韓國 컴퓨터情報學會 論文誌
第13卷 第7號, 2008. 12.

2008-13-7-5-3

귀금속·보석 상품정보 온톨로지 구축에 관한 연구

이 기 영*

A Study on the Development of Ontology based on the Jewelry Brand Information

Ki-Young Lee*

요 약

본 연구에서는 웹 문서에서의 단순 키워드 매칭으로 검색하는 전자상거래시스템의 문제점을 해결하기 위한 방안으로 도메인 온톨로지를 자동으로 생성하고 이를 기반으로 지능형 에이전트기술을 접목함으로서 의사소통이 단일화된 상품검색시스템을 개발한다. 온톨로지 개발은 국제상품분류코드(UNSPSC)와 귀금속·보석 사이트들의 분류정보를 기반으로 대표용어를 추출하고 유사관계 시소러스 적용하여 표준화된 온톨로지를 구축하며 지능형에이전트 기술을 검색 단계에서 접목시켜 사용자에게 정보수집의 효율성을 지원하도록 시맨틱 웹을 지원하는 상거래 시스템을 설계하고 구현한다. 또한 개인화된 검색 환경을 지원하기 위해 사용자 프로파일을 설계하고, 개인화 검색 에이전트와 추론기능을 이용한 검색 환경을 제공함으로서 정보수집의 신속성과 정확한 정보검색이 가능하도록 지원한다.

Abstract

This research is to develop product retrieval system through simplified communication by applying intelligent agent technology based on automatically created domain ontology to present solution on problems with e-commerce system which searches in the web documents with a simple keyword. Ontology development extracts representative term based on classification information of international product classification code(UNSPSC) and jewelry websites that is applied to analogy relationship thesaurus to establish standardized ontology. The intelligent agent technology is applied to retrieval stage to support efficiency of information collection for users by designing and developing e-commerce system supported with semantic web. Moreover, it designs user profile to personalized search environment and provide personalized retrieval agent and retrieval environment with inference function to make available with fast information collection and accurate information search.

▸ Keyword : 온톨로지(Ontology), 시맨틱 웹(Semantic web), 전자상거래(e-Commerce), 프로파일(Profile)

• 제1저자 : 이기영
• 접수일 : 2008. 9. 25, 심사일 : 2008. 11. 5, 심사완료일 : 2008. 12. 24.
* 원광보건대학 영상컨텐츠과 교수
※ 이 논문은 2006년도 원광보건대학 교내연구비 지원에 의해서 수행되었음.

Ⅰ. 서 론

기존의 전자상거래시스템의 물품검색과정은 동의어, 다의어 처리문제와 선별되지 않은 방대한 상거래관련 정보들에서 사용자가 반복적인 피드백 필터링과정등과 같은 많은 시간을 투자하여 정보를 습득하고 검토하여야 하는 어려움이 있다. 또한 검색시스템의 낮은 기능성으로 인하여 수요자와 공급자 모두에게 많은 부하를 요구함으로 결과적으로 전자상거래의 보급과 한계를 가져오게 한다[1]. 시멘틱 웹은 이와 관련된 문제점을 해결할 차세대 웹 기술로 자원간의 관계를 연결시키는 온톨로지 기술이다. 시멘틱 웹과 자동화 에이전트 기술을 접목시켜 의미론적 검색, 상호운용성 보장, 어플리케이션간의 통합 등의 효과를 가져올 수 있다[2]. 따라서 본 논문에서는 국제상품분류코드(UNSPSC)와 귀금속・보석 사이트들의 분류정보를 기반으로 유사관계시소러스 적용하여 표준화된 온톨로지를 구축하고 정보검색 에이전트 기술을 검색 단계에서 접목시켜 사용자에게 정보수집의 효율성을 지원하도록 시맨틱 웹을 지원하는 상거래 시스템을 설계하고 구현한다.

따라서 본 연구에서는 체계적인 정보 표현을 위한 온톨로지 자동 구축 단계를 기술하고 사용자에게 방대한 검색 결과를 제공하는 문제점을 해결하기 위해 물품검색단계를 개인화하는 프로파일을 구성하고 메타 데이터화하여 검색결과의 순위를 차별화함으로서 사용자 검색환경을 개선할 수 있다.

온톨로지 자동 구축 단계는 귀금속·보석 관련분야의 전자상거래 상품정보를 기반으로 용어 관계행렬을 구성하여 대표용어를 추출하고, 용어들 간의 내용분석을 통해 연관정보를 추출하는 과정이며, 개인화 정보검색 단계는 개개인의 차별화된 검색환경을 지원하기 위한 사용자 프로파일를 기반으로 사용자 질의를 분석하여 순위화된 검색결과를 제공하는 과정이다. 이는 상거래 구매주체인 소비자의 편리성을 도모하며 지식과 소프트웨어의 재사용(reuse) 및 프로젝트의 개발기간을 단축하고 소요되는 자원의 절약을 기할 수 있다.

본 논문의 구성은 먼저 2장에서 본 연구의 기초가 되는 상품분류 온톨로지와 시맨틱 웹 에이전트 시스템에 대하여 설명하고, 3장에서는 귀금속・보석 상품정보를 도메인으로 하는 온톨로지 구축방법을 설명한다. 4장에서는 자동 구축된 온톨로지를 기반으로 개인화 상거래시스템을 설계하고 구현한다. 마지막으로 결론 및 향후 연구과제에 대하여 기술한다.

