퍼지지식베이스에서의 효율적인 정보검색을 위한 규칙생성 및 근사추론 알고리듬 설계

김형수*

*제주한라대학 컴퓨터정보계열

요 약

본 논문은 퍼지지식베이스에서 러프 집합과 요인공간이론을 적용하여 최소 결정규칙 생성과 근사추론 연산을 수행하는 두 개의 알고리듬을 제안한다. 최소 결정규칙의 생성은 속성요인에 관련한 상관분석과 베이지안 정리를 응용한 데이터의 분류기법과 리덕트에 의해 수행된다. 이 결정규칙으로 이루어진 최소지식베이스의 탐색공간에서 소속함수와 t-norm의 합성 연산을 정의한 근사추론 방식에 의해 특정 객체를 검색한다. 본 연구의 러프와 퍼지연산 모듈을 수행하는 제안 알고리듬 기법을 객체및 속성수를 증가시키는 시뮬레이션을 통해 다른 검색이론 및 합성연산 방식과 비교하였다. 그 결과 다른 제 방법보다 본 연구에서 제안하는 기법이 특정 객체를 추출하기 위한 검색연산 시간에 있어 보다 빠르게 검색됨을 입증하였다.

Rule Generation and Approximate Inference Algorithms for Efficient Information Retrieval within a Fuzzy Knowledge Base

Hyung-Soo Kim*

ABSTRACT

This paper proposes the two algorithms which generate a minimal decision rule and approximate inference operation, adapted the rough set and the factor space theory in fuzzy knowledge base. The generation of the minimal decision rule is executed by the data classification technique and reduct applying the correlation analysis and the Bayesian theorem related attribute factors. To retrieve the specific object, this paper proposes the approximate inference method defining the membership function and the combination operation of t-norm in the minimal knowledge base composed of decision rule. We compare the suggested algorithms with the other retrieval theories such as possibility theory, factor space theory, *Max-Min*, *Max-product* and *Max-average* composition operations through the simulation generating the object numbers and the attribute values randomly as the memory size grows. With the result of the comparison, we prove that the suggested algorithm technique is faster than the previous ones to retrieve the object in access time.

Ⅰ. 서 론

최근 정보통신 및 웹 관련 기술이 급속도로 발전함에 따라 축적된 대용량의 정보시스템에서 유용한 정보를 실시간으로 빠르게 획득하기 위한 모델링 기법이 그 어느 때보다 필요하다. 과거의 텍스트위주의 정보제공과는 달리, 다양한 멀티미디어 정보에 대한 축적 및 가공을 통한 동기화된 양질의 정보 서비스가 요구된다. 또한 데이터베이스의 양은 점점 대용량되고 있으며 그 데이터 구조 역시 복잡하다. 따라서 이와 같은 대용량의 데이터베이스에서 효율적인 정보 검색을 위해서는 데이터 간의 분류, 캡슐화, 상속, 병행성 및 스키마 진화의 기능을 고려한 적절한 데이터 모델로 지식베이스를 구축하여야 한다[1~3]. 특히, 불확실한 데이터가 있는 퍼지 데이터베이스의 구축에 있어서는 객체 및 속성 스키마에 대한 관계가 잘 정의되어야 하며 특정 튜플 및 도메인을 추출하기 위한 연산의 절차도 효율적으로 모델링이 되어야 한다. 그 동안 데이터베이스의 도메인이 확정적인 스칼라 값인 경우, 그 관계연산은 불로직(boolean logic), 의미망(semantic net), 프레임(frame) 및 술어해석(predicate calculus)의 지식표현으로 수행하였다. 반면에, 불확실성 개념에 대한 정보처리는 퍼지로직(fuzzy logic), 확률로직(probability logic) 및 베이지안로직(bayesian logic)에 기반하여 모델링하였다[4~6].

데이터베이스 내의 퍼지정보에 대한 속성요인들의 합성연산을 통해 특정 객체의 검색을 수행하는 주요이론은 증후이론(evidence theory)[7], 가능이론(possibility theory)[8], 요인공간이론(factor space theory)[9,10], VAGUE모델[11]과 러프집합이론(rough set theory)[12,13] 등이 있다. 러프집합 이론은 결정 속성에 대한 조건 속성의 하한근사(lower approximation) $Apr^*(X)$ 및 상한근사(upper approximation) $Apr^*(X)$의 객체 분류에 기초하여 정보시스템 내의 최소의 지식획득(knowledge acquisition)이 가능하다. 또한 요인공간이론은 획득된 지식베이스 내의 객체(object) u와 속성요인(attribute factor) f에 일치하는 상태공간(state space) $X(f)=\{f(u) \mid u \in U\}$에서 퍼지로직에 의한 합성연산에 의해 특정객체를 검색하는 구조를 제공한다.

이런 점에서 본 논문에서는 퍼지지식베이스에서 특정 객체를 추출하기 위한 지식발견 알고리듬과 퍼지합성연산을 효율적으로 수행하는 알고리듬을 제안하고자 한다. 퍼지질의어에 대한 지식발견은 러프이론의 리덕트(reduct)에 의거하여 최소결정규칙을 생성한다. 다중리덕트 및 부분가능 결정규칙에 대한 기존 러프집합의 문제점을 본 연구에서는, 결정 및 속성요인 간의 상관관계와 베이지안 정리를 적용한 알고리듬을 제안하여 해결한다. 또한 리덕트 결과, 최초의 데이터 특성을 유지한 최소지식베이스에서 특정 객체를 검색하기 위해 퍼지 소속함수의 정의와 t-노름(norm) 퍼지 합성연산을 적용한 알고리듬도 제안한다.

