러프셋에 기반한 정보필터링 웹에이전트 모듈 설계

김 형 수* · 이 상 부*
*제주한라대학 컴퓨터정보계열

Design of Web Agents Module for Information Filtering Based on Rough Sets

Hyung-soo Kim* · Sang-bu Lee*
*Jeju Halla College, Division of Computer Information
E-mail : (khs,sblee)@hc.ac.kr

요 약

본 논문은 대용량의 데이터베이스 내에서 유용한 정보를 검색하기 위해 웹 기반하에 적응형 정보 추출 에이전트 모듈 설계이다. 인터넷을 통한 정보 검색이 일반화됨에 따라 검색시간의 최소화를 기하면서 사용자의 요구조건에 맞는 유용한 정보 제공이 필요하다. 구축되는 지식베이스 시스템의 스키마 구성요소의 도메인이 이진 검색이 가능한 필드 도메인이 있는 가하면 그렇지 않은 불확실한 도메인도 존재한다. 최초의 대용량 지식베이스에서 사용자의 자연어 질의어에 대해 러프셋의 리턱트를 통해 최소지식베이스를 생성한 후, 축소된 스키마의 도메인의 불확실성한 값에 대한 연산을 처리는 퍼지합성 연산처리 모듈에 의해 소프팅 컴퓨팅이 수행토록 설계하였다.

ABSTRACT

This paper surveys the design of the adaptive information filtering agents to retrieve the useful information within a large scale database. As the information retrieval through the Internet is generalized, it is necessary to extract the useful information satisfied the user's request condition to reduce the seeking time. For the first, this module is designed by the Rough reduct to generate the reduced minimal knowledge database considered the user's natural query language in a large scale knowledge database, and also it is executed the soft computing by the fuzzy composite processing to operate the uncertain value of the reduced schema domain.

키워드
Information Filtering, Agents, Rough Sets, Attribute Reduct, Fuzzy Composition.

Ⅰ. 서 론

최근 정보통신 기술 발전에 따라 디지털화된 텍스트를 포함한 멀티미디어 정보로 인해 데이터베이스의 양이 대용량화되고 복잡한 구조를 갖고 있다. 저장되는 자료의 분류, 캡슐화, 상속,병행성 및 스키마 진화에 따른 적절한 데이터 모델을 고려한 논리적인 처리구조의 지식베이스의 생성이무엇보다도 필요하다. 실제 지식베이스의 스칼라도메인은 불린로직에 기초하나, 애매하고 모호한 정보 및 질의에 대한 정보처리는 퍼지로직(fuzzy logic),확률로직(probabilistic logic)및 베지안로직(bayesian logic)에 기반하여 근사적 추론에 의거하여 최적의 튜플등을 추출하였다. 특히, 러프이론(rough sets theory)]은 불확실한 결정속성에 대한 조건 속성의 하한근사(lower approximation)및 상한근사(upper approximation)의 객체분류에 의해 도메인내의 숨겨진 패턴(hidden pattern)의 속성감축(attribute reduct)을 최소지식베이스 생성과 데이터의 중요도 평가(significance evaluation)를 통해 대용량 데이터베이스에서 검색시간의 효율성을 제공한다.

인터넷 사용이 일반화됨에 따라 웹문서의 정보검색은 기계학습, 지식관리, 에이전트, 자료융합, 정보수집 등의 기술과 결합하여 적응가능 학습을 고려하여 정보 추출을 한다[1]. 네트워크 상에서 사용자와 상호 인터페이스를 통해 사용자의 선호 요인을 반복·학습하여 그 결과를 활용하는 귀납적 기계학습(inductive machine learning) 방법을 적용한다. 에이전트(agents)를 이용한 학습은 사용자 행위에 관한 관찰 학습, 사용자의 피드백에 의한

학습, 훈련을 통한 학습과 에이전트간의 도움(advice)을 통한 학습방법에 의해서 필요한 정보를 추출하게 된다[2]. 결과적으로 사용자의 취향을 고려하여 사용자의 만족도를 극대화 할 수 있는 적응성 에이전트(adaptive agents)기능에 의거하여 정보추출(information filtering)이 이루어지게 된다.

이런 관점에서 본 연구는 사용자 인터페이스를 통한 검색 질의어에 대한 속성요인의 분류를 하여 러프이론의 리덕트(reduct)에 의거한 결정규칙의 래퍼(wrapper)의 최소지식베이스를 생성하고, 퍼지합성 연산에 의해 특정 객체를 추출하는 적응성 웹에이전트 모듈을 제안한다.

