Journal of Korea Multimedia Society Vol. 11, No. 11, November 2008(pp. 1566-1574)

온톨로지 기반에서 연관 마이닝 방법을 이용한 지식 추론 알고리즘 연구

황현숙[†], 이준연[††]

요 약

정보 검색에 대한 연구는 방대한 데이터에서 원하는 검색 정보를 제공할 뿐 만 아니라 개인의 취향에 따른 맞춤 검색 및 추론된 지식을 제공하는 데 초점을 두고 있다. 본 논문의 목적은 데이터를 개념화하여 분류 및 정의할 수 있는 온톨로지 구조를 기반으로 숨어있는 지식을 발견하여 개인 맞춤 검색을 제공하는 추론 알고리즘에 대해 연구하는 것이다. 현재의 검색에서는 방대한 데이터에서 너무 많은 검색 결과를 제공하거나 검색 결과를 제공하지 못하는 경우도 발생하고 있다. 이러한 정보 검색의 단점을 보완하기 위해 OWL 온톨로지 제약조건과 연관 마이닝 방법으로 추론된 연관 지식을 SWRL 추론 언어로 표현하여 Jess 엔진을 통한 새로운 지식을 발견하여 효율적인 검색을 지원하는 알고리즘을 제안한다. 식당, 주유소, 제과점 등의 도메인에 따른 개인별 선호 온톨로지를 구축하고, 주유소 개인 선호 데이터를 예제로 하여 연관 및 온톨리지 기반에서 정보를 검색할 때, 연관 및 추론 정보를 제공함을 보여준다.

A Study of a Knowledge Inference Algorithm using an Association Mining Method based on Ontologies

Hyun Suk Hwang[†], Jun Yeon Lee[††]

ABSTRACT

Researches of current information searching focus on providing personalized results as well as matching needed queries in an enormous amount of information. This paper aims at discovering hidden knowledge to provide personalized and inferred search results based on the ontology with categorized concepts and relations among data. The current searching occasionally presents too much redundant information or offers no matching results from large volumes of data. To lessen this disadvantages in the information searching, we propose an inference algorithm that supports associated and inferred searching through the Jess engine based on the OWL ontology constraints and knowledge expressed by SWRL with association rules. After constructing the personalized preference ontology for domains such as restaurants, gas stations, bakeries, and so on, it shows that new knowledge information generated from the ontology and the rules is provided with an example of the domain of gas stations.

Key words: Ontology(온톨로지), Association Mining(연관마이닝), Inference Algorithm(추론알고리즘), User Profile Modeling(사용자 프로파일 모델링), Databases(데이터베이스)

※ 교신저자(Corresponding Author) : 황현숙, 주소 : 부산광역시 남구 대연3동(608-737), 전화 : 051)629-6245, FAX : 051)629-6245, E-mail : hhs@pknu.ac.kr
접수일 : 2008년 6월 18일, 완료일 : 2008년 9월 18일
[†] 정회원, 부경대학교 연구원

[††] 동명대학교 멀티미디어 공학과
(E-mail : jylee@tu.ac.kr)
※ 이 논문은 2006년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(KRF-2006-351-B00013).

1. 서 론

웹과 비즈니스 어플리케이션에서 대부분 데이터베이스를 사용하여 단순 키워드 검색으로 정보를 제공하고 있다. 하지만, 방대한 데이터에서 단순한 키워드 검색으로는 너무 많거나 빈약한 검색 결과를 제공하여 원하는 정보를 발견하는 데 많은 노력과 시간이 필요할 뿐만 아니라 정확한 검색 결과를 제공하기 어려운 단점이 발생하고 있다[1,2]. 즉, 웹 검색에서 클래식 음악회 공연을 검색할 경우, 일단 문화회관 홈페이지를 찾은 후 월별 행사를 검색하거나, 클래식 음악회의 키워드로 검색한 후 검색 결과를 링크해서 정보를 찾게 된다. 즉, "이달의 클래식 음악회 공연은 어떤 종류가 있는가?" 라는 검색을 할 경우, 사용자는 여러 번의 검색을 수행한 후 결과를 발견할 수 있게 된다. 또한, 이러한 방법은 도메인간의 연관 관계를 제공하고 있지 않기 때문에 단순 검색을 수행한 후 사용자가 결과를 분석해야 하는 번거러움이 있다.

그러나 시맨틱 웹의 등장으로 방대한 웹 데이터를 사람뿐만 아니라 컴퓨터 또한 스스로 정보를 이해할 수 있도록 함으로써 검색 작업을 보다 자동적으로 수행할 수 있는 환경으로 변화하고 있다[3]. 이러한 시맨틱 웹에서의 검색은 기존 정보 제공뿐 만 아니라 정보간의 연관성을 발견하여 연관된 정보 및 추론된 정보를 제공할 수 있다. 이러한 의미 기반의 검색을 수행하기 위해서는 데이터의 지식을 표현하기 위해 공통성이 있는 데이터를 개념화하여 분류 및 정의할 수 있는 온톨로지 기반의 데이터 구조를 가져야 한다[4,5]. 웹 검색의 새로운 기술로 등장하고 있는 온톨로지 데이터 기반에서의 데이터 검색 시스템과 효과성에 대한 연구가 진행되고 있다[1,5-8].