본 연구의 제안 알고리듬과 기존 퍼지정보 검색이론인 가능이론및 요인공간이론, 퍼지 합성연산인 Max/Min, Max/Product 및 Max/Average[15]과의 객체 및 속성의 수를 증가하는 시뮬레이션을 통해 검색시간의 효율성을 입증하였다.

Ⅱ. 지식베이스의 구성

2.1 퍼지지식베이스 표현

데이터베이스 내의 지식은 주어진 객체 집합에 대한 속성요인간의 관계구조(relation structure)로 특징화시킬 수 있다. 퍼지지식베이스(fuzzy knowledge base) FKB = 〈 U, C, D, V, f〉으로 구성되며 U, C, D는 비공집합(non empty)이다. 객체 U = {Ui | i = 1,… n}는 조건 속성 C ={Ci | i=1,…, n}, 결정속성 D= {Di | i = 1,…,n}의 도메인과 연관된다. A = C ∪ D로써 모든 속성 집합을 나타내며 C ∩ D = ∅ 이다. V=∪a∈A Va 이며, Va 는 구간길이방법(interval width method)에 의해 분류되는 속성 a ∈ A 대한 퍼지 유한속성 도메인(fuzzy finite attribute domain)이다. f 는 모든 a∈ A 와 Ui ∈ U 대한 f : U × A → V에 대응되는 f(Ui, a)∈ Va 정보함수이다.

지식베이스 FKB의 구성 요소간의 구분불가능관계(indiscernibility relation) IND(B)는 관계 R이 U × U상에서 동치관계(equivalence relation)일 때 IND(B)={(Ui, Uj)∈R | ∀a∈ B, f(Ui, a)=f(Uj,a)} 인 이항관계이다. 전체 집합 U는 기본집합(eleme_ntary set)의 합집합으로 구분 불가능한 관계인 〔Ui〕R 또는 U/B의 동치류(equivalence class)로 분할 할 수 있다. 임의의 객체 Ui ⊆U에 대한 동치류 〔Ui〕R는 지식베이스를 구축함에 있어 가장 기본적인 개념 블록이 된다.

2.2 속성분류

결정속성에 대한 조건속성의 객체로 분류가 되는 동치류는 서로 다른 결정 속성에 속하는 불일치성(inconsistency)을 가질 수 있는 데, 러프집합은 이러한 불일치성에 기초하여 다음과 같은 정의의 상·하한 근사 집합으로 분류한다.

〔정의 1〕 퍼지지식베이스 FKB=〈U, C, D, V, f〉에서 X⊆U이라 할 때 대수적 러프집합(algebraic rough sets)은 X 의 하한 근사 Apr*(X), 상한근사 Apr*(X)와 경계역 BND(X)에 의거하여 (〔Ui〕R,∩,∪,~, Apr*(X), Apr*(X))로 구성된다.

$$Apr^*(X) = \{ Ui \in U \mid [Ui]R \subseteq X \}$$

$$pr^*(X) = \{ Ui \in U \mid [Ui]R \cap X \neq \emptyset \}$$

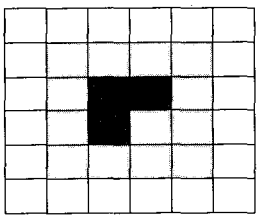
$$BND(X) = Apr^*(X) - Apr^*(X)$$

여기서, 하한근사 Apr*(X)는 X의 부분집합인 모든 기본집합의 합집합으로써 X내에 포함된 가장 큰 합성집합이다. 상한근사 Apr*(X)는 X와 비공집합을 갖는 모든 기본집합의 합집합으로 X를 포함하는 가장 작은 합성집합이다. 러프집합의 하한·상한근사에 관련한 근사정확도 (accuracy of approximation)의 계산을 통해 러프집합을 분류 할 수 있다. 두 집합 X,Y간의 동질성(similarity)과 근접성(closeness)을 나타내기 위해 MZ(Marczewski-Steinhaus)메트릭를 이용한 척도함수 S(X,Y) = | X∩Y | / | X∪Y | 에 의해 근사정확도를 계산한다. S(X,Y)는 집합 X, Y 가 서로 소이면 1이고 동일하면 0이다. 지식베이스에서 결정조건 D에 대한 속성 조건 C의 러프집합의 근사정확도 $\alpha c(X)$ = S(Apr*(X), Apr*(X)) 이다. $\alpha c(X) \in [0,1]$의 값을 가지며, $\alpha c(X)$ = 1이면 C에 대해 X는

경계역이 없는 하한근사역 동치류의 집합임을 나타낸다. 반면에 질의어 X에 대해 속성 분류 시, 대부분 상한근사역의 경계역이 존재하게 되어 근사정확도의 값은 $\alpha c(X) < 1$가 되어 확률적 러프집합(probabilistic rough sets)의 정의에 적용된다.

2.2 근사 결정규칙 생성

전체집합 U의 부분집합인 X는 SQL 질의어에 의거하여 하한 및 상한근사 정의에 기초하여, [그림 1]과 같이 서로소(disjoint)인 양역 POS(X), 음역 NEG(X)과 경계역 BND(X)의 세 근사영역으로 구분된다[14].



(그림 1) 질의어의 근사영역

[그림 1]의 근사영역을 대수적 러프집합의 개념에 의거하여 표현하면 다음과 같다.

$$POS(X) = Apr^{*}(X)$$
$$NEG(X) = U - Apr^{*}(X)$$
$$BND(X) = Apr^{*}(X) - Apr^{*}(X)$$

특히, 객체 동치류 $x \in BND(X)$인 경우, 원소 x는 X 집합의 원소인 지 여부를 확실하게 단정할 수가 없어 조건부확률에 의해 원소여부를 판별하게 된다. 그러나 $BND(X) = \emptyset$ 인 경우에 있어서의 X는 기존의 관계데이터베이스의 관계대수 연산에 의한 이진검색을 통해 객체 검색이 가능하다. 반면에 $BND(X) \neq \emptyset$ 인 경우에 있어서는 X는 근사 결정규칙을 생성하여 객체를 검색하기 위한 확률적 러프집합의 결정임계치(decision threshold value)에 의한 객체 연산의 대상이 된다.