Ⅱ. 관련연구

2.1 상품분류 온톨로지

상품식별코드는 상품을 명확하게 식별하기 위한 코드이지만 상품간의 관련성을 표현하지 못하므로, 상품정보를 체계화하여 분류하는 방법인 국제상품분류코드(UNSPSC), SIC/NAIC, UCC/EAN, HS, eClass 등의 다양한 상품분류코드가 등장하게 되었다. 이는 상품이라는 개념을 나름대로의 관점을 통해서 계층관계로 정의한 단순한 형태의 온톨로지라 할 수 있다. 이중에서 UNSPSC는 그 필요성과 중요성을 인지한 다양한 분야의 기업에서 채택되어 활용되고 있으며 세그먼트(Segment), 패밀리(Family), 클래스(Class), 커머디티(Commodity)의 4개의 계층화된 레벨로 구분되어 8자리의 숫자 값으로 표현된다. 여기에서 세그먼트는 분석을 목적으로 하는 논리적인 집합이며, 패밀리는 일반적으로 인정되는 상품관련성 있는 상품분류이며, 클래스는 사용 혹은 기능을 공유하는 상품그룹을 표현하며 커머디티는 대체 가능한 상품 혹은 서비스그룹을 표현한다. 따라서 UNSPSC 분류 체계를 활용함으로서 우선 구매자와 판매자가 표준화된 상품과 서비스의 계층적인 분류체계를 사용함으로서 전자상거래가 활성화될 수 있고 사용자는 보다 정확한 상품의 검색과 구매가 가능하게 된다[3].

온톨로지를 구축함에 있어 풍부한 상품정보를 기반으로 정보지식의 공유, 재사용, 시스템 연계 측면에서 G2B, 민간 B2B 분야의 콘텐츠의 필요성이 부각되고 이를 활용한 서비스가 획기적으로 증가하고 있다. 따라서 시멘틱 웹을 기반으로 사용자 개인화된 정보 지식을 공유함으로서 차별화된 상품검색을 지원함으로서 전자상거래 시스템의 지능화, e-Marketplace 및 e-비즈니스 시스템의 생산성과 거래비용이 획기적으로 향상될 것으로 기대된다.

2.2 시맨틱 웹 에이전트 시스템

시멘틱 웹 기반의 서비스를 제공하기 위해서는 특정 도메인 지식에 대한 명시적인 명세화 및 지식의 개념과 개념과의 관계를 정형화하는 온톨로지를 통해 이루어진다. 도메인 온톨로지는 정보를 체계화하는 패러다임이다[4]. 기존의 전자상거래 시스템에서 지능형 에이전트의 기술은 주로 상품검색, 판매자 검색, 협상단계에 많이 적용되고 있다. 상품검색단계에서는 지능형에이전트 기술의 개인화 기술을 통해 소비자의

선호도나 관심의 이동을 획득하여 소비자 의도에 맞는 상품을 추천하는 기능이며 판매자 검색단계에서는 비교 쇼핑에이전트 기술을 이용하여 소비자의 소비능력과 비슷한 상품들을 필터링해주는 기능을 이용할 수 있다[5].

따라서, 시맨틱 웹 기술을 상거래시스템에 도입함으로서 웹사이트에 들어오는 고객의 성향과 행태별로 세분화하여 콘텐츠를 보여주거나 서비스를 제공하기 위해서 사용자의 개인적인 취향 및 구매 기록에 따라 페이지를 구성하고 제품을 추천 받을 수 있는 기능들을 제공할 수 있으며 유사한 집단의 패턴정보를 근거로 하여 서비스를 제공할 수 있다.

온톨로지 기술과 에이전트 시스템의 연계를 통한 전자상거래 응용분야의 접목을 위해서는 지능형 에이전트 시스템의 프로세서에서 온톨로지 기술이 접목되어야 하며 시맨틱웹에서 DAML를 지원하는 에이전트가 웹서비스의 기술내용을 찾아서 시나리오를 표시할 수 있다[6].

III. 귀금속·보석 상품정보 온톨로지 구축

도메인 온톨로지는 특정 도메인의 전문적인 용어들의 관계를 나타내는 전문가가 사용가능한 매커니즘이어야 한다. 또한 사용자가 원하는 정보를 같은 도메인의 다른 데이터 집합들과의 일관성 있는 내용과 정보를 찾아 추출할 수 있도록 자동화된 추론이 가능한 메카니즘이어야 한다. 그러나 기존의 수동적인 온톨로지 구축방식은 비용과 시간이 많이 소모되는 단점이 있으므로, 해당 도메인에서 개념(Concept)으로 표현되는 대표용어를 추출하고 유사관계 알고리즘을 통해 자동으로 온톨로지를 구축한다.

3.1 도메인 온톨로지 구축

온톨로지는 특정도메인 개념 및 지식을 명세화하기 위해서 그 지식을 설명하는 표준 용어들을 정의하고 용어들 사이의 계층(taxonmy) 및 연관관계를 정의하는 것이다. 온톨로지는 정의된 키워드를 의미하고 관계는 온톨로지 내에 정의된 개념들의 관계를 의미한다[4,7].

본 논문에서의 온톨로지 개발은 도메인의 특성과 온톨로지 사용 용도를 고려하여 온톨로지 구성영역을 정하고 표준용어들과 용어들 사이의 개념관계를 정의한다[8,9]. 또한 상품분류 온톨로지 및 귀금속·보석 상거래 사이트의 계층구조를 조사하고 도메인 전문가들과의 협의를 통하여 관계들의 구조를 정한다. 온톨로지의 개발은 기존에 사용되었던 소프트웨어 공학기술과 지식기반 시스템개발에 적용되었던 방법론에서 많은 부분을 응용할 수 있다. 기본 온톨로지 구성은 〈그림 1〉와 같이 구성되며 도메인 온톨로지 구축시 문제점인 의미가 애매한 단어의 중의성을 해결하기 위한 알고리즘을 이용하여 메타데이터를 생성하는 방법을 제안한다.