현재 온톨로지 기반의 검색에 대한 연구는 연관 및 추론된 지식 제공에 초점을 맞추고 있으며 온톨로지와 규칙을 통합하는 연구[1,2,9,10]가 대부분이고 이러한 추론에 필요한 규칙은 대부분 실제 응용도메인에서 경험적인 분석으로 사용자가 임의로 정의하고 있다. 본 논문에서는 OWL의 개념화된 추론 뿐만아니라 연관 마이닝 기술을 이용하여 사용할 연관 규칙을 도출한 후 이를 결합하여 새로운 지식을 생성하는 OWL과 연관 마이닝을 결합한 추론 알고리즘을 제시한다. 연관 마이닝 방법은 데이터간의 지지도와 신뢰도를 기반으로 규칙을 발견하는 연관 규칙 기법이 데이터간의 분석에 활용되고 있다[11,12].

본 논문의 구성으로는 2장에서 연관 규칙과 온톨로지 기법에 대한 관련 연구를 기술한다. 3장에서는 OWL과 연관 규칙을 접목한 지식 추론 알고리즘을 제안하며, 4장에서는 제시한 알고리즘을 실생활에 응용할 수 있는 예제로서 사용자 선호 취향 데이터를 사용하여 온톨로지를 구축한 후 연관 마이닝을 이용하여 추출된 지식과 접목하여 새로운 지식을 생성하는 과정을 나타내는 선호기반 지식 검색 시스템을 제시한다. 마지막으로 결론 및 향후 연구 과제를 제시한다.

2. 관련연구

2.1 연관 규칙

데이터 마이닝은 다양한 데이터에 대해 논리적인 구성이 가능하고 유용할 것으로 생각되는 숨겨진 패턴을 발견하기 위한 과정이다[13,14]. Holsheimer [14]는 데이터 마이닝을 다양한 테이터에서 숨은 패턴을 찾아 기업의 의사결정에 도움을 주고 나아가 이러한 결정에 대한 결과를 예측하도록 지원하는 것이라고 하였다. 특히, 마이닝 기법 중에서 연관 규칙은 데이터베이스의 트랜잭션에서 항목 간에 발생하는 규칙을 표현하는 것으로 1993년 처음 소개되었다[11]. 이는 어떤 사건이 발생할 때 그 다음 사건의 관련성을 의미하는 것으로 X=>Y 규칙의 형태로 표현된다. 이러한 연관 규칙의 타당성은 지지도와 신뢰도 척도를 이용하여 판단된다. 지지도는 전체 트랜잭션에 대해 트랜잭션 항목 집합이 차지하는 비율이고 신뢰도는 조건부 트랜잭션 항목 집합에 대해 규칙에 포함되는 모든 항목 집합이 차지하는 비율을 의미한다.

연관 알고리즘 연구로 1993년 Agrawal et. al.[11]은 장바구니 데이터를 대상으로 고객이 구매한 상품간에 연관성이 있는 집합을 발견하는 Association Item Sets(AIS) 알고리즘을 제안하였다. 이는 전체 데이터베이스를 검색하여 최소한의 트랜잭션 개수를 가지는 후보 항목 집합을 발견하였다. 1994년 Agrawal and Srikant[12]는 이전 단계의 빈발 함수 집합을 이용하여 후보 집합을 생성하는 Apriori 알고리즘을 제안하였다. Apriori 알고리즘은 첫 번째 단

계에서는 최소 지지도 이상을 갖는 빈발 항목 집합을 발견하고, 두 번째 단계에서 발견한 빈발 항목 집합의 모든 부분집합을 생성하여 최소 신뢰도 이상인 규칙을 발견한다. 이 알고리즘은 다른 알고리즘에 비해 수행 성능이 좋은 것으로 평가 받아 여러 분야에서 기본 알고리즘으로 활용되고 있다.

2.2 온톨로지 기술

온톨로지는 웹과 제공자 등에 의해 산재되어 있는 데이터를 분석하고 통합하기 위한 데이터 표현 방법으로 이를 공유된 개념화에 대해 정형화되고 명시적인 명세라고 정의한다[4,5].

RDFS(Resource Description Framework Schema Specification)는 온톨로지를 나타내기 위한 가장 기본이 되는 언어로서 RDF 응용 언어를 사용하기 위한 기본적인 어휘를 정의해 두고 있다. 여기서 RDF (Resource Description Framework)는 데이터 모델에서 모든 종류의 정보와 메타데이터를 표시할 수 있는 유용한 언어로써 자원과 문장으로 구성되어 있다. RDF 문장은 subject, predicate(property), object(value) 의 3개의 부분으로 구성되고 이는 특정 주제에 대해 정보를 나타내기 위한 문장으로 표현된다. RDF 구조가 메타 데이터를 표시하는 언어인 XML과의 다른 점은 데이터 간의 의미를 표시할 수 있다는 것이다. 즉 자원과 자원을 연결하여 속성으로써 이들 관계를 연관시킬 수 있다. 이러한 RDF는 컨텐츠를 표시하기 위해서 XML 메타 데이터를 기반으로 하고 있다. RDF는 언어라고 알려져 있지만 일종의 데이터 모델로서 역할을 하고 있다. 따라서 RDFS는 RDF를 포함하고 있는 위치로서 RDF에서 사용되는 자원들의 관계를 나타내고 있는 언어이다.

OWL(Web Ontology Language)은 RDFS와 같은 초기의 온톨로지 언어를 기본으로 작성되고 있다. OWL은 시맨틱 웹에서 온톨로지를 정의하기 위해 W3C에 의해 제안된 표준 온톨로지 언어이다. 서로 다른 형식으로 표현되어 있는 온톨로지간의 정보를 공유하기 위해 W3C에서 온톨로지 표기를 위한 표준 언어로서 정의하고 있다. OWL은 온톨로지의 계층 구조 표현을 위해서 객체 클래스와 클래스간의 상하위 계층 관계를 표현할 수 있으며 사용자가 임의의 속성을 정의하고 정의한 속성을 클래스에 할당할 수 있다. RDFS 보다는 좀 더 표준화 되어 있고 객체간의 관계에서 논리적인 유추가 가능하다. 이는 OWL이 객체와 속성을 정의하고 이에 적용할 수 있는 제약조건들을 포함하고 있기 때문이다. RDFS와 마찬가지로 OWL 또한 RDF 문장으로 정보를 표현할 수 있다.