II. 지식베이스 생성

2.1 러프셋 개념

최초의 데이터베이스는 공집합이 아닌 객체집합 U가 관계 R에 의해 정형화된 속성요인으로 특징화된다. 관계 R은 $U(R=U \times U)$상에서 반사(reflexive),대칭(symmetric) 및 추이(transitive)관계를 만족하는 이항 동치관계(binary equivalence relation)를 만족하면서 근사공간 $Apr = (U,R)$로 U,R의 순서쌍으로 구성한다.

근사공간 $Apr = (U,R)$에서 전체 집합 U는 $[U_i]_R$로 표현하는 기본집합들의 구분 불가능한 관계인 동치류(*equivalence class*)로 분할 할 수 있다. 임의의 객체 $U_i \subseteq U$에 대한 관계 R의 동치류 $[U_i]_R$는 지식베이스를 구축함에 있어 가장 기본적인 개념 블록이 된다. 결정속성에 대한 조건속성의 객체의 분류인 동치류는 서로 다른 결정속성에 속하는 불일치성(inconsistency)을 갖게 된다. 러프셋(rough sets)은 객체 속성간의 불일치성을 분류하는 동치류의 연산관계를 나타내는 $([U_i]R, \cap, \cup, \sim, Apr_*(X), Apr^*(X))$요소로 정의된다[5]

여기서 $Apr_*(X)$, $Apr^*(X)$는 근사공간 $Apr = (U,R)$에서 $X \subseteq U$ 이라 할 때의 X의 하한근사(lower approximation), 상한근사(upper approximation)로 써 X의 부분집합인 모든 기본집합의 합집합, X와 비 공집합을 갖는 모든 기본집합의 합집합을 각각 나타내어 다음과 같이 수식화 된다.

$$Apr_*(X) = \{ U_i \in U \mid [U_i]_R \subseteq X \}$$
$$Apr^*(X) = \{ U_i \in U \mid [U_i]_R \cap X \neq \emptyset \}$$

2.2 지식베이스 표현

의사결정을 수행하기 위한 지식베이스는 사용자의 요구 및 선호도를 고려하여 시스템의 사용자 인터페이스를 통해 필요한 정보들의 생성규칙들의 래퍼(wrapper)모임으로써 최초 데이터베이스에서 정제된 다음의 요소로 구성된다.

$$S = \{ U, C, D, V, f \}$$

(단, 객체집합 $U = \{U_i \mid i = 1, \cdots n\}$, 조건 속성 $C = \{C_i \mid i=1, \cdots, n\}$, 결정속성 $D = \{D_i \mid i=1, \cdots n\}$, V : 유한 퍼지속성 도메인, f : 정보함수)

여기서 U,C,D는 비공집합(non-empty)으로 U에 속한 모든 객체는 조건속성 C와 결정속성 D에 대한 값의 집합과 연관되며 $A = C \cup D$는 모든 속성 집합, $C \cap D = \emptyset$이다. $V = \bigcup_{a \in A} V_a$으로 V_a는 속성 a의 값이며 스칼라 또는 측정가능 요인(measurable factor)으로 구간 길이 방법(interval width method)에 의거하여 속성 값을 분류할 수 있다. $V = \bigcup_{a \in A} V_a$ 이며, V_a 는 구간길이방법에 의해 분류되는 속성 $a \in A$ 대한 퍼지 유한속성 도메인(fuzzy finite attribute domain)이다. f 는 모든 $a \in A$ 와 $U_i \in U$ 대한 $f : U \times A \rightarrow V$에 대응되는 $f(U_i, a) \in V_a$ 정보함수이다.

2.3 속성분류

특정객체를 검색하기 위한 근사공간 $Apr = (U,R)$은 사용자의 요구 조건의 질의어에 의한 객체의 하한 및 상한 근사 정의에 기초하여 양역(positive region) POS(X), 음역(negative region) NEG(X)과 경계역(boundary region) BND(X)의 동치류의 3종류의 속성영역으로 분류할 수 있다. 각 속성영역을 러프셋에 의거하여 분류하여 정의하면 POS(X) = Apr* (X), NEG(X) = U - Apr*(X), BND(X) = Apr*(X) - Apr*(X), Apr*(X) = POS(X) ∪ BND(X)으로 정의할 수 있다. $\forall x \in$ POS(X)는 반드시 X에 속하나 $x \in$ NEG(X)인 경우는 X에 속하지 않음을 나타낸다.