귀금속·보석 사이트의 상품정보데이터는 문서변환 과정을 거쳐 구조화되며 간단한 자연어 처리 과정을 거친 뒤 개념용어를 추출한다. 그리고 용어들 간의 유사관계를 이용하여 대표 개념을 추출하고 이들의 구조를 분석한 결과로서 계층구조를 도출해낸다. 마지막으로 추출된 관계들을 이미 존재하고 있는 온톨로지 개념들과 함께 추가시킨다.

이는 HTML 문서에서 서로 관련있는 용어와 속성 값을 추출해내기 위해, 귀금속·보석 정보를 취급하는 사이트로부터 문서집합을 추출한다. 각각은 보석에 관련된 사이트의 품목정보를 XML 문서로 변환한 문서로 이루어져 있다. 문서에서 개념용어를 추출하고 개념간의 연관관계를 기반으로 내용에 기반한 관계를 도출한다.

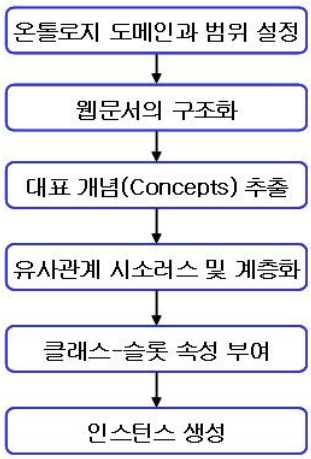

그림 1. 온톨로지 구성
Fig 1. The flowchart of ontology structure

3.1.1 용어사전 구축

귀금속·보석관련 문서로부터 용어를 추출하는 방법은 용어의 출현빈도를 기준으로 시그모이드 함수를 이용하였다. 문서 집합에서 용어를 추출하고 용어와 문서사이의 관련 정도를 가중치로 표현한다. 용어 출현 빈도가 문서 내용을 대표한다는 가설에 근원으로 특정 문서에 대한 용어의 의미 관계를 퍼지화 하였다. 이를 위해 퍼지 소속 함수로 대표적인 비선형

함수인 S자 형태의 시그모이드 함수를 이용하며, 용어의 중요 정도를 측정하고 용어간의 관계 정도를 측정하여 대표 색인을 클러스터링 한다[8,10,11].

〈그림 2〉은 대표 시그모이드 소속 함수를 나타내며 임계 값은 도메인의 특성에 따라 유동적이다.

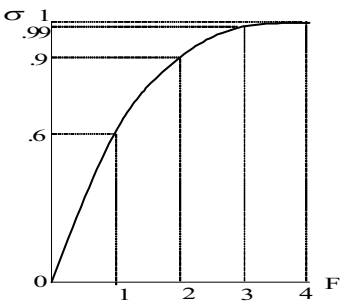

그림 2. 소속함수 (σ)

Fig 2. membership function(σ)

〈그림 2〉은 용어의 빈도수 2이상이면 매우 소속정도가 크며, 1인 경우도 소속정도가 0.6으로 대단히 크다는 의미로 표현된다.

〈그림 2〉의 시그모이드 함수를 통하여 문서집합(D)와 용어 집합(T)의 퍼지 용어 관계를 다음과 같이 표현한다.

$$\mathbf{R} = \begin{array}{c} \\ \mathbf{d}_1 \\ \mathbf{d}_2 \\ \vdots \\ \mathbf{d}_n \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{t}_1 & \mathbf{t}_2 & \cdots & \mathbf{t}_m \end{array} \\ \begin{bmatrix} \mathbf{w}_{11} & \mathbf{w}_{12} & \cdots & \mathbf{w}_{1m} \\ \mathbf{w}_{21} & \mathbf{w}_{22} & \cdots & \mathbf{w}_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{w}_{n1} & \mathbf{w}_{n2} & \cdots & \mathbf{w}_{nm} \end{bmatrix} \end{array}$$

여기서, 문서 집합의 개수는 n이고 문서 집합에서 추출된 용어는 m개이다. 소속함수를 이용하여 문서에서 용어의 소속 정도를 계산하고 용어 가중치에 의한 용어사전을 구축한다.

본 논문을 기술함에 있어 수식과 정의는 실선으로 표현하며 예제에 의한 계산 절차과정은 점선으로 표현한다.

〈예1〉 용어 발생 빈도별 소속 값을 생성하고 문서 집합 (D)에서 용어(T)의 의미를 표현하는 원시문서베이스(R)은 다음과 같다.

$$D=\{d_1,d_2,d_3,d_4,d_5\},\ T=\{t_1,t_2,t_3,\cdots,t_9\}$$

$$=\left\{\begin{array}{l}\text{목걸이, 여성용, 장신구, 주얼리,}\\ \text{토드말린, 구리, 모조보석, 순보석, 음이온}\end{array}\right\}$$

$$R=\begin{array}{c} \\ d1 \\ d2 \\ d3 \\ d4 \\ d5 \end{array}\begin{array}{c}\begin{array}{ccccccccc} t1 & t2 & t3 & t4 & t5 & t6 & t7 & t8 & t9 \end{array}\\ \begin{bmatrix} 0.98 & 0.85 & 0.50 & 0.80 & 0.00 & 0.50 & 0.50 & 0.80 & 0.00 \\ 1.00 & 0.70 & 0.50 & 0.00 & 0.90 & 0.75 & 0.50 & 0.00 & 0.94 \\ 0.94 & 0.00 & 0.80 & 1.00 & 0.00 & 0.00 & 0.80 & 1.00 & 0.00 \\ 0.89 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.74 & 0.00 & 0.00 & 0.50 & 0.00 & 0.00 & 0.00 & 0.70 & 0.00 \end{bmatrix}\end{array}$$

3.2 대표용어 행렬 구성

본 논문에서는 도메인 의존적인이며 웹 문서를 대표하는 개념행렬을 생성하며 이를 기반으로 유사관계 시소러스를 구성하는 방법을 제안한다.