Protege 시스템은 지식 기반 시스템을 구축하기 위한 환경으로 의학 분야를 도메인으로 지식 기반 응용 시스템 개발로 시작하여 현재는 OWL 온톨로지를 비롯하여 더욱 더 범용적인 범위로 확대되어 지식 공학의 툴로써 개발되고 있다[15]. Protégé/OWL는 Racer, Jess(Java Expert System Shell) 등의 추론 엔진과 통합을 지원하고 있고 시맨틱 웹의 규칙 언어인 SWRL과의 통합으로 분석적, 의미적, 추론적인 결과를 가져올 수 있는 통합 환경을 제공하고 있다[16].

3. 연관규칙과 온톨로지 기반의 지식 추론 알고리즘

연관 규칙을 이용하여 온톨로지 기반에서 새로운 지식을 생성하기 위한 알고리즘을 제시한다. 그림 1은 임의의 도메인에서 속성들의 선호 빈도수를 이용하여

```
Algorithm Association-based Inference Search
n : the total number of attributes
k : index for each step
Domain(i, j) : j-th attributes for domain i
Ontology(i) : the defined ontology of the domain
RDF(i, j) : j-th individuals for the defined ontology
RuleSet(Rk) : k-th Rule sets generated from Rk
MINk(support): k-th minimum support
MINk(confidence): k-th minimum confidence
Rk : rule sets with MINk(support) and MINk(confidence)
Create RDF contents from the defined Ontology(i)
For(k=1, n; k++) do begin
  define MINk(support), MINk(confidence);
  Rk=Association_RuleItem();
End
Create RuleSet(Rk) and modify Ontology
        based on the SWRL language
Infer new knowledge and modify RDF contents
        from the RuleSet(Rk)
For(j=1, n; j++) do begin
    Set user's request values for each attribute j
End
Execute the inference search with the generated RDF
Return results of inferred search
```

그림 1. 연관 규칙 기반 지식 추론 알고리즘

연관 규칙을 발견한 후 SWRL 사용하여 규칙을 생성한 후 추론 엔진을 이용하여 새로운 지식을 추론하는 전체 과정을 나타낸 것이다. 우선, 온톨로지와 RDF 개인 프로파일을 작성하고 개인 선호 데이터베이스 파일을 사용하여 연관집합을 추출한다. 다음, SWRL 언어를 사용하여 추출된 규칙을 변환하여 온톨로지의 제약조건으로 삽입한다. 이러한 규칙을 이용하여 추론을 수행한 후 RDF 파일을 갱신함으로써 새로운 지식이 생성된다. 사용자는 새롭게 갱신된 RDF 프로파일을 이용하여 추론검색을 수행할 수 있다.

그림 2는 후보 집합, 빈발 집합을 생성한 후 연관 규칙 집합을 생성하는 전체 과정이다. 첫 번째는 후보 항목 집합을 생성하는 Candidate_Item() 함수로 이에 대한 결과는 C_k에 테이블에 저장된다. 두 번째는 후보 항목 집합인 C_k에 대해 최소 지지도 이상을 가지는 빈발 항목 집합을 생성하는 Frequent_Item() 함수로서 이의 결과는 F_k에 저장된다. 마지막으로 빈발 항목 집합인 F_k에 대해 각 단계에서 최소 신뢰도 이상을 가지는 연관 규칙 집합을 생성하는 Rule_Item() 함수로서 이의 결과는 F_k에 저장된다.

그림 3은 후보, 빈발, 규칙 집합을 생성하기 위한 세부 알고리즘이다. C_k는 후보 항목 집합을 생성하는 함수로서, 속성이 n개일 경우, 2개의 속성 집합부터 n개까지의 속성 집합을 추출한다. 각 k 단계 속성 집합별로 데이터베이스에 있는 n개의 트랜잭션까지 검색 후 후보 항목 집합을 생성한다. 이때, n개의 항목에서 k개의 항목 필드는 인덱스를 변경하면서 그룹 연산자 (group by)와 카운트 함수를 사용하여 추출한다. 두 번째로 F_k는 후보 항목 집합인 C_k에 대해 최소 지지도 이상을 가지는 빈발 항목 집합을 생성하는 함

```
Algorithm Association_RuleItem()
trans_tbl : a transaction table of Domain(i, j)
tr_cnt : the number of the transaction
Ck : k-th candidate sets with k items and count field
Fk : k-th frequent sets with k items and count field
For (k=1; n; k++) do
        Ck=Candidate_Item(trans_tbl, k, n)
        Fk=Frequent_Item(Ck, k, n);
End
For (k=2; Fk≠ø ; k++) do
        Rk=Rule_Item(Fk, tr_cnt);
End
```

