적인 기반을 마련한다[1]. 현재까지 연구되어 구현된 데이터 모델로는 대표적으로 관계형 모델(relational model)을 비롯하여, 객체 지향적 모델(object oriented model), 망 모델(network model), 계층 모델(hierarchical model) 등이 있다. 그런데, 이들 모델은 본질적으로 실세계에 존재하는 데이터들을 무결성을 유지하면서(consistency constraints), 데이터 자체와 그 구조(structural data model)를 표현한 것들이다. 이들 수학적인 데이터 모델들의 기반이 되는 본질은 개체(object)와 그 속성(attribute), 그리고 그 속성 값(attribute value)으로 요약된다.

데이터베이스에 정리된 각 데이터들은 원시 자료(raw data)로서 활용 가능성은 높지만, 데이터를 분석(analysis)하고 종합(summarization)하기에는 또 다른 이론적인 기반이 필요하다. 본 논문에서는 데이터 모델의 논리적인 기반을 떠나서, 본질적인 데이터(개체와 속성, 그리고 속성 값)를 대상으로 데이터를 분석(data analysis model)하는 모델을 제안하고자 한다. 이렇게 본질적인 데이터를 표현하는 모델은 이미 오래 전부터 연구가 진행되어 왔는데, 대표적인 모델이 속성 집합 모델(property set model)이다. 이 모델은 단순한 데이터의 표현은 물론, 지식의 표현에까지 사용이 되는 수학 모델이다.

본 논문에서 접근하고자 하는 방식은 기존의 속성 집합 모델에 퍼지 집합(fuzzy set)[2] 이론을 도입한 퍼지 속성 집합(fuzzy property set)을 이용한 데이터 분석 모델이다. 따라서, 본 논문의 구성은 이러한 기반 지식을 소개하고, 이를 대상으로 질의 결과 데이터가 속성 공간에 어떻게 대응이 되고, 이렇게 표현된 모델에서 분석을 위한 이론적인 배경을 설명할 것이다.

속성 집합 모델

이것은 실세계의 데이터를 수학적인 모델로 표현하는 방식으로 수학에서 논하는 집합(set)과 관계(relation), 그리고 함수(function)로 구성이 된다[3].

기본 구성

- 실세계에는 개체(object)의 집합인 Θ 가 있다.
- 그리고, 그 개체를 표현하는 성격(attribute)들의 집합 AT 가 있으며,

- 각 성격이 가질 수 있는 값(value)들의 집합 VAL이 있다.

$$VAL = \cup_{a \in AT} VAL_a$$

- 그리고, 여기에서 정보(information)란 각각의 개체가 어떠한 성격에 어떠한 값을 가지고 있는가 하는 함수(function)로 표현이 되며, 이를 정보 함수(information function) I_f라 한다.

$$I_f : \Theta \times AT \rightarrow VAL$$

이러한 기본 구성을 바탕으로 다음과 같이 속성 집합 모델이 정의된다.

정의 1 속성 집합 모델 (property set model)

- 실세계 개체 집합의 부분 집합으로 개체 집합 Θ가 정의된다.

$$\Theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$$

- 어떠한 개체(object)가 가질 수 있는 속성(property)들의 집합을 정의한다.

$$\Pi = \{\pi_1, \pi_2, \ldots, \pi_n\}$$

- 여기에서 속성이란, 성격-값의 관계(relation)로 정의된다.

$$\pi = (a, \upsilon),\ a \in AT,\ \upsilon \in VAL$$

- 개체(object)와 속성(property) 사이에는 연관된 관계(relation) 집합인 has-property를 정의할 수 있다.

$${}_{\theta}P_{\pi} = (\theta, \pi) \in P$$

이렇게 정의된 속성 집합 모델을 기반으로, 개체(object)는 속성들의 집합으로 정의될 수 있으며, 속성(property)은 다시 개체들의 집합으로 정의될 수 있다. 즉, 다음과 같이 정의할 수 있을 것이다.

