의미적 연결 관계에 기반한 전자 카탈로그 검색용 유사도 척도

(A New Similarity Measure for e-Catalog Retrieval Based on Semantic Relationship)

서 광 훈 † 이 상 구 ††
(Kwang-hun Seo) (Sang-goo Lee)

요 약 전자 상거래의 발달과 함께 B2B Market Place의 등장과 통합으로 전자 상거래의 중심 단위인 전자 카탈로그의 양도 급증하고 있다. 이러한 전자 카탈로그의 정보의 질적, 양적 증가는 상품 정보 검색의 난이도를 높이고 있다. 특히, 대량 거래를 하는 상품 전문가의 의사 결정을 위해 단일 분류 체계가 아닌 다양한 분류체계 내에서의 상품 정보 검색을 지원하는 시스템의 필요성이 증가하고 있다. 하지만 기존의 검색 시스템은 일반 문서 검색 시스템이 대다수이며, 이러한 전자 카탈로그의 특성을 반영하지 못하고 있어 이를 지원하기에는 한계가 있다. 따라서 본 논문에서는 전자 카탈로그가 지니고 있는 속성적, 어휘적인 특성을 반영하고 의미적 연결관계에 기반한 검색을 통하여 해당 요구 사항을 충족시킬 수 있는 시스템의 토대를 마련하고자 하였다. 이를 위해, 전자 카탈로그의 특징을 반영한 전자 카탈로그 기본 모델을 제시하고, 검색을 결과 제시를 위한 유사도 평가 요소를 도출하였으며, 정확성 향상을 위해 이를 어휘적 특성을 고려한 데이타 확장 모델 및 어휘 기반 유사도 평가 요소로 확장하였다. 그리고 제시한 모델을 통해 의미적 연결 관계에 기반한 전자 카탈로그 유사도 평가 함수를 제시하고 이를 전자 카탈로그 정보 검색시스템으로 구현하고 검증하였다.

키워드 : 전자카탈로그, 전자상거래, 유사도 평가함수, 상품, 정보검색, 의미적 연결관계

Abstract The e-Marketplace is growing rapidly and providing a more complex relationship between providers and consumers. In recent years, e-Marketplace integration or cooperation issues have become an important issue in e-Business. The e-Catalog is a key factor in e-Business, which means an e-Catalog System needs to contain more large data and requires a more efficient retrieval system.

This paper focuses on designing an efficient retrieval system for very large e-Catalogs of large e-Marketplaces. For this reason, a new similarity measure for e-Catalog retrieval based on semantic relationships was proposed. Our achievement is this: first, a new abstract e-Catalog data model based on semantic relationships was designed. Second, the model was extended by considering lexical features (Especially, focus on Korean). Third, the factors affecting similarity with the model was defined. Fourth, from the factors, we finally defined a new similarity measure, realized the system and verified it through experimentation.

Key words : e-Catalogs, e-Business, IR, Semantic, Ontology

· 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT 연구센터 지원 사업(IITA-2007-C1090-0701-0031)의 연구결과로 수행되었음

† 정 회 원 : 서울대학교 전기.컴퓨터공학부
longago@europa.snu.ac.kr
†† 종신회원 : 서울대학교 전기.컴퓨터공학부 교수
sglee@europa.snu.ac.kr

논문접수 : 2005년 2월 14일
심사완료 : 2007년 10월 4일

정보과학회논문지: 데이타베이스 제34권 제6호(2007.12)

1. 서 론

인터넷 발달과 함께 전자 상거래는 이제 삶의 한 부분으로 자리 잡기 시작했다. 전자 상거래 환경에서는 상품 및 서비스에 대한 정보를 구조화된 형태로 담고 있는 전자 카탈로그에 기반하여 대다수의 거래 행위가 이루어진다. 전자 카탈로그를 중심으로 이를 단순화 해보면, 구매자가 상품을 고르는 행위는 유통사의 전자 카탈로그 시스템상에서 자신에게 적합한 상품 정보를 구성하고 있는 하나의 단위 상품 정보(=전자 카탈로그)를 찾는 것(Retrieval)으로, 판매자가 물품을 시장에 내어 놓는 것은 해당 시스템 상에 새로운 전자 카탈로그를 배포(Publish) 하는 것으로 볼 수 있다. 따라서, 전자 상거래는 하나의 정보 검색의 연장선 상에 놓여 있다 하겠다.