그림 2. 연관 규칙 생성 알고리즘

```
Algorithm Candidate_Item(trans_tbl, k, n)
For(i1=1; n-(k-1); i1++) do begin
  for(i2=i1+1; n-(k-2); i2++) do begin
                ... ... ... ...
  for(ik=ik+1; n; ik++) do begin
   insert into Ck
    select Item1, Item2, Item3, ··· , Itemk, count(*)
      from trans_tbl group by Item1,Item2,Item3,···,
Itemk
End
Algorithm Frequent_Item(Ck, k, n)
insert into Fk
  select C.Item1,C.Item2, ···,C.Itemk,C.cnt from Ck as C,
         Fk-1 I1, ···,Fk-1 Ik
  where C.cnt ≥ Mink(support)
END
Algorithm Rule_Item(Fk, tr_cnt)
insert into Rk
  select item1,item2, ···,itemi as "itemi =>", itemi+1, ··· ,
     itemk, cnt/tr_count as support,
     cnt/(select cnt from fi where
       fi.item1=fk.item1, ···, fi.itemi=fk.itemi)
            as confidence from fk
  where confidence ≥ MINk(confidence)
END
```

그림 3. 후보,빈발,규칙 항목 집합 생성 알고리즘

수로서 각 단계별 속성집합을 생성한다. 세 번째로 빈발 항목 집합인 F_k에 대해 각 단계별로 최소 신뢰도 이상을 가지는 규칙 항목 집합 R_k를 생성한다.

4. 개인 선호 기반 지식 검색 시스템

4.1 사용자 선호 기반 검색 시스템 구성

그림 4는 도메인별로 구축되어 있는 데이터베이스에서 선호 테이블을 작성하여 연관 마이닝 엔진을 통해 연관 규칙 생성하고, 선호 온톨로지를 작성하고 지식을 추론하여 지식 기반 검색 결과를 제공하는 시스템 구조이다. 이 검색 시스템은 도메인별로 개인이 선호하는 정보를 등록한 후, 연관 및 온톨로지를 이용하여 지식을 추론하여 축약되고 연관된 검색을 지원한다. 여기에서 도메인이란 주유소, 음식점, 제과점 등이 될 수 있다.

각 도메인별로 구축되어 있는 데이터베이스에서 관련 도메인에서 사용자의 선호 속성 정보를 저장하기 위한 테이블 구조를 생성한다. 선호 테이블 구조는 한 개 또는 다수의 도메인을 이용하여 작성될 수

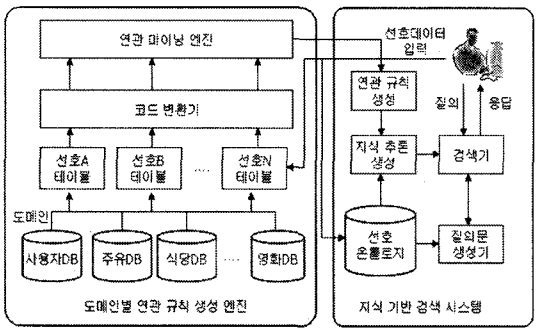


그림 4. 선호 기반 검색 시스템 구성

있다. 즉, 사용자 선호 테이블은 도메인별 취향 분석에 사용될 속성들을 선별하여 구성한다. 예를 들어, 선호 주유소 테이블속성에는 "사용자, 소유차종, 선호 주유회사, 선호가격, 세차장 유무" 등으로 구성될 수 있다. 선호 테이블에는 사용자의 질의문 인터페이스를 통해 선호 데이터 값이 입력되어 진다. 입력된 선호 데이터는 코드 변환기를 통해 코드화되어 연관 마이닝 엔진에 의해 연관 규칙이 생성된다.

4.2 사용자 선호 온톨로지 구축

사용자 선호 온톨로지를 작성하여 사용자의 선호 데이터를 입력하여 인스턴스를 생성한 후 시맨틱 웹 규칙 언어와 규칙 엔진을 이용하여 새로운 규칙을 생성하여 연관되고 추론된 지식 검색을 지원하게 된다. 그림 5는 선호 온톨로지 구성으로 크게 User, ServiceType, Event 객체로 분류하고 ServiceType 객체는 도메인을 세부적으로 분류하기 위한 객체로서 정보 제공만을 위한 InformationService와 정보제공 및 판매를 위한 SellingService 로 분류한다. SellingService는 GasOilStation, Restaurant, Bread,

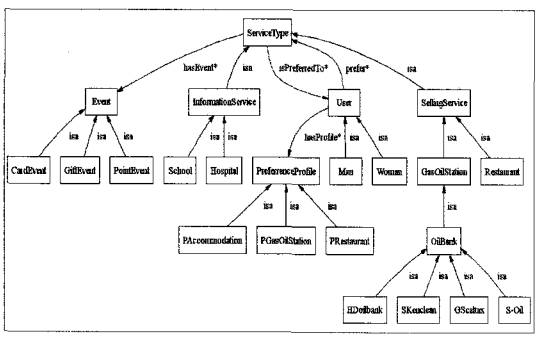


그림 5. 개인 선호 온톨로지 구성

Mart 등과 같이 용도에 따른 다수의 도메인으로 분류한다. 예를 들어 SellingService 객체에서 Restaurant 객체를 한식, 중식, 양식 등의 세부 객체로 분류할 수 있다. PreferenceProfile 객체는 사용자가 도메인에 따라 선호하는 속성을 분류하기 위해 설정한다. User 객체는 Man, Woman 로 분류한다.

객체간의 연관 속성을 나타내기 위해 User와 ServiceType 간에 prefer, isPreferredTo 라는 객체 속성을 정의하고 User와 PreferenceProfile 간에 hasProfile 이라는 객체 속성을 정의한다. prefer 객체 속성은 사용자가 세부적으로 분류화된 선호하고 있는 서비스유형 객체간의 연결을 위한 것으로 주유 선호유형, 음식 선호유형 등을 예로 들 수 있다. hasProfile 객체 속성은 도메인에 따라 구체화한 속성들에 대한 개인의 선호 성향을 연결하기 위한 속성으로 주유소의 경우 선호하는 가격, 세차장 유/무 등에 대한 속성들이 될 수 있다.