특히, 경계역 $x \in$ BND(X)인 경우에 있어서는 x가 X에 속하는지 여부는 확실하게 단정할 수 없다. 그러나 BND(X) = ∅이면 X는 명확한 이진검색을 통한 객체 검색이 이루워지나, BND(X) ≠ ∅인 경우에 있어서는 X는 근사 결정규칙을 생성하여 객체를 검색하기 위해 확률적 러프셋(probabilistic rough sets)을 적용한다[5].

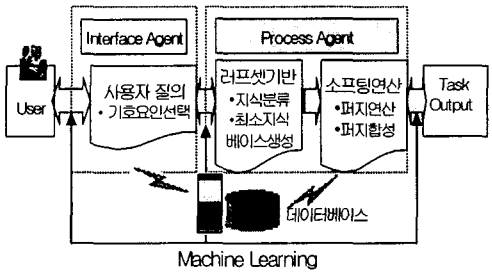
본 연구에서는 이러한 경계역 객체에 대한 동치류의 분류는 베지안정리(Bayesian theorem)를 적용, δ-절단 집합에 의거하여 근사결정규칙의 객체 검색공간의 최소지식베이스를 생성한다.

III. 에이전트 모듈설계

3.1 제안 에이전트 구조

인터넷 상에서 정보처리를 수행하는 에이전트는 각종 정보의 검색, 필터링, 추출 및 통합 기능의 에이전트로 대별된다. 사이트마다 html,xml의 웹문서의 표현 구조의 이질성과 가변성, 자연어 처리의 제한성으로 인해 정보 시스템의 확장성을 고려하여 HLRT[6],WHIRL[7]에서 도입한 것처럼 질의 인터페이스(query interface) 및 래퍼귀납

(wrapper induction)의 기계학습을 통한 자동처리 프로그래밍 기법을 채용하여 정보가공을 수행한다. 본 연구의 정보필터링 웹에이전트의 구조는 [그림 1]과 같다. 웹환경하에서 사용자의 질의영역으로써 기호요인을 입력하는 인터페이스 에이전트(interface agents)와 입력정보를 데이터베이스와 연동하여 러프셋과 퍼지이론에 의거한 소프팅 연산을 수행하는 프로세스 에이전트(process agents)로 구성, 학습에 의한 반응의 연속적인 기계학습을 수행하여 객체를 추출토록 설계하였다.



[그림 1] 제안 웹에이전트 설계

3.2 에이전트별 기능

3.2.1 사용자 질의모드

웹 상에서의 자료의 양이 기하급수적으로 늘어남에 따라 데이터베이스의 양이 대용량화되고, 효율적인 정보검색을 위한 도구 역시, html인식,언어처리기반, 온토로지(ontology)기반, 레퍼추론 및 모델링기반 도구등 다양하다. 특히 자연어처리기반 도구는 기구축된 웹사이트에서 자연어로 작성된 문서에서 관련된 자료를 추출하기 위한 구문 및 의미론적 제약조건으로 학습될 추출규칙이 고려되어야 한다. 또한 퍼지언어 요소에 대한 정형화도 이루어져야 한다. 본 연구에서의 사용자인터페이스 에이전트의 질의어모드(q)는 다음과 같다.

```
RETRIEVE object_name
FROM DB
WHERE c_fac=C and d_fac=D (dec_value=a)
```

이 모드는 $P[Ui \mid q]a$ 형 결정임계치 a와 질의어의 선택절 q 를 만족하는 Ui 객체를 검색하라는 의미이다. 질의어모드(q)는 선택절(select clause)로써 퍼지 언어변수(fuz_lin)와 퍼지 변형자(fuz_mod)를 사용하여 아래 형식과 같이 러프셋을 적용하기 위한 DB 내의 k 개 조건속성 및 결정속성의 결합이다.

```
<select clause>::=
        ([c_att] <fuz_mod><fuz_lin>)^k
      ∧[d_att] <fuz_mod><fuz_lin>)^k
```

3.2.2 결정규칙 추출

사용자 인터페이스 에이전트의 질의어 모드에 따라 웹문서들의 형태소 분석에 의거하여 생성되는 레퍼규칙의 모임으로써 최초의 지식베이스 $S = \{ U , C , D , V , f \}$는 방대하며 결정속성(D)에 따른 여분(superfluous)의 조건속성(C)이 존재한다. 따라서 이러한 검색하고자 하는 질의어에 따라 여분 또는 중복되는 속성을 제거하여 슬립화된 지식베이스를 속성리덕트(attribute reduct)에 통해 생성한다. 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 기본집합의 하한근사의 합집합 $POS_c(D) = \cup_{x \in D^*} C_*(X)$ 일 때, 속성요인 a 의 여분속성 여부결정은 아래와 같다.