보조정리 1

속성 집합 모델에서, 속성은 해당하는 속성을 가진 개체들의 집합으로 정의되고, 개체는 자신이 가진 속성들의 집합으로 정의된다.

$$[\pi] = \{\theta \mid \theta \in \Theta, (\theta, \pi) \in P\}$$

$$[\theta] = \{\pi \mid \pi \in \Pi, (\theta, \pi) \in P\}$$

속성 집합 모델에서는 속성들로서 정의되는 개념 공간(conceptual space)을 정의할 수 있다. 속성 공간은 각 속성들이 하나의 축을 정의하는데, 이렇게 할 경우, 각각의 개체들은 고유의 위치를 지정 받게 된다.

정의 2 단위 개념(atomic concept)

속성들로서 정의되는 개념 공간(conceptual space)에서, 각각의 고유 위치를 정의하는 방식으로 특정한 속성들 혹은 속성의 여집합들의 교집합으로 정의된다.

$$[c_0] = [\pi_1]^c \cap [\pi_2]^c \cap \ldots \cap [\pi_n]^c,\ [\pi]^c = \Pi - [\pi]$$

$$[c_1] = [\pi_1] \cap [\pi_2]^c \cap \ldots \cap [\pi_n]^c$$

$$\ldots\ldots$$

$$[c_n] = [\pi_1] \cap [\pi_2] \cap \ldots \cap [\pi_n]$$

여기에서 정의되는 단위 개념은 개념 공간(conceptual space)이 정의됨에 따라서 정의되는 것이며, 개념 공간은 몇몇 개의 속성이 정의됨에 따라 정의되는 것이다. 또한, 정의되는 단위 개념의 개수는 개념 공간을 정의하는 속성들의 여집합(power set) 개수와 같다. 즉, 다음과 같은 식이 성립된다.

$$\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$$

$$N = 2^n - 1$$

$$n = card(\Pi),\ card(A) = \text{cardinality of set A}$$

개념 공간에서 정의되는 각각의 개체들의 위치는 다음과 같다.

보조정리 2

개념 공간 안에서 특정한 개체 θ는 다음과 같이 정의되는 단위 개념(atomic concept)에 위치하게 된다.

$$c_0 = \left(\bigcap_{\pi \in \theta} f_{\pi_i} f \right) \cap \left(\bigcap_{\pi \notin \theta} f_{\pi_i} f^c \right)$$

퍼지 속성 집합 모델

퍼지 속성 집합은 앞에서 정의한 속성 집합 모델에 퍼지 집합 이론(fuzzy set theory)을 도입하여 확장한 것이다. 퍼지 이론을 도입하는 방안으로 우선 정의된 개념 공간의 구분이 모호해지도록 정의하는 방안을 고려할 수 있을 것이다.

따라서, 각 단위 개념(atomic concept)은 퍼지한 속성의 정의에 있어, 특정한 멤버 값(β)을 가지고 참여하게 된다. 즉, 어떠한 퍼지한 성질을 가진 속성이 퍼지 집합(fuzzy set)으로 정의되고, 이름이 B라 가정한다면, 어떤 단위 개념 c는 멤버 값 β로서 B의 원소로 정의될 수 있다.

$$(\beta, c) \in B, \text{ where } \mu_B(c) = \beta$$

따라서, 각각의 단위 개념들은 퍼지한 성격을 가지지 않지만, 개념 공간은 퍼지한 성격을 가진 속성들의 집합으로 정의될 수 있다. 퍼지 속성 집합 모델에서 각 속성들은 이러한 멤버 값을 가진 단위 개념들의 합집합으로 정의된다.

$$[\pi] = \{(\beta_i, c_i) \mid 0 \le \beta_i \le 1,\ c_i \in \mathcal{C}\}$$

정의된 개념 공간에서 각 개체들은 어떻게 정의가 될까? 개념 공간의 속성상 개체는 특정한 단위 개념에 위치하게 된다. 다만, 각각의 단위 개념이 모여 퍼지한 속성을 정의하게 되고, 각 개체는 이렇게 정의된 퍼지한 속성들 혹은 속성의 여집합들의 교집합으로 정의될 것이다.