4.3 지식 기반 주유소 선호 검색

4.3.1 주유소 선호 테이블 구성

주유소 도메인을 예제로 하여 연관 집합을 추출하기 위해 표 1과 같이 사용자와 주유소간의 선호 속성을 가진 테이블 구조를 생성한다. 사용자의 차종, 연령, 성별 등은 선호하는 주유회사와 선호가격 등과 연관성이 있는 것으로 가정하고 이들을 속성으로 구

표 1. 주유소 선호 속성 데이터 및 코드

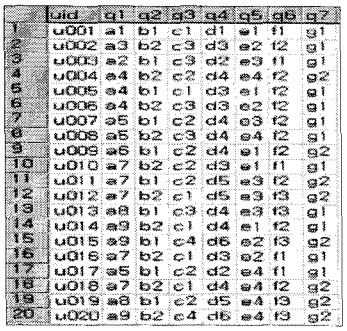
속 성	내용 및 코드
UserID	u001, u002, u003, …, u020
Age	a1(20세미만), a2(20-25), a3(26-30), a4(31-35), a5(36-40), a6(41-45), a7(46-50), a8(51-55), a9(56세이상)
Man/Woman	b1(Man) , b2(Woman)
Company	c1(Hyundai), c2(Samsung), c3(Kia), c4(Ssangyong)
Capacity	d1(1000), d2(1800), d3(2000), d4(2500), d5(3000), d6(3500)
Company	e1(SKenclean), e2(GScaltax), e3(S-oil), e4(HDoilbank)
Price	f1(평균미만),f2(평균가격),f3(평균이상)
Washing	g1(세차장유), g2(세차장무)

성한다. 선호 테이블에 데이터를 입력한 후 연관 마이닝 엔진을 사용하려면 데이터를 코드로 변환하여야 한다. a,b,c, …는 속성별로 부여한 코드를 의미한다.

연관 집합 생성을 위해서 3장에서 제시한 다중항목 연관알고리즘을 기본으로 하여 SQL-Server 데이터 베이스의 프로시저를 이용하여 연관 집합을 생성한다. 다중 항목 연관 알고리즘은 장바구니 분석에서 주로 사용한 단일 항목 필드 구조의 확장된 입력 데이터 구조로서 사용자별로 다수의 항목을 가진 데이터베이스 기반 알고리즘이다[17]. 연관 규칙 집합을 생성할 때, 우선 후보 항목집합을 생성한 후 최소 지지도 이상인 집합에 대해 빈발 함수를 생성한다. 생성한 빈발함수에 대해 최소 신뢰도 이상인 집합이 마지막으로 연관 규칙의 대상이 된다. 각 단계에서 지지도와 신뢰도의 한계를 어떻게 설정하는 가에 따라 연관 규칙의 성립이 결정되는 데, 본 논문에서는 각 단계의 연관 집합 생성을 보고 휴리스틱한 방법으로 선정한다. 그림 6은 표 1의 테이블 구조에 따라 생성한 코드화된 데이터 집합이다.

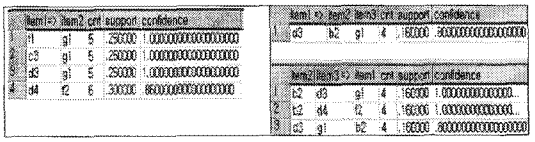
표 2는 단계별로 생성된 연관 집합 및 선정된 규칙의 개수를 나타내고, S_m은 최소 지지도, C_m은 최소 신뢰도를 의미한다. 우선, 모두 30개의 항목에 대해 선호 카운트를 구하는 1단계 후보집합을 생성한다 (C_1). 생성한 후보집합에서 카운트가 5이상, 즉 최소 지지도가 25%인 항목을 빈발항목으로 취한다(F_1). 다음으로, 2개의 항목으로 구성되는 후보집합(C_2)은 모두 109개의 레코드가 구성되어서 카운트를 4이상인 항목을 2개의 빈발항목집합으로 취한다(F_2). 이때, 최소 지지도를 20%로 하여 생성된 빈발 항목 집합은 26개의 레코드이다. 다음 단계로 2개의 빈발 항목 집합 26개의 레코드와 3개의 빈발항목 집합 3개를 대상으로 연관규칙 집합을 생성하면 26개의 항목 집합에서 생성한 52개의 규칙집합이 생성된다. 3개의 항목으로 구성되는 가능한 후보 집합으로 모두 13개의 레코드가 생성되어 카운트를 4이상으로 하여 9개의 빈발항목 집합이 생성된다(F_3). 마지막 단계로 2개의 연관 항목집합에서 최소 지지도와 최소 신뢰도를 각각 0.2와 0.8로 하여 4개의 규칙집합이 선정되고, 3개의 연관 항목집합에서 최소 지지도와 신뢰도를 각각 0.16과 0.8로 하여 3개의 연관 규칙 집합이 추출된다.