[정의 2] 전체 객체 집합 U의 원소 x, y 가 주어지고 이항관계 R이 다음과 같은 사상 r 로 정의된다고 가정하면 이 r(x) 는 x 의 모든 R-관련된(R-related)원소를 나타낸다.

$$r : U \rightarrow 2^{U}$$
$$x \in U \rightarrow r(x) = \{y \in U \mid xRy\}$$

이 r(x)에 기초로 하여 U에 관한 이항연산 R에 대한 대수적 러프집합의 확장인 확률적 러프집합의 상 · 하한근사는 다음과 같이 계산된다.

$$Apr_{*\delta}(X) = \left\{ x \in U \mid \frac{|\overline{X} \cap r(x)|}{|r(x)|} \leq \delta \right\}$$
$$Apr^{*}{}_{\delta}(X) = \left\{ x \in U \mid \frac{|X \cap r(x)|}{|r(x)|} > \delta \right\}$$

여기서 인수 $\delta \in [0, 0.5)$는 근사 결정규칙을 생성하기 위한 결정임계치(decision threshold value)이다. $\delta = 0$인 경우는 $Apr^{*}0(X) = Apr^{*}(X)$, $Apr^{*}0(X) = Apr^{*}(X)$ 관계가 성립하며, 대수적 러프집합 연산 성질과 같다. BND(X)의 영역이 이 확률적 러프집합 연산의 기본 대상이 된다. δ의 값의 지정에 따라서, 결정규칙의 하한근사의 양역 POS(X)가 되어 객체 검색의 대상으로 결정된다. 본 연구에서는 경계역에 속한 객체의 탐색(mining)을 위한 분류는 베이지안 정리(bayesian theorem)를 이용하였다. δ-절단 집합에 의해 경계역에 속한 객체 동치류가 상 · 하한

근사로 분류되어 최종 결정규칙이 생성된다. 이 결정규칙의 집합은 여분속성이 제거된 최소지식베이스로써 요인공간 이론의 검색 상태공간(retrieval state space)이 된다.

2.3 검색 상태 공간

검색 상태 공간은 데이터베이스의 객체와 속성 그리고 객체와 속성요인간의 관계(relation)로 구성되는 일종의 개념적 스키마(conceptual schema)이다. 이 검색 상태 공간에서 특정 객체 또는 튜플의 검색은 전형적인 데이터베이스 조작언어인 SQL의 select-from-where 구조의 질의어에 의해 검색된다. 검색을 위한 객체 u는 속성 요인 f에 관련(relevant) 되어 u 에서 f 로의 사상(mapping)인 f(u)로 그 관계가 표현된다[9,10].

U, V 를 객체와 속성요인들의 전체 집합이라 할 때, 임의의 u∈U 에 대하여 u에 관련된 모든 요인들을 v가 포함할 때 (U, V]를 좌대칭쌍(left-matched pair)이라 한다. 좌대칭쌍 (U, V]에서의 U, V 간 일반적 관계 R은 다음과 같이 정의된다.

[정의 3] u가 f에 관련되면 R(u, f)=1이다. D(f)={u∈U | R(u, f)=1}, V(u)={f∈U | R(u, f)=1}일 때 요인 f∈V는 객체 u 에 대해 f(u)가 산출되는 사상이다.

$$f : D(f) \rightarrow X(f)$$
$$u \rightarrow F(u)$$

여기서 X(f) = {f(u) | u∈U}는 전체 상태공간이다. 이 X(f)의 각 원소는 f의 상태(state)로써 논리곱(∧), 논리합(∨) 및 논리부정(~)의 일반적 불대수(boolean algebra)의 기본 연산을 만족한다. 검색 상태 공간의 특성에 따라 요인들의 도메인은 정수, 실수 및 용어집합등의 측정가능요인(measurable factor), 스칼라요인(scalar factor), 정도요인 (degree factor), 스위칭요인(switching factor)의 값으로 구분된다.

2.4 요인 연산의 성질

2.4.1 기본 요인

영요인(zero factor)은 상태공간 X(0)={θ}를 만족하는 경우로써 기호 0 으로 나타낸다. 여기서 기호 θ은 검색을 위한 원소가 없는 공상태(empty state)임을 말한다. 영요인을 제외한 부분요인이 없는 요인 f 는 원자요인(atom factor)이 된다. 특히, 원자요인 T에 대하여 임의의 원자 s, t 에 대하여 $f_s \wedge f_t = 0$ 이면 독립요인(independent factor)의 관계가 성립한다고 한다. 상태 공간의 집합군 $\{F_t\}_{t\in T}$ 가 독립이면 $X(V_{t\in T} f_t) = \Pi_{t\in T} X(f_t)$ 이다. 여기서 Π 는 카티션 곱 연산자(cartesian product operator)이다. 요인 g가 다음과 같은 조건을 만족할 때 f의 진부분요인(proper subfactor)이라 하며 f 〉 g 로 표기한다.