따라서, 특정한 퍼지 속성이 어떠한 개체를 정의함에 있어 어느 정도의 신뢰도(degree of confidence)를 갖는가 하는 척도가 필요하게 되며, 이를 α라 정의한다.

정의 3 개체 정의에 있어 퍼지 속성의 신뢰도

개체는 고유한 단위 개념(atomic concept)의 위치를 지정 받는다. 퍼지 속성 집합 모델에서 단위 개념은 퍼지 속성들 혹은 속성의 여집합들의 교집합으로 정의되며, 이때 특정 속성이 어떠한 개체를 정의함에 있어 사용되는 신뢰도는 다음과 같이 정의된다.

$$(\alpha, [\pi]) = (\alpha, \{(\beta_i, c_i) \mid 0 \le \beta_i \le 1,\ c_i \in \mathcal{C}\})$$

$$= \{(\delta_i, c_i) \mid \delta_i = 1 - \alpha + \beta_i(2\alpha - 1),\ c_i \in \mathcal{C}\}$$

여기에서 만약 α가 1이라면, $[\pi]$의 속성을 완전히 긍정하는 것이고, α가 1이라면 $[\pi]$의 여집합

속성을 인정하게 되는 것이다.

여기에서 각각의 개체는 정의에 따라 다음과 같이 퍼지한 속성들의 교집합으로 정의된다.

보조정리 3

개체(object) θ는 각 신뢰도를 가진 속성들의 교집합으로 정의된다. 이 교집합은 퍼지 개념 공간(fuzzy conceptual space)에서 퍼지 단위 개념(fuzzy atomic concept)을 정의하는 것이다.

$$\theta=\bigcap_{i=1}^{n}\{(\alpha_i,[\pi_i])\}$$

위의 보조 정리 3과 정의 3에 따라서, 각각의 개체는 단위 개념(atomic concept)의 교집합으로 재정의 될 수 있고, 이 때, 새로이 정의되는 δ_{ij} 는 다음과 같이 정의된다.

$$\delta_{ij} = 1 - \alpha_i+\beta_j(2\alpha_i - 1)$$

따라서, 퍼지 속성 집합 모델에 있어, 개체는 다음과 같이 정의된다.

$$\theta=\bigcap_{i=1}^{n}\{\left(\alpha_i,\bigcup_{j=0}^{N}\{(\beta_j,c_j)\}\right)\}=\bigcap_{i=1}^{n}\left(\bigcup_{j=0}^{N}\{(\delta_{ij},c_j)\}\right)$$

그런데, 퍼지 집합 이론에서 교집합을 수행함에 있어 그 멤버 값은 항상 작은 쪽(minimum)을 택하게 되어 있으므로, 결국 퍼지 속성 집합 모델에서 개체의 정의는 다음과 같이 정의된다.

보조정리 4

퍼지 속성 집합 모델에서 개체(object) θ는 다음과 같이 단위 개념(atomic concept)의 합집합으로 정의된다.

$$\theta=\bigcup_{j=0}^{N}\left(\bigcap_{i=1}^{n}\{(\delta_{ij},c_j)\}\right)=\bigcup_{j=0}^{N}\{(min\{\delta_{ij}\},c_j)\}$$

데이터 분석 알고리즘 개요

퍼지 속성 집합 모델(Fuzzy Property Set Model)은 데이터베이스 시스템에서 사용자의 질의(user query) 응답 데이터를 분석하는데 응용이 될 수 있다. 즉, 사용자의 질의 결과 데이터를 퍼지 개념 공간에 대응시켜, 데이터의 분포를 조사하여 질의 결과 데이터의 성향을 원하는 관점에서 분석할 수 있다.