그림 7은 마지막으로 선정된 연관 규칙 집합이다. 이렇게 생성된 연관 집합의 코드를 실제 데이터로 변환하여 해석하면 "주유 선호 가격이 평균미만이거나 차종이 Kia, 또는 배기량이 2000cc 일 경우는 세차장이 있기를 원한다. 배기량이 2500 cc 이면 평균가격의 유가를 선호한다" 라는 의미로 해석할 수 있다. "배기량이 2000cc 이상일 경우 여성 고객이자 세차장이 있는 주유소를 선호한다. 여성고객이자 배기량이 2000cc 일 경우 세차장이 있는 주유소를 선호하고, 배기량이 2000cc 이고 세차장이 있는 경우는 여



	uid	g1	g2	g3	g4	g5	g6	g7
1	u001	a1	b1	c1	d1	e1	f1	g1
2	u002	a3	b2	c3	d3	e2	f2	g1
3	u003	a2	b1	c3	d2	e3	f1	g1
4	u004	a4	b2	c2	d4	e4	f2	g2
5	u005	a4	b1	c1	d3	e1	f2	g1
6	u006	a4	b2	c3	d3	e2	f2	g1
7	u007	a5	b1	c2	d4	e3	f2	g1
8	u008	a5	b2	c3	d4	e4	f2	g1
9	u009	a6	b1	c2	d4	e1	f2	g2
10	u010	a7	b2	c2	d3	e1	f1	g1
11	u011	a7	b1	c2	d5	e3	f2	g2
12	u012	a7	b2	c1	d5	e3	f3	g2
13	u013	a8	b1	c3	d4	e3	f3	g1
14	u014	a9	b2	c1	d4	e1	f2	g1
15	u015	a9	b1	c4	d6	e2	f3	g2
16	u016	a7	b2	c1	d3	e2	f1	g1
17	u017	a5	b1	c2	d2	e4	f1	g1
18	u018	a7	b2	c1	d4	e4	f2	g2
19	u019	a8	b1	c2	d5	e4	f3	g2
20	u020	a9	b2	c4	d6	e4	f3	g2

그림 6. 코드화된 선호 데이터



	item1	item2	cnt	support	confidence
1	f1	g1	5	.25000	1.00000000000000000
2	c3	g1	5	.25000	1.00000000000000000
3	a3	g1	5	.25000	1.00000000000000000
4	g4	f2	6	.30000	.85000000000000000

	item1	item2	item3	cnt	support	confidence
1	a3	b2	g1	4	.16000	.80000000000000000

	item2	item3	item1	cnt	support	confidence
1	b2	a3	g1	4	.16000	1.00000000000000000
2	b2	d4	f2	4	.16000	1.00000000000000000
3	c3	g1	b2	4	.16000	.80000000000000000

그림 7. 지지도와 신뢰도에 따른 연관 규칙 집합

표 2. 단계별 항목 집합 생성 결과

개수 \ 단계	1단계 후보	1단계 빈발	2단계 후보	2단계 빈발	3단계 후보	3단계 빈발
연관	30	16 (S_m:0.25)	109	26 (S_m:0.2)	13	9 (S_m:0.2)
규칙	-	-	-	52	-	18
선정된 규칙	-	-	-	4 (S_m:0.2 C_m:0.8)	-	4 (S_m:0.16 C_m:0.8)

성 운전자가 선호한다. 마지막으로 여성고객이자 2500 cc 인 경우는 평균가격을 선호한다"라고 해석할 수 있다.

4.3.2 OWL 제약조건 및 RDF 인스턴스

주유소 도메인을 대상으로 온톨로지 구조를 OWL 코드로 변환하여 객체간의 관계를 정의하고 제약조건을 생성한다. 객체 속성 hasProfile의 Domain은 User 객체이고 Range는 PreferenceProfile이고, prefer 객체 속성은 Domain은 User이고 Range는 ServiceType으로 정의한다. 객체 정의로는 OilBank는 GasOilStation의 서브 클래스이고, Preference Profile은 PGasOilStation의 서브 클래스이다. 또한, 설정한 제약조건으로는 Information Service와 SellingService 의 서브 객체는 서로 disJoint 하고, User 인디비쥬얼은 hasProfile 객체 관계 속성에서 PreferenceProfile 인디비쥬얼로만 대응되어야 하고 prefer 객체 관계 속성에서도 ServiceType 객체의 인디비쥬얼로만 대응되어야 함으로써 universal 제약조건을 부여한다. 그리고 hasProfile 객체는 적어도 1개 이상의 PrefermeceProfile 인디비쥬얼을 가지도록 한다.

그림 8은 주유 선호 데이터를 가지고 OWL 언어를 이용하여 작성한 RDF 인스턴스이다. 이는 사용자 u001에 대한 주유 선호 RDF 파일의 일부를 나타내고 있다. 생성한 RDF 예제 파일은 데이터 속성과 객체간의 연결을 담당하고 있는 객체 속성인 prefer과 hasProfile 속성에 대한 값을 포함하고 있다. u001은 선호하는 주유 유형은 Skenclean 이고 주유 프로필에는 가격은 평균이하, 세차장이 있는 것을 선호하는 것으로 되어 있다.

```
<prefer>
  <SKenclean rdf:ID="SKencleanService">
    <isPreferredTo>
      <User rdf:ID="un001">
        <hasProfile>
          <PGasOilStation rdf:ID="pgu001">
            <carwashing rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean">
            true</carwashing>
            <preferenceprice rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
            average below</preferenceprice>
          </PGasOilStation>
        </hasProfile>
        <u-mfemale rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
          Man</u-mfemale>
        <u-carcompany rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
                Hyundai</u-carcompany>
        <u-identification rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
          u001</u-identification>
        <u-carcapacity rdf:datatype="http://www.w3.org/2001/XMLSchema#inf">
          1000</u-carcapacity>
        <u-age rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
          19</u-age>
        <prefer rdf:resource="#SKencleanService"/>
      </User>
    </isPreferredTo>
          <serviceName rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
            SKenclean</serviceName>
    <serviceType rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
      GasOilStation</serviceType>
  </SKenclean>
</prefer>
```