∃Y(Y≠ϕ & Y≠{0}) w.r.t X(f) = X(g) × Y

2.4.2 요인연산

요인 f, g 의 요인곱(conjunctive of factor)인 h (f ∧ g)는 요인 h 가 f, g 의 최대공약 부요인(greatest common subfactor)으로써 만일 f ≥ h, g ≥ h 이고, h ≥ e 관계가 성립하는 임의의 요인 e 가 존재하는 경우이다. 반면에 요인합(disjunction of factor)인 h (f ∨ g)는 h ≥ g 이고, e≥f, e ≥ g일 때 e ≥ h 를 만족하는 e 가

존재할 때이다. f, g의 요인차(difference of factor) h = f - g는 (f ∧ g) ∨ h = f 이면서 h ∧ g = 0 일 때다. f 의 요인여(complement of factor)는 f^c =1 - f 이며, 여기서 1 은 완전요인(complete factor)이다. 한편 요인공간 (U, V)에서 F ⊂ V일 때 집합군 $\{X(f)\}_{f\in F}$ 는 U 상에서 다음의 공리를 만족한다.

① F = F(∨, ∧, c, 1, 0)는 완전 불대수이다.

② X(0) = {θ}

③ ∀T ⊂ F 이고 ($^{\forall}s$, $t \in T$) (s ≠ t ⇒ s ∧t= 0) => $\vee_{f\in T} f = \prod_{f\in T} f$

Ⅲ .최소결정규칙의 생성

3.1 여분속성의 제거

데이터베이스에서 속성의 의존성 분석 (dependency analysis)을 통해 효율적인 정보 검색이 가능하다. 본 연구에서는 그 의존성 분석을 조건속성과 결정속성의 이 변량모집단(bivariate population)의 상관계수 계산에 의거하여 계산한다. 특정 결정속성에 대해 상관관계가 높은 조건속성부터 그 여분성을 먼저 조사하면서 계속적으로 속성리덕트(attribute reduct)에 첨가하는 알고리듬을 제안하여 다수의 속성리덕트를 생성하는 러프집합의 제한성을 개선하였다.

최초의 퍼지지식베이스 FKS = (U, C, D, V, f)에서는 결정속성에 따라 중복(redundant) 또는 여분(superfluous)의 조건속성이 있을 수 있다. 시스템 내에서 이러한 조건속성의 필수 불가결, 중복 또는 여분 속성의 판별은 속성리덕트 절차에 의해 결정된다. 여분속성을 제거하는 속성리덕트의 결과인 REDD(C)는 최소 데이터 탐색공간인 감소 정보시스템으로서, 주어진 최초 정보시스템의 데이터의 특성과 패턴이 그대로 보존된다. FKS에서 객체 U에 부가적인 정보를 제공하지 않는 조건속성 C를 발견하여 제거하기 위한 속성리덕트 정의는 다음과 같다.

[정의 4] C*, D*는 각각 관계 INC(C), INC(D)의 동치류의 집합이다. 근사공간 Apr = (U, INC(C))에서 C*에 대한 분할 D*의 모든 기본 집합의 하한근사의 합집합 POSc(D) = Ux ∈D* C*(X) 일 때, 속성요인 a 의 여분속성 여부결정은 아래와 같다.

① If POSc(D) = POS(c-{a})(D) then
a : D-superfluous attribute(for all a∈C) else
a : D-indispensable attribute(for all a∈C)

② Subset C′⊆C, C′ is attribute reduct iff
C′ is D-indispensable attribute and POSc(D) = POSc′(D)

여분속성 여부를 결정하는 [정의 4]를 기초로 하여 조건속성 C와 결정속성 D 요인간의 의존도 (degree of dependency) rc(D) = | POSC(D) | / | U | 이며, C $\xrightarrow{\gamma}$ D(단, rc(D) ∈ [0,1])로 표현한다

[정의 5] 결정속성 D에 대한 조건속성 C의 부분집합인 C′ ⊆ C가 속성리덕트가 될 필요충분조건은 다음과 같다.

① rc′(D) = rc(D)

② For all a∈C′, rC′(D) ≠ rc′-{a} (D)

[정의 5]에서 ①의 조건 rc′(D) = rc(D)는 속성 C′가 D에 대한 의존도의 보존으로써 최초 정보시스템의 속성의 특성을 유지하고 있음을 나타낸다. 또한, ②의 조건을 만족하는 C′가 존재하면 이 C′는 최소 지식베이스의 속성리덕트 REDD(C)가 된다. 이 경우 다수의 속성리덕트가 존재 시, 조건속성 C에 있는 모든 D-필수 불가결 속성의 집합은 C 에 대한 D-코아(core) CORD(C) 집합이 되며, CORD(C) = ∩ REDD(C) 이다.

3.2 결정규칙의 생성

속성리덕트 결과 감소 정보시스템에서 속성값의 리덕트인 도메인리덕트를 통해 다시 최소 결정규칙(minimal decision rule)을 생성할 수 있다. 이 최소 결정규칙은 데이터베이스 내의 속성의 중복성이 없이 조건속성의 수를 최소화하면서 주어진 정보 시스템에서 데이터 검색을 위한 상태공간을 이룬다. 결정규칙 R은 하한근사역의 부분(partial) 결정규칙과 경계역 BNDC(D)의 부분가능(partial possibly) 결정규칙으로 구성되어 해당 객체의 패턴을 나타내게 되며, 조건속성 Ci (i=1,....,n) 와 결정속성 D의 도메인 Vii의 요인곱의 결합이다.

R : (C1 = Vi1)∧(C2 = Vi2)∧ .... ∧ (Cn =Vm) → (D = Vii)

조건속성의 분류 결과 Ei(C)에 있는 객체가 때로는 서로 다른 결정속성 분할 Ej(D)에 속할 경우, 어느 결정영역에도 속할 수 없는 결정모순(decision conflict)에 빠지게 된다. 퍼지질의어에 대한 근사추론을 행하기 위해서는 하한근사의 부분 결정규칙 외에 결정모순의 상한근사역에 있는 객체를 근사적으로 검색할 수 있는 도메인리덕트 가 필요하다.