대표용어 집합은 원시 문서베이스를 이용해서 생성하며 용어 집합의 부분집합이다. 본 논문에서는 시소러스를 구축하기 위해 각 용어의 관계 값을 도출하는 방법을 적용하였으며 문서집합 특성을 이용하여 각 용어의 관계 값을 생성한다. 이는 두개의 퍼지 집합 사이의 동치관계를 논리적 동치인 불리언 대수를 퍼지 집합에 적용하기 위해 다음과 같이 퍼지 소속 함수로 표현한다[8,11,12].

$$\left.\begin{array}{l} u_{A\cong B}(w)=\max\{\min(u_A(w),u_B(w)),\\ \qquad \min(1-u_A(w),1-u_B(w))\} \\ t_i=u_{w_i\equiv w_j}=\dfrac{1}{|d|}\displaystyle\sum_{k=1}^{d}u_{w_i\equiv w_j}(D_k) \end{array}\right] \text{--(식3-1)}$$

단, $u_A(w)$: 임의의 원소w가 퍼지 집합A에 속할 정도
$u_B(w)$: 임의의 원소w가 퍼지 집합B에 속할 정도
t_i : 용어i가 문서집합(도메인)에서의 관계 정도
$|d|$: 전체 문서의 개수
$u_{w_i\equiv w_j}(D_k)$: 문서k에서 용어i,j 사이의 유사정도

대표용어집합은 각 용어의 퍼지 값에 대하여 $\alpha-cut$을 적용함으로써 도메인 영역에서 문서를 분류하기에 부적합한 용어를 제거할 수 있는 장점을 갖고 있다. 본 논문에서는 (식 3-1)의 소속 함수를 이용하여 각 용어가 도메인 전체 영역에서의 관계 정도를 평가하는 방법을 이용한다.

즉, $u_A(w)=u_B(w)$일 경우를 평가하는 방법[11,12]을 응용하여 도메인에서 색인어를 평가하고 $\alpha-cut$에 의한 대표용어 집합을 생성하였다. 문서와 대표용어의 퍼지 관계를 표현하는 대표용어 집합 기반의 문서베이스는 원시 문서 베이스에

서 문서 내용을 대표할 수 있는 대표용어만을 추출하여 구성한다. 여기서, 대표용어 집합(R_r)은 다음과 같다.

$$R_r = \begin{array}{c} \\ d_1 \\ d_2 \\ \vdots \\ d_n \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & \cdots & t_r \end{array} \\ \begin{pmatrix} I_{11} & I_{12} & \cdots & I_{1r} \\ I_{21} & I_{22} & \cdots & I_{2r} \\ \vdots & \vdots & \cdots & \vdots \\ I_{n1} & I_{n2} & \cdots & I_{nr} \end{pmatrix} \end{array}$$

단, r은 축소행렬의 첨자로 $r < m$이다.

〈예2〉 〈예1〉에서 구한 원시 문서베이스(R)에서 (식 3-1)를 이용하고 $\alpha-cut$을 0.8로 적용했을 경우 대표용어행렬은 다음과 같이 얻을 수 있다.

원시베이스(R)에서 색인어 t_1의 계산 절차는 다음과 같다.

$$\begin{aligned} \mu_{t_1}(w) = & max\ \{min(0.98,\ 0.98), min(1-0.98, 1-0.98))\} + \\ & max\ \{min(1.00,\ 1.00), min(1-1.00, 1-1.00))\} + \\ & max\ \{in(0.94,\ 0.94), min(1-0.94, 1-0.94))\} + \\ & max\ \{min(0.89,\ 0.89), min(1-0.89, 1-0.89))\} + \\ & max\ \{min(0.74,\ 0.74), min(1-0.74, 1-0.74))\} \\ = & 4.55 \end{aligned}$$

$t_1 = 4.55/5 = 0.91$ 이므로, 계산결과가 0.8이상인 용어가 대표용어로 구성된다. 따라서, 다음과 같이 $(t_1, t_2, t_5, t_8, t_9)$로 구성된 대표행렬을 얻을 수 있다.

$$R_r = \begin{array}{c} \\ d1 \\ d2 \\ d3 \\ d4 \\ d5 \end{array} \begin{array}{c} \begin{array}{ccccc} r1 & r2 & r3 & r4 & r5 \end{array} \\ \begin{pmatrix} 0.98 & 0.85 & 0.00 & 0.80 & 0.00 \\ 1.00 & 0.70 & 0.90 & 0.00 & 0.94 \\ 0.94 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.89 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.74 & 0.00 & 0.00 & 0.00 & 0.00 \end{pmatrix} \end{array}$$

3.3 유사관계시소러스

퍼지 관계성을 표현하는 퍼지 호환관계(tolerance, compatibility relation)는 다음과 같이 반사와 대칭 성질을 만족하고 전이관계는 만족하지 않는다.

1) 반사관계 : $\mu_{\leqq}(x,x) = 1$
2) 대칭관계 : $\mu_{\leqq}(x,y) = \mu_{\leqq}(y,x)$
3) 전이관계 : $\mu_{\leqq}(x,z) \geqq \min\{\mu_{\leqq}(x,y), \mu_{\leqq}(y,z)\}$

μ : *membership function*

일반적인 집합(crisp set)에서 호환관계 $R(X,X)$이 주어졌을 때 하나의 호환 클래스 집합A는 A에 속하는 임의의 x, y에 대하여 x, y의 관계가 R에 속하면 X의 부분집합이다[11,12,13,14].