그림 8. RDF 인스턴스

4.3.3 새로운 선호 규칙 생성

데이터마이닝에서 추출한 규칙을 SWRL 언어를 사용하여 입력한 후 Jess engine 을 이용하여 새로운 선호 규칙을 생성한다. 이를 위해 Protégé/OWL에서 SWRL Tab을 이용하여 Jess 엔진을 통하여 OWL 지식과 SWRL 규칙을 결합하여 새로운 추론된 지식을 표시한다. 그림 9는 연관 마이닝에서 도출된 연관 규칙을 SWRL 을 사용하여 작성한 규칙으로 추론을 수행한 결과 20개의 속성들이 추론되었다. 여기서 PGU10, PGU13번 인디비듀얼의 carwahing 속성 값이 새로운 값으로 추론되었다. 즉, 추론된 속성 중에서 객체 속성 hasProfile 에 의해 연관되는 PGasOilStation 클래스의 인디비쥬얼 중에서 PGU013의 원래 carwashing 값이 true값인데 추론되어 false값이 새롭게 생성되었다. 이는 u013이 본인의 profile을 이용하여 추론 검색을 할 경우, 비록 세차장이 있는 주유소를 검색하기를 원하지만 규칙에 의해 세차장이 없는 주유소 또한 이 사용자에게는 유용한 정보가 될 수 있음을 암시한다.

4.3.4 온톨로지 기반 검색

온톨로지 기반에서 검색 시스템은 계층화된 검색 및 연관된 검색을 지원할 뿐 아니라 추론 검색을 지원함으로써 데이터베이스로부터의 검색보다 장점이

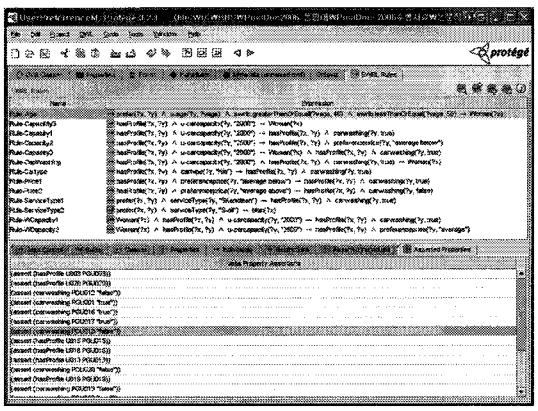


그림 9. SWRL 규칙기반 추론 결과

있다. 일반적인 데이터베이스로 부터의 검색에서는 복잡한 검색 질의를 할 경우 테이블간의 조인으로 인한 검색 속도가 늦어질 수 있다. 또한, 사용자가 입력한 키워드에 매칭되지 않을 경우에는 데이터 검색 결과를 제공할 수가 없다. 이러한 관점에서 온톨로지 기반 검색의 장점은 다음과 같다.

첫째, 객체간의 연관 검색을 수행하여 검색 속도를 높일 수 있다. 다수의 객체간의 연결에 의한 복잡한 질의문에 대한 검색은 데이터베이스의 조인으로 인해 검색 속도가 늦어 질 수 있다. 온톨로지 기반으로 객체 간의 연관되는 인디비쥬얼을 검색한 후 실제 데이터가 있는 한 개의 테이블을 검색함으로써 다수의 조인 없이 데이터를 검색할 수 있어 검색 속도를 높일 수 있다. 또한 데이터 베이스에서 정의한 연관성 뿐 만 아니라 온톨로지에 새로운 연관성을 생성함으로써 데이터베이스의 구조를 변경함으로써 발생하는 비용을 절감할 수 있다. 따라서 온톨로지 기반의 웹 검색에서는 다수의 검색어로 여러 번의 검색 과정을 거치는 복잡한 과정이 줄어들기 때문에 데이터베이스만으로 부터의 검색 속도 보다 빠른 검색 결과를 제공한다.

둘째, 온톨로지의 계층적인 구조를 이용하여 상위계층 기반의 검색을 제공함으로써 검색 결과가 없는 경우에 상위계층에서 의미적인 연관성을 가진 검색 결과를 제공할 수 있다. 예를 들어, 선호하는 주유타입이 "SKenclean" 이고 가격이 "1600원" 이상인 주유소를 검색했을 경우 검색 결과가 없다면 주유타입에 대한 동일한 계층에 대해 가격이 "1600원"이상인 주유소 검색 결과를 표시한다면 사용자는 검색 회수를 줄일 수 있다.

셋째, 추론 검색으로, OWL 제약조건, 생성한 규칙, 입력한 데이터를 이용하여 새로운 규칙 즉, 지식을 구축함으로써 사용자에게 추론된 검색 결과를 제공할 수 있다. 이러한 추론 검색은 사용자가 입력한 검색어에 대해 검색 결과가 없을 경우 추론된 검색 결과를 제공함으로써 사용자에게 유용한 정보를 제시할 수 있다.

5. 결 론

인터넷 사용이 교육, 쇼핑, 문화, 홍보 등 모든 분야에서 보편적으로 사용되고 있음에 따라 방대한 데이터가 축척되어 있어 사용자들은 이러한 데이터에서 원하는 정보를 찾고자 많은 시간을 투자하고 있다. 이러한 추세에서 현재, 웹에 있는 데이터에서 숨은 지식 찾아내려는 연구가 웹 마이닝, 지식 추론 알고리즘, 온톨로지 툴 개발 등의 분야에서 진행되고 있다.