간단하게 설명하자면, 우선 모든 데이터는 본질적으로 데이터 자체의 집합(개체, object, Θ)과 속성의 집합(property, Π), 그리고, 그들의 관계(has-property, P)로 간단하게 정의될 수 있다. 그렇다면, 이들을 대상으로 다음과 같은 순서로 사용자 질의 결과 데이터를 분석할 수 있다. 물론, 데이터베이스 시스템에서는 각각의 데이터 모델을 대상으로 자신의 데이터베이스를 유지 관리하겠지만, 이들은 본질적으로 앞에서 정의한 것처럼, 개체와 속성 그리고 그들의 상관 관계로 다시 재정의 할 수 있음을 가정한다.

단계 1 사용자가 적절한 질의를 문의하면, 데이터베이스 시스템은 저장된 데이터를 대상으로 요구 조건에 맞는 데이터를 빠짐없이 추려내어 출력한다.
이 때, 데이터 모델에 따라서는 중복된 데이터를 다시 정리하는 과정이 있지만, 여기에서는 중복된 데이터라도 각각 출력한다고 가정한다.[1]

단계 2 분석을 원하는 속성들을 대상으로 개념 공간(conceptual space)을 형성한다. 이 때, 물론 각각의 개체들은 자신들이 취하고 있는 속성들이 각각 다르겠지만, 분석을 위해서 몇 가지 관점에서 속성을 고를 것이고, 이 경우에도 중복을 허용한다.

단계 3 각 개체들을 주어진 개념 공간에 대응 시키는 과정을 수행한다. 이 때, 중복된 데이터는 중복해서 특정한 단위 개념(atomic concept) 위치에 위치시키도록 한다.

단계 4 개념 공간 내 개체 분포를 조사하여 데이터의 성향을 유추한다. 즉, 공간 내에서 각 개체들의 위치를 조사하여, 이들의 전체적인 데이터 성향을 분석하여 유추하도록 하는 과정으로 데이터 분석을 수행한다.

이 때, 개념 공간을 이루는 속성들은 흔히 질의 과정에서 프로젝션(Projection)에 해당하는 속성들을 지칭한다. 즉, 일련의 데이터들을 일정한 속성을 기준으로 분석하기 위해서는 데이터들을 원하는 속성만을 기준으로 추려내는 조건이 필요하고, 이 때 질의 결과 데이터는 프로젝션된 속성들이 형성하는 개념 공간 내에 위치하게 된다. 따라서, 각각의 개체가 가지고 있는 속성들의 집합은 모두 제각기 다를지 모르겠지만, 질의 결과 정리된 데이터들은 모두 같은 속성을 기준으로 표현될 것이다.

데이터 분석 모델

임의의 두개 퍼지 집합의 유클리드 거리(Euclidean Distance)는 각 퍼지 집합을 이루고 있는 원소들의 멤버 값을 대상으로 거리를 측정한다.

보조정리 4에 따라서 각 개체(object)들은 단위 개념과 어떠한 퍼지 집합의 멤버 값의 관계(relation)를 원소로 가지는 퍼지 집합으로 정의가 된다. 따라서, 임의의 두개 개체의 유클리드 거리는 다음과 같이 정의될 수 있다.

$$\delta(\theta,\theta')=\sqrt{\left(\sum_{i=0}^{N}\left(\delta_i-\delta_i'\right)^2\right)}$$

그런데, 거리 측정에 필요한 δ_i는 어떠한 단위 개념(atomic concept)이 특정한 개체를 퍼지 개념 공간 내에서 정의함에 있어 다시 정의되는 멤버 값이므로, 이는 다시 다음과 같이 표현될 수 있다.

$$\delta(\theta,\theta')=\sqrt{\left(\sum_{i=0}^{N}\left(\mu_\theta(c_i)-\mu_{\theta'}(c_i')\right)^2\right)}$$

여기에서 유클리드 거리가 가까우면 가까울수록 두개의 개체는 같은 성향을 나타낸다. 퍼지

[1] 예를 들어, 관계형 데이터베이스(relational database)에서는 질의 결과 데이터를 대상으로 중복된 데이터를 제거하는 과정을 질의 처리 마지막 과정으로 정의한다.