본 연구에서의 도메인리덕트는 [정의 2]의 확률적 러프집합에 기초하여, 다음 P[j | Ei(C)] 조건부확률인 베이지안 정리를 적용하여 결정속성을 생성하였는 데 해당 알고리듬은 [그림 2]와 같다.

$$P[j \mid E_i(C)] = \frac{P[E_i(C) \mid j]Q_j}{P[E_i(C) \mid 1]Q_1 + \ldots\ldots + P[E_i(Cm) \mid m]Q_m}$$

여기서, Qj는 최초의 정보시스템 FKS의 전체 객체에서 각각의 결정 분할 Ej(D) 객체가 존재할 확률이다. 조건부 확률 P[Ei(C) | j]는 경계역 BNDC(D)에 있는 임의의 조건속성 분류 Ei(C)의 객체가 각각의 결정속성 분할 Ej(D)에서 발견될 확률이다. 또한, P[j | Ei(C)]는 조건속성의 분류 Ei(C)의 객체가 서로 다른 결정속성 분할 Ej(D)에서 발견할 확률이다.

```
Input : Attribute Reduct RED_D(C)
Output: Minimal Knowledge Base (U,C,F)
Attribute classification_procedure()
  Do /* Calculation of Quotient set E_i(C), E_j(D) */
     classify  E_i(C) ;        /* 조건속성의 분류 */
        U/C = ∪ E_i(C) ;
     partition  E_j (D) ;   /* 결정속성의 분류 */
        U/D = ∪ E_j (D) ;
  Do /* decide boundary region  E_i(C) w.r.t E_j (D)  */
  R = Ø;  r_i = Ø ; /* 부분(R_i)및 가능(r_i)결정규칙의 초기화 */
  For i=1 to n
  For j=1 to m
   If  E_i(C)⊂E_j (D) then { E_i(C)⊂POS(D):R_i =R∪R_i ;}
    else If E_i(C)⊆E_j(D) then { E_i(C)⊂BND(D) : r_i  = r∪r_i ;}
          else E_i(C)⊂NEG(D)
         Endif
  Endif
Decision Rule_Procedure()
  select r_i ;  /* 경계역 BND(E_i(C)) 객체 선택 */
  Do /*generate  approximate  decision  rule  for  E_i(C)⊂
BND(D)*/
  set  α :   /*  α : 결정 임계치    */
     calculate Q_j w.r.t E_j (D) & P [E_i(C)  |  j] in E_j (D) :
     calculate probability P[j  |  E_i(C)]:
     decide w_j^i = Max P[j  |  E_i(C)]:
  Do /*E_i(C)의  경계역 객체 E_j (D)에 대한 하한근사 여부조사*/
  If 1-α ≤  w_j^i  then  /* 하한근사 시 근사 결정규칙 R_i 채택 */
     {E_i(C) ⊂E_j (D) ; R =R∪R_i ;}
    else E_i(C)⊂NEG(D)
  Endif
```

(그림 2) 알고리듬 1 : 결정규칙 생성

[그림 2]의 〈알고리듬 1〉은 우선 최초의 퍼지지식베이스 FKB = 〈 U , C , D , V , f 〉에서

객체 검색 질의어에 따라, 결정 속성에 대한 조건 속성의 여분속성을 제거한 [정의 5]의 속성리덕트 결과가 입력 요소가 된다. 러프집합의 다중 리덕트 생성에 대한 제한점은 결정 및 조건속성에 대한 공분산의 값에 따라 코아(core) 리덕트로서 유일한 최적 리덕트를 취한다.

〈알고리듬 1〉에서 전반부인 속성분류 모듈인 Attribute classification_procedure()은 리덕트 결과에 대한 [그림 1]의 질의어 근사영역의 상·하한근사 객체에 대한 결정 및 조건속성의 분류 과정이다. 새로운 감소된 지식베이스를 얻기 위한 결정규칙의 생성은 하한근사역에 속한 부분 결정규칙을 제외한 경계역에 속한 객체들의 부분 가능규칙에 대한 처리가 요구된다.

본 연구에서는 이 경계역 객체에 대한 최종 결정규칙의 생성은 [정의 2]의 확률적 러프집합 정의에 기초하여 P[j | Ei(C)]의 베이지안 정리를 적용하여 결정임계치 α-절단집합에 의해 결정된다. 이 과정은 〈알고리듬 1〉의 후반부의 결정규칙 생성 모듈인 Decision Rule _Procedure()에서 수행된다. 이와 같은 〈알고리듬 1〉의 결정규칙 생성을 위한 전·후반부의 데이터 분류 및 베이지안 정리에 의거한 확률적 러프집합에 적용한 결과, 감소된 최소 지식베이스로써의 새로운 검색 상태 공간인 질의어의 개념 C의 탐색공간(description frame) (U, C, F)를 생성한다. 여기서 F는 질의어의 개념 C를 만족하는 리덕트의 결과로써 여분속성이 없는 속성요인의 집합이다. 또한 U는 C에 의한 객체 전체집합으로써 베이지안 정리 적용 결과, 각 원소는 $\{\forall u_1, u_2 \in U, (\exists f \in F) \mid f(u_1) \neq f(u_2)\}$의 단사(injection) 조건을 만족한다.