또한 임의의 호환 클래스(compatibility class) 내에서 완전히 포함되지 않는 호환 클래스를 최대 호환 클래스라 하고, X에 대하여 관계R에 의해 생성된 모든 최대 호환 클래스를 R에 대하여 X의 완전 cover라고 한다. 반면, 퍼지 집합에서도 퍼지 호환 관계 R이 임의의 집합 A에 주어지면, 호환관계를 만족하는 부분 집합들로 분할될 수 있는데 이와 같이 얻어진 부분집합들은 퍼지 호환 클래스(fuzzy compatibility class)라 한다. 퍼지 호환 관계에 $\alpha-cut$을 적용하여 생성된 α-호환 클래스 A_i는 다음과 같이 정의된다.

$$\mu_R(x,y) \geq \alpha,\ \forall x, y \in A_i$$

즉, 임의의x, y 에 대하여x, y의 관계가 특정 α값 이상이면X의 부분집합으로 구성되고 이렇게 구성된 모든 호환 클래스들을 최대 호환 클래스 또는 완전 α-cover라고 한다.

퍼지 호환관계(tolerance relation)의 특성을 만족하는 유사관계 행렬을 정의함으로서 용어를 분류할 수 있으며 호환관계를 만족하고 $\alpha-cut$을 적용하여 유사관계시소러스를 구성할 수 있다.

대표용어로 구성된 문서베이스를 기반으로 용어 사이의 퍼지 관련 정도를 나타내는 시소러스를 구성한다. 시소러스는 다음 (식 3-2)과 같이 문서베이스와 원시 문서베이스의 퍼지 관계곱 연산을 이용한다.

$$\left.\begin{array}{c} S_r = R^T \otimes R_r \\ s_{ij} = \underset{i,j=1.n}{\vee} (\min(w_{in}, I_{nj}), (\min(1-w_{in}), (1-I_{nj}))) \end{array}\right] \text{--(식 3-2)}$$

s_{ij} : 색인어i와 대표용어j의 유사 정도
w_{in} : 문서n에서 용어i의 중요 정도
I_{nj} : 대표용어행렬 문서n에서 용어 j의 중요정도

퍼지 관계곱 연산은 특정 문서에서 용어 간 동시 출현 빈도가 많을수록 유사성이 높다는 가정 하에 (식 3-2)의 동시 출현 빈도를 고려하였다.

여기서, 유사관계 시소러스S_r은 다음과 같다.

$$S_r = \begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_m \end{array} \begin{array}{c} \begin{array}{cccc} r_1 & r_2 & \cdots & r_r \end{array} \\ \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1r} \\ s_{21} & s_{22} & \cdots & s_{2r} \\ \vdots & \vdots & \cdots & \vdots \\ s_{m1} & s_{m2} & \cdots & s_{mr} \end{pmatrix} \end{array}$$

대표용어 집합 기반의 시소러스는 용어의 의미를 정의하기 위한 관점에서 구축하였으므로 개념 행렬(concept matrices)이라 할 수 있다.

〈예3〉 〈예2〉의 대표용어 행렬과 원시 문서베이스의 퍼지 관계를 (식 3-2)을 이용하면 유사관계 시소러스를 다음과 같이 구할 수 있다.

유사관계 시소러스(S_r)의 원소s_{23}의 계산 절차는

$$\begin{aligned} s_{23} = & \; max(min(0.85, 0.00),\ min(1-0.85,\ 1-0.00)) + \\ & max(min(0.70,\ 0.90), min(1-0.70,\ 1-0.90)) + \\ & max(min(0.00,\ 0.00), min(1-0.00,\ 1-0.00)) + \\ & max(min(0.00,\ 0.00), min(1-0.00,\ 1-0.00)) + \\ & max(min(0.00,\ 0.00), min(1-0.00,\ 1-0.00))) \\ = & \; 3.85 \end{aligned}$$

이므로$s_{23} = 3.85/5 = 0.77$되며, 다른 원소도 같은 방법으로 계산하면 다음과 같은 행렬을 구할 수 있다.

$$S_r = \begin{array}{c} \\ t1 \\ t2 \\ t3 \\ t4 \\ t5 \\ t6 \\ t7 \\ t8 \\ t9 \end{array} \begin{array}{c} \begin{array}{ccccc} r1 & r2 & r3 & r4 & r5 \end{array} \\ \begin{pmatrix} 0.91 & 0.40 & 0.27 & 0.51 & 0.28 \\ 0.40 & 0.91 & 0.77 & 0.48 & 0.77 \\ 0.43 & 0.64 & 0.64 & 0.62 & 0.64 \\ 0.47 & 0.52 & 0.36 & 0.86 & 0.35 \\ 0.27 & 0.77 & 0.98 & 0.32 & 0.98 \\ 0.34 & 0.84 & 0.85 & 0.41 & 0.85 \\ 0.43 & 0.64 & 0.64 & 0.62 & 0.64 \\ 0.51 & 0.48 & 0.32 & 0.90 & 0.31 \\ 0.28 & 0.77 & 0.98 & 0.31 & 0.99 \end{pmatrix} \end{array}$$

3.4 메타데이터 구축

시멘틱 웹을 실현하기 위한 주요 요소인 온톨로지 구축 작업을 위한 툴로 OntoEdir, OilEd, Protege 등의 제품들이 개발되어 사용되고 있다. 이중에서 GUI 편집환경을 갖는 Protege는 온톨로지 저작툴로서 국내외에서 많이 이용되고 있다. 따라서 본 연구에서는 Protege를 이용하여 온톨로지를 구축하고 이를 이용한 지식처리가 가능하도록 하였다.

온톨로지 도메인과 범위 설정은 귀금속·보석 상거래 사이트 및 국제상품분류코드를 기반으로 조사한 내용으로 한정하였다. 중요 용어목록 설정은 귀금속·보석관련 웹사이트를 검색로봇을 통해 수집하고 용어들에서 대표용어집합을 생성하였다. 또한 용어간의 관계정도에 따른 유사관계시소러스를 생성하였다. 유사관계 시소러스를 통해 용어를 계층화하고 관련 산업체 종사자와의 피드백을 통해 클래스 계층구조를 결정하였다. 먼저 클래스로 신상품, Gold 명품, 웨딩주얼리, 베이비 주얼리, 캐릭터주얼리, 남성주얼리 등으로 구분하였다. 〈그림 3〉는 클래스의 계층구조를 나타내는 화면이다.