전자 상거래 환경이 B2B Market Place 와 같이 복잡하고 다양한 거래 환경을 지원하는 시장으로 성장하였고, 이들 간의 연동 및 통합이 이슈화되면서 시스템에서 처리해야 할 전자 카탈로그의 규모가 커졌다. 또한 개별적으로 성장한 Market Place의 특성 상 시스템 별로 상이한 상품 분류 체계를 가지고 있어 이들 간의 연동과 통합을 더 복잡하게 하고 있다. 이러한 규모적 성장과 시스템간 상이한 분류체계는 다품종의 상품 거래 시 필요한 전자 카탈로그 검색의 복잡성을 더욱 심화시키고 있지만, 기존의 전자 카탈로그 검색 시스템은 웹 검색의 기법을 그대로 적용한 경우가 많아 처리에 한계를 보이고 있다. 즉, 전자 카탈로그는 다량의 정보를 포함하는 일반 문서와 달리 적은 량의 정보가 다양한 속성으로 구분하여 저장하고 있어 검색 효율성이 떨어질 수 있으며, Market Place 환경에서 필요로 하는 상품 속성 기반의 분류 정보 검색 등 다면적인 정보 검색 방식을 충족시키기 어렵다. 따라서, 전자 카탈로그의 특성을 반영하고, 이러한 요구 사항을 수용할 수 있는 전자 카탈로그 검색 시스템의 고안이 필요하다 하겠다.

이러한 전자 카탈로그 검색 시스템의 구축을 위해서는 질의어와 검색 대상간의 유사도 평가를 위한 요소 및 평가 함수를 제시하여, 검색 시스템에 적용하는 방법론이 필수적이다. 따라서 본 논문은 이를 위해 전자 카탈로그 내포하고 있는 의미적 연결 관계와 특수성에 기반하여 새로운 데이타 모델 및 유사도 평가 함수를 제시하는데 그 목적을 두었다.

본 논문의 연구 내용 및 범위는 다음의 네 가지로 요약된다.

첫째, 의미적 연결 관계 기반의 전자 카탈로그 논리적인 데이타 모델을 제시하였다.

둘째, 제시한 데이타 모델을 이용한 전자 카탈로그 유사도 평가 요소를 제시하였다.

셋째, 전자 카탈로그의 어휘적 특성을 반영하는 확장된 데이타 모델을 제시하고, 이를 반영하기 위하여 관련한 유사도 평가 요소를 추가하였다.

넷째, 의미적 연결관계 기반 시스템을 위한 유사도 평가 함수의 제시 및 구현하고 실험을 통해 해당 알고리즘의 유효성을 단계별로 검증하였다.

2. 관련 연구

초창기 전자 카탈로그 관련 연구는 '종이 카탈로그를 어떻게 웹에 표현 하는가?' 혹은 '상품 정보를 어떻게 DB화하는가? '에 관한 관심이 높았으며 이는 전자 카탈로그에 대한 데이타 모델에 관한 연구[1,2]와 UI에 관한 연구[3] 등으로 지속 되어 오고 있다. 근래에는 Market Place 의 등장과 이들 간 연동이 이슈화됨에 따라 상품 분류에 대한 연구가 쟁점이 되고 있다. 전자 카탈로그 구축을 위한 분류체계 모델에 관한 연구[4]나 상품의 자동 분류 및 분류체계 간 통합에 관한 연구[5-8], 룰 기반의 분류 체계간 통합에 관한 연구[9]등이 제시되고 있다. 또한 온톨로지에 대한 연구와 함께 분류체계의 의미 기반 모델[10], 상품 온톨로지 구축에 관한 연구[11]가 부각되고 있다.