Ⅳ. 근사추론의 컴퓨팅

4.1 퍼지로직으로의 변환

퍼지 및 러프집합 이론은 불확실한 정보를 효율적으로 모델링을 할 수 있는 논리적 연산구조를 제공한다. 우선 퍼지집합의 퍼지로직에 적용하기 위한 소속함수는 볼록정규화(convex normalized)의 퍼지집합과 구분연속함수(piecewise continuous function)의 성질을 만족해야 한다. 본 연구에서의 소속함수 $\mu_{c_i}(x)$ 는 측정 가능한 수치 도메인을 갖는 속성요인들의 최대수(q*), 최소수(q*)에 의해 좌삼각퍼지수(left triangular fuzzy number)의 형태의 소속함수로 다음과 같이 정의하였다.

$$\mu_{c_i}(x) = \begin{cases} 0 & x \in (-\infty, q_*] \\ (x - q_*)/(q^* - q_*) & x \in (q_*, q^*] \end{cases}$$

이 $\mu_{c_i}(x)$ 는 퍼지 개념 α에 대한 소속의 정도이며, 퍼지집합의 근사적인 값을 구하는 데 적용된다. 특히, 도메인 값을 결정하는 속성요인의 양(positive)·음(negative) 개념에 따라 양의 개념의 속성인 경우는 단조증가 함수를 취하고, 반면에 음개념의 속성의 소속함수의 값은 1-$\mu_{c_i}(x)$ 를 취함으로써 단조감소 함수의 값이 되도록 하였다.

Input : Retrieval State Space(U , C , F)
Output: Specific Object retrieval
Reduct_Generation_Procedure()
Decision Rule_Procedure()
Approximate Inference_ Procedure()
Define a concept for $U=\{U_i \mid i \in \{1, n\}\}$
Determine α *in* C
Determine $a_j(u_i)(\ i \in \{1, n\}, j \in \{1, m\})$ *& Complete factor*
$1(u_i)$
$=(a_1(u_1), a_2(u_2), \ldots\ldots, a_m(u_i))$
Convert Fuzzy representation extension $\alpha_{C_i}(a_j)(u_i)$
w.r.t α
Calculate m-ary triangle norm T_m
$\alpha_{C_i}(1)(u_1, u_2, \ldots, u_n)$
$= T_m(\alpha_{C_i}(a_1)(u_1), \alpha_{C_i}(a_2)(u_2), \ldots, \alpha_{C_i}(a_m)(u_n))$
$= \alpha_{C_i}(a_1)(u_1) \cdot \alpha_{C_i}(a_2)(u_2) \cdot \ldots \cdot \alpha_{C_i}(a_m)(u_n)$
Calculate Feedbback extension $1^{-1}(\alpha_{c_i}(1))$ *w.r.t* α
$1^{-1}(\alpha_{c_i}(1))(u_i)$
$= \alpha_{c_i}(1)(1(u_i))$
$= T_m(\alpha_{c_i}(a_1)(a_1(u_i)), \alpha_{c_i}(a_2)(a_2(u_i))$
$\ldots\ldots, \alpha_{c_i}(a_m)(a_m(u_i)))$.
Retrieve Max U_i *by 1st Maximum membership principle*

(그림 3) 알고리듬 2 : 근사추론

4.2 근사추론에 의한 객체검색

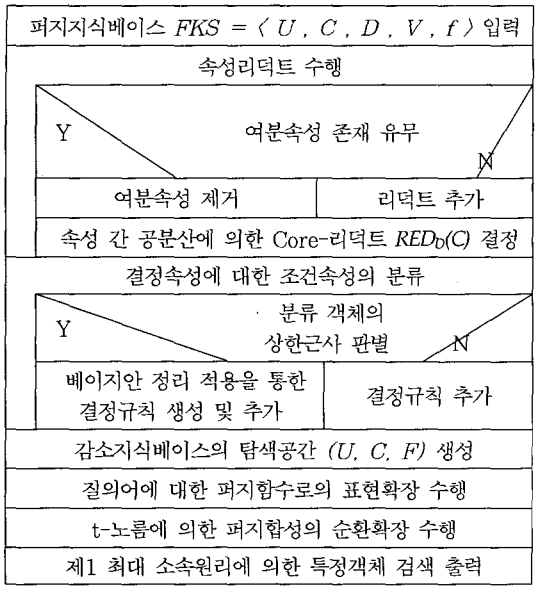
근사추론에 의한 특정 객체를 검색하기 위한 절차는 〈알고리듬 1〉의 결과, 생성된 최소지식베이스의 검색공간 (U, C, F)에서 퍼지 질의어 C에 대한 속성요인 F의 퍼지 도메인에 대한 정형화된 알고리듬이 필요하다. 이런 점에서 본 연구에서는 퍼지 개념 $\alpha \in C$ 의 근사 추론 과정으로써 표현확장(representation extension) 및 순환확장(feedback extension)의 개념을 정의한 〔그림 3〕 근사추론 〈알고리듬 2〉를 제안하였다. 근사적 추론모듈로써 Approximate Inference_ Procedure()는 탐색공간 (U, C, F)에서의 퍼지함수로의 정형화를 나타내는 표현확장및 순환확장 연산으로 구성된다.

퍼지집합 $\tilde{A} \in F(U)$의 확장 개념인 질의어 $\alpha \in C$ 가 주어질 경우, 우선 α의 표현 확장 $f(\tilde{A})$는 $f(\tilde{A}) \in F(X(f))$으로 요인 상태공간 X(f)의 퍼지 부분집합이면서 삼각퍼지수 $\mu_{c_i}(x)$ 성질을 갖게 된다.

$$f(\tilde{A}): \ X(f) \rightarrow [0,1]$$
$$x \ \rightarrow f(\tilde{A})(x) = V_{f(u)=x}\tilde{A}(u)$$

또한, 순환확장 $1^{-1}(\alpha_{c_i}(1))$는 〈알고리듬 2〉에 나타낸 것처럼 m 개의 조건속성 요인에 대한 n 개의 각 객체의 표현확장 값인 $\mu_{c_i}(x)$ 의 t-노름 의 합성연산이다. 이 순환확장은 가능이론(possibility theory)에 있어서 각 속성간의 유사정도를 고려한 가능분포의 사영 일반화(projection generation)의 연산과 동일한 카티션 곱연산이다.