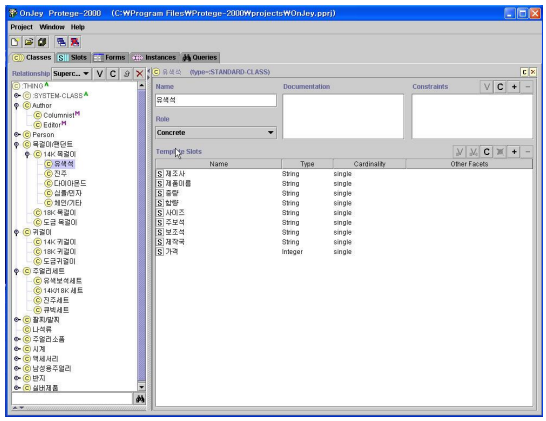

그림 3. protege-2000에서의 클래스 계층구조
Fig 3. class hierarchy on protege-2000

〈그림 3〉의 계층구조를 기반으로 클래스의 슬롯과 속성을 표현하였다. 각 클래스의 서브클래스와 슬롯은 〈그림 4〉와 같이 구성한다.

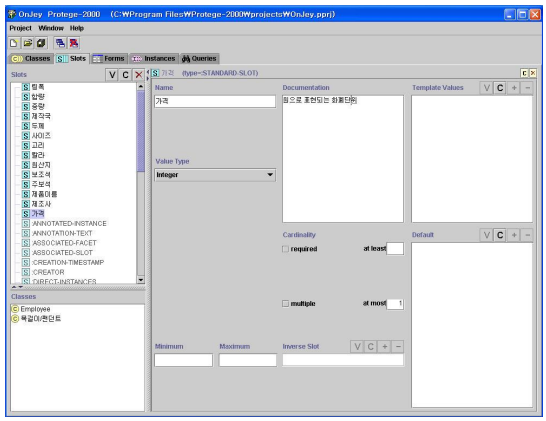

그림 4. 클래스-슬롯의 속성
Fig 4. classes and slots properties

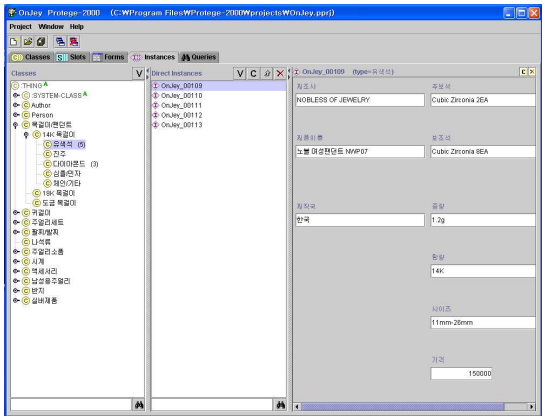

그림 5. 인스턴스 생성
Fig 5. instances on protege-2000

또한, 작성된 온톨로지 생성모듈에 의한 인스턴스를 생성한다. 〈그림 5〉은 자동 생성된 인스턴스의 생성 화면이다.

3.5 에이전트 개발

시멘틱 검색은 도메인 지식을 표현하는 온톨로지를 기반으로 웹페이지의 숨겨진 의미를 추론하여 향상된 검색이 가능하도록 지원한다. 온톨로지 기술과 에이전트 시스템의 연계를 통한 의미정보 추론 및 의미검색을 지원하기 위하여 온톨로지 표현언어는 도메인 관련 전문가가 도메인체계를 잡아야 하며 귀금속·보석 산업의 종사가가 직접 피드백을 주어야 하는 체계가 이루어져야 한다.

〈그림 6〉와 같이 지능형 에이전트시스템의 프로세스에서 온톨로지 기술이 접목될 것이다.

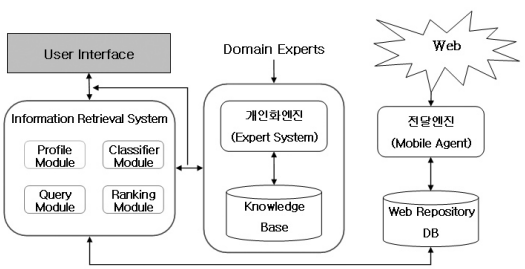

그림 6. 지능형 에이전트 시스템의 프로세스
Fig 6. process of intelligent agent system

3.5.1 프로파일 확장 검색 에이전트

지능형 에이전트시스템에서 개인화 검색을 실행하기 위한 가장 기본적인 데이터를 제공하는 사용자 프로파일의 구성과 사용자 기호학습모듈을 설계하여 시멘틱 검색시스템을 사용하는 사용자 개개인의 검색행위에 대한 데이터 및 사용자 기호 학습을 통하여 개인화된 사용자 검색을 지원하게 된다.