집합에서 유클리드 거리를 이용하여 유사도(similarity)를 다음과 같이 정의할 수 있다.

$$\sigma(\theta,\theta') = 1 - \sqrt{\frac{\left(\sum_{i=0}^{N}\left(\mu_{\theta}(c_i) - \mu_{\theta'}(c_i')\right)^2\right)}{N}}$$

모호성을 표현하는 수학적인 모델로서는 퍼지 집합 이론이 이용되지만, 모호 데이터 분석을 위해서는 러프 집합 이론(Rough Set Theory)[4]이 편리하다.

러프 집합을 정의함에 있어서는 두 개체의 유사도가 중요한 역할을 하고, 임의의 두개의 개체의 유사도가 τ보다는 크다고 하고, 이들의 관계 집합을 $^{\tau}C$ 라 정의한다면, $^{\tau}C$ 는 다음과 같이 정의된다.

정의 4 유사도가 τ보다 큰 개체들의 관계 집합

$$(\theta, \theta') \in {}^{\tau}C \Leftrightarrow \sigma(\theta, \theta') \geq \tau$$

이 관계 집합을 이용하여 러프 집합을 정의할 수 있다. 전체 데이터를 대상으로 분석을 수행하므로, 그 대상은 각각의 개체가 아닌, 개체들의 멱집합(power set)이 될 것이다. 따라서, 러프 집합은 개체들의 멱집합을 대상으로 정의되며, 그 내용은 다음과 같다.

$$T \in 2^{\Theta}$$

$$low(T) = \{\theta | (\theta \in T) \wedge \forall(\theta, \theta') \in {}^{\tau}C \wedge (\theta' \in T)\}$$

$$upp(T) = \{\theta | (\theta \in T) \vee ((\theta, \theta') \in {}^{\tau}C \wedge (\exists\theta' \in T)\}$$

$$bnd(T) = upp(T) - low(T)$$

결론 및 향후 연구 방향

본 논문에서는 개념적인 퍼지 속성 집합 모델을 이용하여, 데이터베이스의 데이터를 분석하는 방법론에 관하여 논하였다. 분석을 위한 데이터를 위해 일정한 개념 공간을 형성하고, 이들의 분포를 알아보기 위해서는 유클리드 거리를 측정하는 방법과 러프 집합을 정의하는 방법을 도입하여 분석을 수행한다면 원하는 분석 결과를 얻을 수 있을 것이다.

하지만, 이를 구현함에 있어서는 모든 개체들의 유클리드 거리를 계산해야 함은 물론, 분석하는 조건에 따라서 그때마다 러프 집합을 재정의 해야 하는 문제가 생길 수 있다. 그래서, 단순히 개념 공간 내에서 출현 빈도를 계산하는 방법을 고려해야 할 것이다.

참고 문헌

[1] 임정훈, 이진호, 이전영, "객체-집합을 기반으로 하는 데이터 모델의 구조적인 정의", 1995년 한국정보과학회 추계학술대회 발표 논문집, pp. 105~108 1995

[2] Lotfi A. Zadeh, "Fuzzy Sets", Inf. Control 8, pp. 338~353, 1965

[3] Michael Hadjmichael, S.K. Michael Wong, "The Fuzzy Property Set Model: A Fuzzy Knowledge Representation for Inductive learning", Proc. Of the 3rd IEEE Conference on Fuzzy System, pp. 684~689, 1994

[4] Z. Pawlak, "Rough Sets", International Journal of Information and Computer Science, 11, pp. 344-356, 1982