본 연구에서 제안하고 있는 〈알고리듬 1, 2〉를 종합하여 최초의 퍼지지식베이스 FKS = 〈 U , C , D , V , f 〉에서 최종 특정 객체를 검색하기 위한 일련의 연산의 순서도는 〔그림 4〕의 N-S 차트(Nassi-Schneiderman Chart)와 같다. U의 특정 객체의 검색은 질의어 $\alpha \in C$ 에 대한 각 속성요인 F의 분류를 통한 결정규칙의 생성과 퍼지함수로의 표현확장과 도메인의 순환확장의 합성연산 결과, 제1최대 소속원리(1st Maximum membership principle)에 의해 최종 객체를 검색하게 된다.

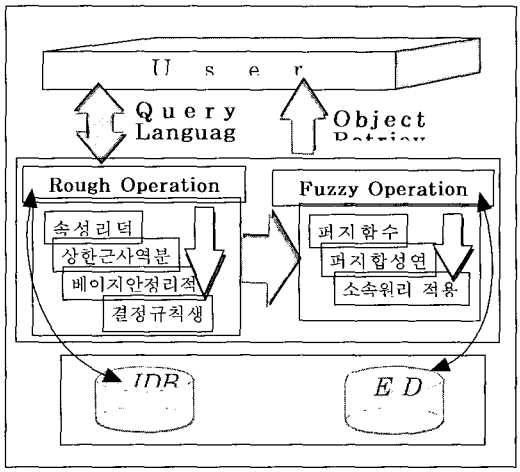


(그림 4) 제안 알고리듬1, 2의 연산처리 순서도

4.3 제 연구와의 비교

대용량화된 데이터베이스 내에서 유용한 정보만을 탐색(mining)하여 추출(extract)하고자 하는 여러 기법이 소개되고 있으며, 특히 웹 상에서 특정 객체를 추론 검색하고자 하는 기계학습을 통한 정보검색형 지능형 에이전트에 대한 연구도 활발하다[16,17].

이런 점에서 본 연구는 퍼지데이터베이스 내에서 검색 시간의 효율성을 고려하여 감소된 지식베이스의 생성과 퍼지 연산의 정형화를 위한 〈알고리듬 1, 2〉를 제안하였다. 이에 관련한 객체검색 구조는 [그림 5]와 같다. 우선 최초의 퍼지지식베이스 FKS=〈U, C, D, V, f〉가 저장된 IDB(identity database)에서 질의어의 속성요인에 대한 리덕트와 결정규칙을 생성하는 러프연산모듈(rough operation module)의 전단계와 감소된 지식베이스 (U, C, F)의 EDB(extension database)의 데이터를 이용하여 퍼지근사추론을 수행하는 후단계의 퍼지연산모듈 구조로 구성되어 객체를 검색하게 된다.

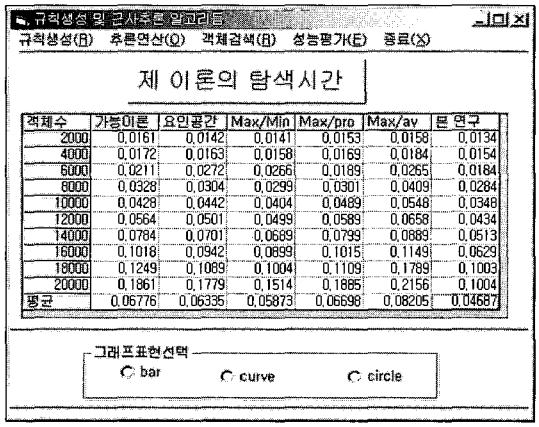


(그림 5) 객체 검색 구조

퍼지정보에 대한 정보 검색의 이론과 그 모델링의 방법은 데이터베이스 내의 데이터의 특성에 따라 각각 그 특징이 있다. 이런 점에서 본 연구에서는 제안 〈알고리듬 1, 2〉의 러프 및 퍼지 연산 모델링 방법과 기존의 퍼지 정보 검색 이론인 가능이론, 요인공간이론, Max/Min, Max/Product, Max/Average 퍼지합성 연산과의 객체 검색시간을 비교하였다. 가능이론의 유사정도(similarity degree)와 연구에서의 결정임계치 δ는 각각 1과 0.5로 가정하였다.

시뮬레이션의 명령어 코드는 VB 6.0 코드로 933MHz 펜티엄 Ⅲ 컴퓨터에서 수행하였으며 그 결과는 [그림 6]과 같다. 20개의 속성요인이 있는 객체 수를 2,000에서 2,000개씩 증가시켜 20,000×20 매트릭스가 되는 퍼지지식베이스의 정보량까지 크기를 점진적으로 증가시키는 시뮬레이션을 행하였다. 각 객체에 대한 도메인은 0에서 1 사이의 값을 갖도록 난수를 발생하였다. 총

검색시간의 계산은 최종 색인 객체가 검색되는 검색시간의 100번 시행한 결과의 평균으로 하였다.