사용자 프로파일은 사용자 개인의 선호도뿐만이 아니라 다양한 검색패턴이 저장되고 사용자 검색 결과집합에서 대표용어와의 내용기반 확장 기법을 이용한다. 프로파일 확장은 다음 (식 3-3)과 같이 프로파일용어집합(P)과 대표용어(S_r) 사이의 퍼지 합성에 의해 생성되며 이와 같은 과정을 통해 사용자 프로파일이 도메인 영역으로 확장된다.

$$P_r = P^T \otimes R_r$$
$$s_{ij} = \bigvee_{i,j=1.n} (\min(w_{in}, I_{nj}), (\min(1-w_{in}), (1-I_{nj})) \quad \text{--(식 3-3)}$$

s_{ij} : 프로파일 용어i와 대표용어j의 유사 정도
w_{in} : 프로파일 문서n에서 용어i의 중요 정도
I_{nj} : 대표용어행렬 문서n에서 용어 j의 중요정도

3.5.2 질의확장 에이전트

시멘틱 웹 질의 모듈은 질의 분석기로부터 검색질의어와 질의어 의미를 전달받아 시멘틱 검색 모듈이 처리할 수 있는 RDQL로 재구성한다. 사용자에게 검색 질의를 전달받아 의미분석을 수행하기 위해 질의를 확장하고 재구성된 트리플 패턴은 사용자가 요구하는 검색 결과를 바인딩하기 위한 변수 및 검색 조건을 표현함으로서 개인화 검색 결과를 지원한다. 사용자 정보 요구에 대한 질의는 도메인 지식을 확장하기 위해 시소러스와 퍼지 합성을 통해 확장된 질의베이스로 구성된다. 질의베이스(Q_r)는 다음 식(3-4)와 같이 사용자에 의해서 표현된 질의(Q)와 대표용어 집합과 문서와의 퍼지 관계를 나타내는 문서베이스(S_r) 사이의 퍼지 합성에 의해 생성되며 의미분석을 통한 확장질의로 재구성한다.

$$Q_r = Q \circ S_r$$
$$\mu_{Q \circ S_r}(x,z) = \underset{y \in Y}{Max} \{ Min \{ \mu_Q(x,y), \mu_{S_r}(y,z) \} \} \quad \text{-- 식(3-4)}$$

단, $\mu_Q(x,y)$: 사용자 질의와 색인어와의 관계
$\mu_{S_r}(y,z)$: 문서와 축소용어 집합과의 관계정도
$\mu_{Q \circ S_r}(x,z)$: 질의와 축소용어 집합과의 관계정도

Ⅳ. 시멘틱 웹 기반 개인형 상품검색에이전트 개발

4.1 시스템 개발

본 연구에서는 사용자 관심도를 문서에서의 단순 키워드 매칭으로 검색하는 시스템의 문제점을 해결하기 위한 방안으로 도메인 온톨로지를 자동으로 생성하고 이를 기반으로 지능형 에이전트기술을 접목한 시스템을 설계하고 구현한다.

본 연구에서는 나라장터(http://www.g2b.go.kr)를 포함한 전문쇼핑몰 13개의 상품정보를 등록하고, 대표용어를 추출하여 내용분석을 수행함으로서 도메인 온톨로지를 구축하였다.

〈그림 7〉은 시스템의 전체적인 구조이다.

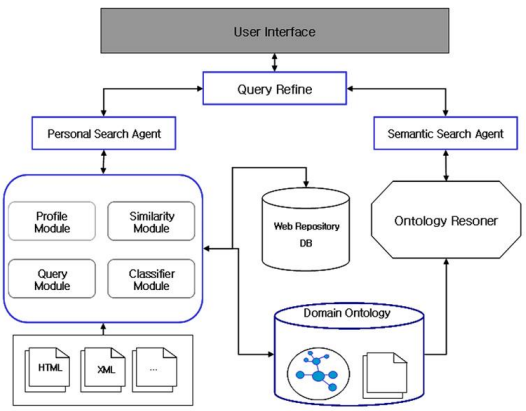

그림 7. 시멘틱 웹 기반 개인형 검색 에이전트
Fig 7. personalized retrieval agent for semantic web

〈그림 7〉에서 웹 검색 에이전트 모듈과 온톨로지 생성 및 관리를 위한 모듈로 나눌 수 있다. 온톨로지 기반의 검색시스템은 사용자 인터페이스, 온톨로지 저장소, 온톨로지 생성엔진 및 웹서버 엔진으로 구성된다. 본 연구에서는 도메인 용어 생성 및 유사관계 시소러스를 기반으로 용어 계층화를 수행하였으며 온톨로지 구축을 위한 저작도구로 protege-2000을 이용하였다.

4.2 검색 에이전트 구현

프로그램 개발은 웹서버가 여러 플랫폼에서 작동함으로 프로그램 도구로 자바를 이용하였다. 그리고 RDF 검색을 위하여 Jena를 이용하였다. 프로그램에 사용된 프로그램 도구로는 자바를 중심으로 하였고 인터페이스 부분은 Jsp를 이용하였다.

〈그림 8〉는 상품등록을 통한 내용을 분석하는 과정인 내용분석모듈이며, 상품을 등록하여 내용분석을 통한 계층화 인터페이스이다.

이와 같은 과정을 통하여 상품 용어를 분류하고 대표 개념을 생성하는 과정을 통해 계층구조에 따른 온톨로지를 구현한다.

그림 8. 상품 등록 및 내용분석 모듈
Fig 8. registration and content analysis module

〈그림 9〉은 (식 3-3)의 사용자 프로파일 의미정보를 내용분석하여 재구성하고 사용자 질의를 확장하여 검색을 수행하는 시멘틱 검색 화면이다.

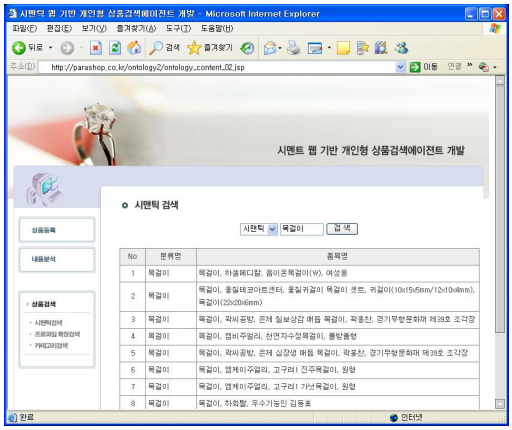

그림 9. 프로파일 확장을 통한 시멘틱 검색 모듈
Fig 9. a semantic retrieval module through profile expansion

〈그림 10〉는 기존의 클래스 카테고리 정보를 이용하여 검색할 수 있는 카테고리 검색 환경을 나타내는 인터페이스이다.