본 논문에서는 이러한 지식 추출을 위하여 연관 규칙 알고리즘을 이용하여 생성된 새로운 규칙을 온톨로지와 연계함으로써 개념화된 새로운 지식을 사용자에게 제공함으로써 효과적인 검색을 지원받을 수 있는 추론 알고리즘을 구축하였다. 지식을 개념화하기 위한 데이터 구조로 OWL 온톨로지를 사용하고, 이러한 구조에서 SWRL 규칙과 융합하여 새로운 규칙을 생성하여 개념화된 지식을 도출한다. 기존 온톨로지 기반의 검색이 개념화 된 구조화에서 연관된 지식을 제공할 수 있는 장점과 신뢰성 있는 연관 규칙과의 결합으로 개념화된 새로운 지식을 제공할 수 있으므로 사용자에게 더 효과적인 검색 정보를 제공할 수 있다. 본 논문에서는 주유소 관련 예제 데이터를 사용하여 사용자 선호 온톨로지를 구축한 후 지식 추론에 의해 사용자에게 효과적인 정보를 추가적으로 보여줄 수 있음을 제시하였다.

제안 알고리즘은 온톨로지 기반의 검색 장점인 연관 및 추론 검색이 실생활의 검색과 밀접한 개인 선호 데이터를 이용함으로써 방대한 데이터에서 사용자에게 연관되고 추론된 정보를 제공할 수 있는 시스템을 제공하였다는 점에서 의미를 가진다. 향후, 이러한 검색 알고리즘을 바탕으로 실제 사용자의 선호 데이터를 수집하여 주기적으로 새로운 지식을 제공할 수 있는 시스템의 구현과 온톨로지와 연관 규칙 기반검색의 효율성에 대한 지속적인 연구가 필요하다.

참 고 문 헌

[1] R.Guha, R.McCool, and E.Miller, "Semantic Search," *Proceedings of the 12th International Conference on World Wide Web*, pp. 779-830, 2003.

[2] V.Sugumaran and S.Vijayan "A Semantic-Based Approach to Component Retrieval," *The Database for Advances in Information Systems*, Vol.34, No.3, pp. 8-24, 2003.

[3] T.Berners-Lee, J.Hendler, and O.lassila, "The

Semantic Web," *Scientific American*, Vol.284, No.5, pp. 34-43, 2001.

[4] G.Fischer and J.Ostwald, "Knowledge Management : Problems, Promises, Realities, and Challenges," *IEEE Intelligent Systems*, Vol.16, No.1, pp. 60-72, 2001.

[5] M.F.Uschold and R.J.Jasper, "A Framework for Understanding and Classifying Ontology Applications," *Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods(KRR5)*, pp. 1-12, 1999.

[6] Xizo Hang Wang, Tau Gu, Da Qing Zhang, and Hung Keng Pung, "Ontology Based Context Modeling and Reasoning using OWL," *IEEE International Conference on Pervasive Computing and Communication*, pp. 18-22, 2004.

[7] H.Knublauch, "Ontology-Driven Software Development in the Context of the Semantic Web: An Example Scenario with Protégé/OWL," *International Workshop on the Model-Driven Semantic Web*, 2004.

[8] B.Berendt, A.Hotho, and G.Stumme, "Semantic Web Mining and the Representation, Analysis, and Evolution of Web Space," *Proceeding of the 1st Workshop on Representation and Analysis of Web Space*, 2005.

[9] R.Rosati, "On the Decidability and Complexity of Integration Ontologies and Rules," *Journal of Web Semantics*, Vol.3, No.1, pp. 61-73, 2005.

[10] G.Christine, "Web Rules for Health Care and Life Sciences: Use Cases and Requirements," *International World Wide Web Conference*, 2006.

[11] R.Agrawal, T.Imielinski, and A.Swami, "Mining Association Rules between Sets of Items in Large Databases," *Proceeding of ACM SIGMOD Conference on Management of Data*, pp. 207-216, 1993.

[12] R.Agrawal and R.Srikant, "Fast Algorithms for Mining Association Rules," *Proceeding of the 20th VLDB Conference*, pp. 487-499, 1994.

[13] A.Pieter and D.Zantinge, *Data Mining*, Addison-Wesley, 1996.

[14] M.Holsheimer, "A Perspective on Databases and Data Mining," *Proceeding of 1st International Conference on Knowledge Discovery and Data Mining*, pp. 150-155, 1995.

[15] G.Christine, "Combining SWRL rules and OWL ontologies with Protégé OWL Plugin, Jess, and Racer," *Proceeding of 7th International Protégé Conference*, Bethesda, MD, 2004.

[16] O.Martin, K.Holger, T.Samson, and M.Mark, "Writhing Rules for the Semantic Web Using SWRL and JESS," *8th International Protege Conference*, 2005.

[17] 황현숙, 박규석 "연관 규칙 기반의 상품 검색 데이터베이스 최적화 연구," 멀티미디어학회논문지, 제7권, 제2호, pp. 145-155, 2004.

황 현 숙

2001년 부경대학교 경영정보학 박사
2003년 8월~2004 7월 미국 UMKC Post Doc. 연수과정 수행
2006년 9월~2007년 8월 학술진흥재단 국내 Post Doc. 연수과정 수행
2008년 3월~현재 부경대학교 BK21 사업단 연구원
관심분야 : u-방재시스템, LBS/GIS 시스템, 온톨로지, 데이터마이닝, 유비쿼터스 센서 네트워크

이 준 연

1992년~1995년 Microsoft Inc.
2000년 중앙대학교 컴퓨터공학과 박사
2000년 3월~현재 동명대학교 멀티미디어공학과 부교수
관심분야 : 센서 네트워크, 위치기반 서비스, 헬스케어 시스템