(그림 6) 제 이론과의 검색시간 비교

(그림 6)에서 나타난 것처럼 데이터베이스 내의 검색 객체의 수가 커짐에 따라, Max/Average 합성연산은 선형적 증가, 가능 및 요인공간이론과 Max/Product 합성연산은 점진적 증가를 하고 있으나, Max/Min 연산은 완만한 증가를 하고 있다. 이는 Max/Product 합성연산이 전역 속성요인을 고려하는 다른 연산 모델과 비교하여 합성연산의 복잡성에 기인한다고 볼 수가 있다. 완만한 증가를 하고 있는 Max/Min합성 연산을 포함한 다른 검색이론과 본 연구모델과 비교하여 보면 본 연구 모델이 여타의 연구이론 검색시간 보다 빠르다는 것을 알 수 있다. 이는 20,000×20 매트릭스가 되는 퍼지지식베이스의 정보량까지 그 크기가 증가하더라도 본 연구에서 제안하고 있는 모델링 기법이 최초의 퍼지지식베이스 스키마의 성질을 보존하면서 도메인 리덕트를 통한 최소지식베이스 검색 공간에서 객체 검색을 수행함으로써 효율적인 정보검색을 위한 기법이라 할 수 있다.

V. 결 론

대용량의 데이터베이스에서 특정 질의어에 대해 전역 속성을 대상으로 정보 검색을 위해 데이터 조작를 가공할 경우에는 때로는 메모리 및 처리 속도로 인해 NP-Complete 문제를 야기할 수 있다. 그래서 정제되고 일관성이 있는 데이터베이스의 구축이 요구되고 자료 조작을 위한 효율적인 모델링 방법도 요구된다. 이런 관점에서 본 연구에서는 퍼지데이터베이스 내에서 특정 질의어에 대한 객체에 대한 속성 요인의 분류를 통한 효율적인 정보 검색 알고리듬을 제안하여 기존 검색이론과 퍼지합성 연산과 특정 객체 검색시간을 비교하였다. 본 연구의 제안 알고리듬은 러프집합 및 퍼지로직에 기초하고 있다. 러프집합 이론의 다중 리덕트의 문제점은 조건 및 결정속성에 대한 상관계수의 공분산에 의해 해결하였다. 데이터 분류시의 경계역 객체의 규칙생성은 베이지안 정리를 적용하여 하한근사 여부를 결정하여 규칙을 생성하였다. 또한 퍼지연산의 모델링은 사분위수에 의한 퍼지 소속함수를 정의하여 퍼지함수로의 표현 및 순환확장 연산을 정의하여 최대 소속함수 원리에 의해 특정 객체를 검색토록 하였다.

본 연구 결과는 RDB, OODB 및 ORDB내에서의 검색 모델링, 패턴 및 영상분류, 정밀제어를 위한 규칙 생성, 웹 상에서의 정보필터링을 통한 실시간 정보 검색엔진의 설계와 지능형 에이전트의 구축에 적용할 수 있으리라 여긴다.

참고문헌

[1] 김형수, 김홍기, 이상부, "확률적 러프이론에 기반한 퍼지정보의 검색", 한국정보과학회

(B), 제 25권, 제 9호, pp. 1431-1441, 1998.

[2] R. G. Cattell, *Object Data Management :Object-Oriented and Extended Relational Database Systems*, Revised Edition, Addsion-Wesley, Reading, MA, 1994.

[3] W. Kim, Introduction to *Object-Oriented Databases*, The MIT-Press, Cambridge, MA, 1990.

[4] L. A. Zadeh, "The role of Fuzzy Logic in the Management of Uncertainty in Expert Systems", FSSII, pp. 199-227, 1983.

[5] N, J. Nilsson, "Probabilistic Logic", Artificial Intelligence 28, pp. 71-87, 1986.

[6] K. A. Anderson, & J. N. Hooker, "Bayesian Logic", Decision Support Systems 11, pp 191-210, 1994.

[7] S. Parsons, "Current Approaches to Handling Imperfect Information in Data and Knowledge Bases", IEEE Transaction Knowledge and Engineering, Vol. 8, NO. 3, pp. 353-372, 1996.

[8] D. Dubois & H. Prade, *Possibility Theory : An Approach to Computerized Processing of Uncertainty*, Plenum Press, New York, 1988.

[9] P. Z. Wang, "A factor space approach to knowledge representation" Fuzzy sets and systems 36, pp. 113-124, 1990

[10] H. X. Li & V. C. Yen, *Fuzzy sets and Fuzzy Decision-Making*, CRC Press. Inc. 1995

[11] A. Motro, "VAGUE : A User Interface to Relational Database that permit vague queries", ACM transaction on Office Information Systems, Vol. 6, No. 3, pp. 187-214, 1988.

[12] Z.Pawlak, Rough sets : *Theoretical Aspects of Reasoning about Data*, A Kluwer Academy Publisher. 1991.

[13] Z. Pawlak, "Rough Sets Present state and Further prospects", Intelligent Automation and Soft Computing, Vol. 2, No. 2, pp. 96-102, 1996.

[14] T. Y. Lin & N. Cercone, *Rough sets and Data mining : Analysis of imprecise data*, Kluwer Academic Publishers, 1997.

[15] G.J.Klir and T.A.Folger, *Fuzzy sets, Uncertainty and Information*, Prentice Hall, New Jersey, pp. 71-94, 1988.

[16] Y. Chien, "A Bayesian model for collaborative filtering", Proc. of Uncertainty in Artificial Intelligence, Morgan-Kaufmann, 1998.

[17] L. Chen & K. Sycara, "WebMate: A personal agent for browsing and searching", Proc, 2nd Int. Conf. on Autonomous Agents and Multi-Agent Systems, pp. 132-139, 1998.

김 형 수

1998년 충북대학교 전자계산학과 졸(이학박사)

1991년 숭실대학교 정보산업과졸 졸(이학석사)

1985년 성균관대학교 정보처리과 졸(경영학석사)

1981년 제주대학교 수학교육과 졸(이학사)

1992년~현재 제주한라대학 컴퓨터정보계열교수

관심분야 : 퍼지 및 러프이론, 인공지능, 멀티미디어컨텐츠, 웹에이전트 시스템