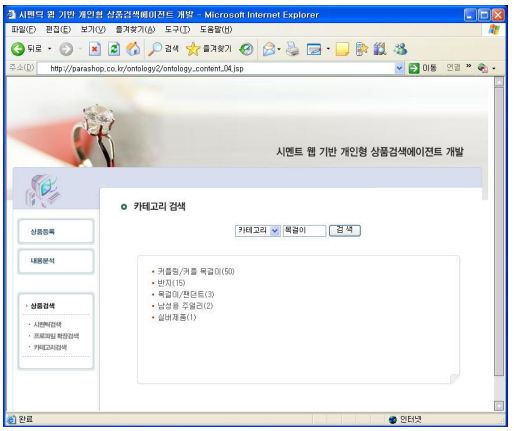

그림 10. 카테고리 검색
Fig 10. classification retrieval

V. 결 론

기존의 전자상거래시스템의 물품검색과정은 동의어, 다의어 처리문제와 선별되지 않은 방대한 상거래관련 정보들에서 사용자가 반복적인 피드백 필터링 과정 등과 같은 많은 시간을 투자하여 정보를 습득하고 검토하여야 하는 어려움이 있다. 또한 검색시스템의 낮은 기능성으로 인하여 수요자와 공급자 모두에게 많은 부하를 요구함으로 결과적으로 전자상거래의 보급과 한계를 가져오게 한다. 따라서 본 논문에서는 도메인 온톨로지를 자동으로 생성하고 이를 기반으로 지능형 에이전트기술을 접목함으로서 의사소통이 단일화된 상품검색시스템을 개발하였다. 또한, 사용자 프로파일과 유사관계 시소러스와의 매핑을 통해 프로파일을 확장하고 개인화 검색을 지원하기 위해 적용하였다.

향후 연구 방향은 온톨로지 용어사전 구축에 있어 도메인 특성에 매개변수를 자동화하는 연구와 대표용어 집합에 대한 신뢰성을 향상시키는 연구를 계속할 것이며 전문가 집단의 팩트를 효율적으로 사용자 정보와 매핑하는 연구를 수행하여 자동화되고 최적화된 검색 결과 순위를 제공할 계획이다.

참고문헌

〔1〕 권혁철, "시멘틱 웹의 가능성과 한계", 지식정보인프라, 통권 15호, pp.15-19, 2004.

〔2〕 최옥경, 한상용, 오길록, "자동화된 통합 프레임워크를 위한 시멘틱 웹 기반의 정보검색시스템", 정보처리학회 논문지, 제13-C권 제1호, 2006.2, pp. 129-136.

〔3〕 노상규, 박진수, "인터넷 진화의 열쇠 온톨로지 웹 2.0에서 3.0으로" 가즈토이, 2007.2.

〔4〕 T.R.Gruber, "Toward Principles for the design of ontologies used for Knowledge Sharing", Int. J.Human Computer Studies, Vol.43, pp.907-928, 1995.

〔5〕 이은석, "멀티 에이전트 기술의 실세계 시스템으로의 응용", 정보과학회지, 15권 3호, pp.17-28, 1997.

〔6〕 윤지웅, "인터넷 환경하에서 지능형에이전트 현황 및 전망", 정보통신정책연구원, 1998.12.

〔7〕 구미숙, "데이터마이닝 기법을 이용한 XML 문서의 온톨로지 반자동 생성", 정보처리학회논문지, 제13-D권 제3호, pp.299-308, 2006.6.

〔8〕 은희주, "퍼지함수와 관계성을 적용한 질의 확장 및 문서분류 시스템", 전북대학교 대학원 박사학위논문, 2003.8.

〔9〕 김창민, 김용기, "퍼지 관계곱 기반 퍼지정보 검색 시스템 구현", 정보처리학회 논문지, 제8-B권 제2호, pp. 115-122, 2001.4.

〔10〕 Laszlo T. Koczy, T. D. Gedeon, "Information retrieval by fuzzy relations and hierarchical co-occurrence," Part Ⅰ. TR97-01, Dept. of Info. Eng., School of Comp. Sci. & Eng., UNSW, 1997.

〔11〕 Laszlo T. Koczy, T. D. Gedeon, "Information retrieval by fuzzy relations and hierarchical co-occurrence," Part Ⅱ. TR97-03, Dept. of Info. Eng., School of Comp. Sci. & Eng., UNSW, 1997.

〔12〕 이광형, 오길록, "퍼지 이론 및 응용", 홍릉 과학 출판사, 1991.

〔13〕 Shyi-Ming Chen, Yih-Jen Horng, "Fuzzy Query Processing for Document Retrieval Based on Extended Fuzzy Concept Networks," IEEE Transactions on Systems, MAN. and CyberNetics-Part B: CyberNetics, Vol. 29, No. 1, February, 1999.

〔14〕 Shyi-Ming Chen, Jeng-Yih Wang, "Document Retrieval Using Knowledge-Based Fuzzy Information Retrieval Techniques," IEEE Transactions on Systems, MAN. and CyberNetics, Vol. 25, No. 5, May, 1995.

저 자 소 개

이기영(Ki-Young Lee)

1992년 2월 광주대학교 컴퓨터과학과 졸업(이학사)

1994년 2월 전북대학교 전산통계학과 졸업(이학석사)

2005년 2월 전북대학교 전산통계학과(이학박사)

1998년 3월~현재 원광보건대학 영상컨텐츠과 교수

※ 관심분야 : 퍼지 클러스터링, 정보검색, 데이터 마이닝