여분속성을 제거하는 속성리덕트의 결과인 $RED_D(C)$는 최소 데이터 탐색공간인 감소 정보시스템으로서, 주어진 최초 정보시스템의 데이터의 특성과 패턴이 그대로 보존된다.

```
Attribute reduct{}
 Decision_Redundant
   {if POS_c(D) = POS_(c-{a})(D) then
    a : D-superfluous attribute(for all a∈C);
        else
    a : D-indispensable attribute(for all a∈C);}
 Get_Base
   {Subset C'⊆C, C' is attribute reduct iff
       C' is D-indispensable attribute and
       POS_c(D) = POS_c'(D);}
```

3.2.3 근사추론 합성

웹 문서의 어휘 및 문장분석을 통해 생성된 속성요인의 값에 따라 객체에 대한 상·하한 근사공간으로 분류하여 속성리덕트를 수행한 결과 축소된(reduced) 결정규칙(R)의 지식베이스를 생성한다. 이 결정규칙은 하한근사의 부분(partial) 결정규칙과 경계역 $BND_C(D)$의 부분가능(partial possibly) 결정규칙으로 구성되며 조건속성 C_i $(i=1,....,n)$ 와 결정속성 D의 도메인 V_{ii}의 요인곱의 결합이다.

$$R:(C1=V_{i1}) \wedge (C2=V_{i2}) \wedge .... \wedge (Cn = V_{in}) \rightarrow (D = V_{ii})$$

조건속성의 분류 결과 $E_i(C)$에 있는 객체가 때로는 서로 다른 결정속성 분할 $E_j(D)$에 속할 경우, 어느 결정영역에도 속할 수 없는 결정모순(decision conflict)에 빠지게 된다. 퍼지질의어에 대한 근사추론을 행하기 위해서는 결정모순의 상한근사역에 있는 객체를 근사적으로 검색할 수 있는 도메인리덕트가 필요하다.

본 연구에서의 도메인리덕트는의 확률적 러프셋에 기초하여, 다음 $P[j \mid E_i(C)]$ 조건부확률인 베이지안 정리를 적용하여 결정속성을 생성한다.

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

여기서, Q_j는 최초의 정보시스템 *FKS*의 전체 객체에서 각각의 결정 분할 $E_j(D)$ 객체가 존재할 확률이다. 조건부 확률 $P[E_i(C) \mid j]$는 경계역 $BND_C(D)$에 있는 임의의 조건속성 분류 $E_i(C)$의 객체가 각각의 결정속성 분할 $E_j(D)$에서 발견될 확률이다. 또한, $P[j \mid E_i(C)]$는 조건속성의 분류 $E_i(C)$의 객체가 서로 다른 결정속성 분할 $E_j(D)$에서 발견할 확률이다. 최종적으로 생성된 생성된 최소지식베이스 내의 속성 도메인의 퍼지 언어의 소속함수의 값는 좌삼각퍼지수(left triangular fuzzy number)이며 max/min 합성 연산에 의한 [0,1]의 퍼지수로 정량화 한다.

3.3 정보추출 연산

가변적이고 비정형화된 웹문서에서 유용한 정보의 획득하기 위해서는 사용자의 자연어 질의에 따른 소위 의미론적 웹(semantic web)과 온톨로지(ontology) 기능의 지식표현과 계층구조의 생성이 필요한다. 그러나 웹상에서 자동적으로 단시간에 유용한 정보를 제공하기 위한 정보검색 연구가 진행되고 있지만, 대부분의 자연어 질의에 형태소분석, 품사태킹, 구문 및 의미분석의 자동 생성의 어려움이 산재하고 있다. 또한 생성된 지식베이스에서 적절한 추론 기법의 정형화의 어려움이 있다. 특히 웹 정보추출의 다양성과 가변성으로 인해 래핑기술에 의한 효율적인 래퍼관리가 요구되며 자동및 반자동 프로그램으로 관리된다[8].

```
Schema1{}
 Query_rule{
   get()[ method:post;
          url:" http://www.jejupansion.co.kr";
          param:"for"="input condition & decision attribute";
                "type"="main text"; ]}
   generate_knowledgebase{
          attribute_matching=describe condition description ;}
          program coding{.....}

Schema2{}
 Attribute_reduct{
  Classification(){classify  Ei(C) ;  U/C = ∪ Ei(C) ;
  partition   Ej (D) ;  U/D = ∪ Ej (D) ;}
  Decision(){
     For i=1 to n
       R = Ø ;
       For j=1 to m
     If  Ei(C)⊂Ej (D) then { Ei(C)⊂POS(D);Ri =R∪Ri ;
     else If Ei(C)⊂Ej(D) then   Ei(C)⊂BND(D;
             else Ei(C)⊂NEG(D;}
  Boundary(){
   select RND(Ei(C));
   set   δ ;    /* δ : decision value */
   calculate Qj w.r.t Ej (D) & P [Ei(C)  | j] in Ej (D) ;
   calculate probability P[j  | Ei(C)];
   decide wi^j = Max P[j  | Ei(C)];
   If  α ≤ wi^j  then  {Ei(C) ⊂Ej (D) ; Ri =R∪Ri ;
   Convert Fuzzy expansion  w.r.t  α;
   Calculate m-ary Triangle norm Tm ;
Retrieve Max Ui by 1st Maximum membership principle;}
```