하지만, 검색 관련 연구는 아직 미흡한 편이다. 상용화 된 전자 카탈로그 검색 시스템은 주로 일반적인 문서 검색 모델을 그대로 적용하였거나, 직접적인 데이타베이스 검색에 따르는 경우가 많아, 주로 상품명이나 분류명에 국한한 검색만을 지원하는 경우가 많으며 검색 결과의 유사도 평가 또한 전자 카탈로그의 특성을 고려한 모델은 찾기 어렵다. 특히, 다품종 거래에 있어서는 상품 정보와 연계하여 분류 및 속성에 관한 검색이 필요하지만, 상품 혹은 속성을 통한 분류의 검색 등 상이한 검색 범주 간의 검색을 지원하는 시스템이나 관련 연구는 찾아 보기 어려웠다. 그리고, 의미적 연결 관계에 대한 관련 연구 또한 전자 카탈로그 상의 의미적 연결 정보에 대한 모델링에 관한 연구는 제시되고 있으나 검색에 대한 방법론의 제시는 미미한 상태다. 온톨로지 환경에 있어서도 주로 탐색(Navigation)[12] 혹은 온톨로지 질의어 모델을 통한 검색[13]에 관한 연구는 있으나 키워드를 통한 검색 모델은 찾아보기 어려웠다.

따라서 본 연구에서는 상이한 검색 범주 간의 검색 등을 지원할 수 있도록 의미적 연결 관계 기반의 전자 카탈로그의 데이타 모델을 구성하고, 키워드 기반의 효율적인 전자 카탈로그 검색 시스템 구현을 위해 데이타 모델 확장 및 유사도 평가함수 고안에 초점을 두었으며 이를 기존 정보 검색 모델을 기반으로 구현함으로써 관련 연구에 대한 초석이 되고자 하였다.

또한 본 연구에서는 주요 쟁점이 되고 있는 분류 문제를 상품 키워드 기반의 검색으로 접근하고자 하였다. [5]의 경우 상품 정보를 속성별로 규격화하여 연구 성과를 높였지만 학습된 분류체계와 동일하게 속성별로 규격화된 정보만을 지원하는데 반해, 본 연구는 일반 키워드를 사용함으로써 이러한 제약을 풀었다. 또한, 기존의 자동 분류 연구가 주로 단일 분류 체계만을 지원하지만, 본 연구에서는 다양한 분류로의 검색을 지원하도록 고안되었다.

3. 문제 정의 - 데이타 모델링

3.1 전자 카탈로그 모델

전자 카탈로그는 상품(Product), 속성(Attribute), 분류(Category) 정보로 구성 되어 있으며, 이들 각각은 자신을 정의하는 속성과 속성값의 벡터 집합으로 정의할 수 있다. 또한 분류가 상품 정보의 집합으로 정의되고, 상품이 여러 속성의 집합으로 정의되며, 속성은 상품이 가지는 속성값의 집합으로 정의되는 것과 같이 서로 포함 할 수 있는 상호 연결 관계를 지니고 있다. 본 논문에서는 이러한 상호 연결관계를 그림 1과 같이 '의미적 연결 관계'라고 정의하고 세부적인 전자 카탈로그 구성을 Def. 1과 같이 정의하였다.

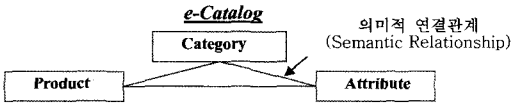
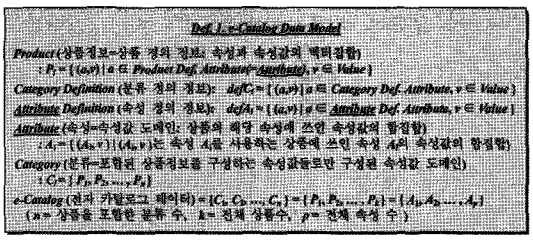


그림 1 전자 카탈로그 요소 간의 연결관계 도식



3.2 Semantic Data Model

앞선 정의에 따라 전자 카탈로그의 정보는 단위 요소(Entity)와 이들 간의 의미적인 연결 관계에서 추출 가능한 정보로 다시 추상화 할 수 있다. 즉, 상품/분류/속성의 정의 정보를 Entity로, 의미적 연결관계에 놓여 있는 정보를 Related Entity Information이라고 정의하면 다음과 같으며 이를 통해 하나의 요소 정보를 관련 요소의 정보로도 표현 된다.

Entity = *Product* | *Category* | *Attribute*

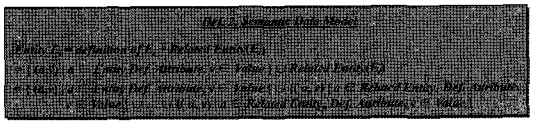
Related Entity Information

: *Related Entity* (E_i) = *Entity* E_i와 연관관계를 맺고 있는 *Entity*의 정의 정보

예를 들어, 상품 정보의 Related Entity Information을 수식화하면 다음과 같다.

Related Entity (Product) = *Category Def. Information* ∪ *Attribute Def. Information*

이를 본 논문에서는 Semantic Data Model이라 하고 Def. 2와 같이 정의하였다.



3.3 Semantic Data Model에 기반한 유사도 평가 요소

일반 문서의 정보 검색을 위한 데이타 모델과 비교할 때, 앞서 제시한 Semantic Data Model은 그림 2와 같이 다양한 속성으로 구성되어 있는 전자 카탈로그의 특성(Multi Attributed Entity Information)을 포괄하고 의미적 연결 관계를 지니는 정보(Related Entity Information)로 확장한다는 점에서 차이점을 지닌다.

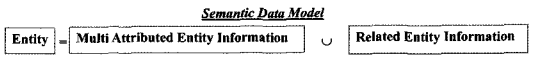


그림 2 Semantic Data Model 도식

여기에서, 하나의 요소를 정의하는 정보가 여러 개의 속성으로 구분하여 구성한다. 이는 각각의 단위 속성에 따라 속성에 포함된 속성값이 해당 카탈로그를 대표할 수 있는 자질이 서로 상이할 것이라는 가정에 따른 것으로 [5]에서 전자 카탈로그 분류에 유효한 영향력을 끼침을 보인 바 있다. 그리고 의미적 연결관계를 지니는 정보 또한 구분하여 반영한다. 이는 그 영향력이 상이할 것이라는 가정에 따른 것으로 예를 들면 'TV' 이라는 키워드로 분류를 검색 할 때, 분류명에 쓰인 'TV'와, 자동차 상품의 'TV' 유무를 구분하기 위해 속성에 쓰인 'TV'는 검색 결과에 주는 의미가 다르다는데 기인한다. 이에 따라, 질의어와 검색 대상간의 유사도 평가 요소를 속성적 요소(Attribute factor)와 의미적 연결 관계적 요소(Related Entity factor)로 나누어 볼 수 있다고 보고 Def. 3과 같이 정의하였다.

3.4 전자 카탈로그의 어휘적 환경에 따른 모델 확장

일반 문서 검색의 경우, 문서 자체가 가지는 정보량이 방대하기 때문에, 문서 전체 정보를 담기 보다 그 문서를 특정 지우는 키워드를 추출하는 방식이 주로 채택되어 왔지만, 전자 카탈로그는 단위 속성별로 구분된 정보로 지니는 정보의 량은 매우 적기 때문에 이러한 접근 방법론은 한계가 있다. 따라서, 이를 극복하기 위해 정보량을 늘려주는 접근 방법론이 필요하다. [5]에서는 제시된 어휘 확장 모델은 어휘를 어휘 확장함수에 따라 공백이나 쉼표(,) 등으로 구분되는 단위로 분리하고, 복합명사로 유추 되는 경우에는 단순명사를 추출하며, 불용어를 배제시키는 등의 방법으로 어휘자원을 풍부하게 하거나, 검색에 유효한 어휘자원을 추출하는 방식을 제시하였다. 하지만, 단일 분류 체계 만으로의 분류를 지원하는 해당 자동 분류 시스템과 본 논문에서 고안하고자 하는 검색 시스템은 목적이나 처리 방식이 상이하여 이를 다음과 같이 변경하여 차용하였다.