[그림 2] 정보추출 제안 알고리듬

본 고에서는 [그림 2]와 같이 특정 URL에 제한된 html인식 W4F[9]의 레퍼규칙 생성기법에 의거하여 지식베이스의 생성과 러프셋에 근사추론 알고리듬에 의거한 객체추출 알고리듬을 제안한다. Schema1{}은 사용자 인터페이스 에이전트의 질의모드로 검색하고자 하는 URL에서 조건 및 결정속성에 대한 검색조건의 결정규칙의 지식베이스의 생성단계이다. Schema2{}는 최소지식베이스를 생성하기 위한 속성리덕트, 근사추론 및 퍼지 합성에 의거한 객체 검색과정으로 제안된 알고리듬에 의해 생성된 지식베이스에서 특정 객체를 검색하기 위해 소속함수 및 t-노름의 퍼지 연산을 정의하여 근사적 추론을 행하여 제 1 최대원리(first maximum membership principle)에 의거하여 특정 객체를 결정하였다

IV. 결 론

인터넷을 활용한 정보검색이 일반화된 요즈음 방대한 웹데이터베이스 내에서 자연어 질의어에 의한 검색을 통해 유용한 정보의 최단 시간내 획득이 요구되어진다. 그러나 웹문서의 가변성과 이질성, 의미분석의 모호성으로 인해 문서의 구조적 체계화에 어려움이 있다. 최근 웹상에서의 정보통합과 수집, 텍스트마이닝 및 적응성 정보추출 기법이 시맨틱 웹에서의 온톨로지 활용기술에 대한 연구가 활발하나 실험적 수준이라 여길 수 있다.. 본 연구에서는 제한적이지만 모호한 정보의 처리를 위한 러프셋에 기반하여 html인식도구인 W4F에 의거한 래퍼 생성규칙과 근사추론 정보추출 알고리듬을 제안하였다. 본 알고리듬은 자연어 질의에 대한 웹 데이터베이스의 구축에 논리적인 구조의 정형화에 참고가 되리라 여긴다. 또한 20,000×10 매트릭스 정적데이터베이스에서 근사추론 알고리듬의 정보검색의 효율성을 시뮬레이션을 통해 제시하였던 바[10], 본 연구에서 제안 알고리듬을 웹 환경의 가변적 변화에 따른 적응성 에이전트를 탑재한 시스템 구축이 과제로 남겨둔다.

참고문헌

[1] N.Kushmerick and B.Thomas, "Adaptive information extraction:Core technologies for information agents: An AgentLink perspective",Lecture Notes in Computer Science 2586, 2003.

[2] P.MAES,"Agent that reduce work and information overload", Communications of the ACM, Vol.37,No.7, 1994.

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

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

[5] W.Ziarko.& N.Shan,"Knowledge discovery as a search for classification, Workshop on Rough sets and Database Mining ,23rd Annual Computer Science, CSC'95.

[6] N.Kushmerick, D.Weld, R.Doorenbos," Wrapper Induction for Information Extraction", International Joint Conference on Artificial Intelligence,pp.729-735, 1997.

[7] W.Cohen, "A Web based Information System that Reasions with Structured Collections of Text: 2nd International Conference on Autonomous Agents,pp.400-407, 1998.

[8] K.Lerman, S.N,Minton, C.A.Knoblock, "Wrapper maintenance:A Machine Learning Approach",Journal of A.I Research,Vol18, pp.149-181, 2003.

[9] A.Sahuguet,F.Azavant,"Building intelligent Web Applications using Lightweight Wrapper", Data Knowledge Engineering.36(3), pp.283-316, 2001.

[10] K.Hyuns-soo, K.Hong_gi, "Algorithm for Knowledge Discovery based on Data Classification and Approximate Inference", KFIS, Vol2, No.2, pp.27-32, 2001.