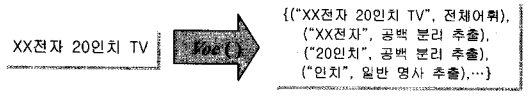
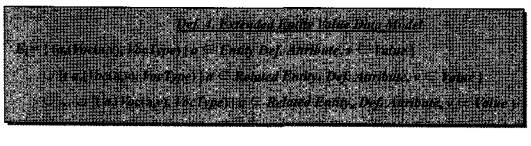


그림 3 어휘 확장 함수의 적용 예

먼저 어휘 확장에 따른 결과 키워드가 검색 상 지니는 중요도가 상이하다고 보고 어휘 확장함수(*Voc( )*)를 그림 3의 예와 같이 어휘 및 확장 방식에 대한 부가 정보(*VocType*)를 함께 도출해 내는 것으로 재정의하였다. 또한, 단순 상품 정보만을 대상으로 하는 [5]와 달리 분류의 설명이나 속성의 설명과 같이 일반 문서처럼 긴 문장의 정보를 지니는 경우도 있어 이 경우는 확장이 아닌 키워드를 추출하는 것이 효율적이다. 따라서, 함수의 수행 방식이 속성에 따라 상이하도록 어휘 확장 함수(*Voc( )*)의 파라메터를 추가하였다.

이렇게 재정의된 어휘 확장 함수(*Voc(Attribute, Value)*)를 이용하여 기존의 데이타 모델이 지니고 있는 속성값(Value)을 확장한 모델을 다음의 Def. 4로 정의하였다. 여기에서 *VocType*은 어휘 확장 함수가 도출하는 어휘 확장 방식에 대한 부가 정보를 의미한다.



또한, 키워드 검색을 목적으로 하고 있으므로 사용자의 질의어(q) 역시 어휘 확장 함수의 속성에 대한 파라메터 값을 'query' 상수로 하여 추출 되는 질의어 확장 집합(Q)으로 재 정의한다. 이에 따라 확장된 질의어 키워드 모델을 다음의 Def. 5와 같이 정의하였다.

끝으로, '20인치'라는 어휘가 질의어와 일치하는 것과 '20인치'에서 확장된 '인치'라는 어휘에서 일치하는 것이 주는 의미가 다를 것이라는 점에 착안하여, 추출 방식에 따라 어휘가 대상을 표현하는 표현력이 차이를 지닐 것으로 가정하였다. 이에 따라, 해당 어휘에 대한 확장 방식 정보 (VocType)를 추가적인 유사도 평가 요소인 어휘적 요소 (Vocabulary factor)로 다음의Def. 6과 같이 정의하였다.

4. 유사도 평가 함수

4.1 유사도 평가 함수의 정의

3장에서 정의된 유사도 평가 요소를 평가 함수화 하기 위해 가중치 상수로 재정리하면 표 1과 같으며, 이에 기반한 유사도 평가함수 및 순위화 알고리즘은 다음과 같이 정의 된다.

표 1 유사도 평가 요소 및 가중치 상수의 정의

유사도 평가 요소	가중치 상수
Attribute factor	Attribute weight = ***A***
Related Entity factor	Related Entity weight = ***R***
Vocabulary factor	VocType weight = ***V***

1) 검색을 위한 키워드형 질의어를 Query ***Q***, 최종 검색 대상을 Entity ***E***로 정의한다.

2) 질의어 ***Q***는 Def. 5에 의해 어휘 확장된 질의 키워드 ***q***의 집합으로 재정의된다.

$$Q \rightarrow Voc(\text{'query'}, Q) \rightarrow \{(q, VocType) \mid q = \text{확장된 질의어 키워드}\}$$

3) 검색 대상 ***E***는 Def. 4에 의해 다음과 같이 확장된 집합으로 재정의 된다.

$$E = \{(a,(Voc(a,v), VocType) \mid a \in Entity\ Def.\ Attribute, v \in Value\} \cup \{(a,(Voc(a,v), VocType) \mid a \in Related\ Entity1\ Def.\ Attribute, v \in Value\} \cup \ldots \cup \{(a,(Voc(a,v), VocType) \mid a \in Related$$

***Entityn Def. Attribute, v* ∈ *Value* }**

4) 또한 위와 같이 재 정의된 ***E***는 다음과 같은 집합으로 재정의 할 수 있다.

***E* = { (*value, EntityType, AttType, VocType*) }**

- ***value*** **: 어휘 확장 함수(*Voc()*) 에 의해 도출된 키워드**
- ***Entity Type*** **:** ***value*** **가 소속 되있던 Entity 정보**
- ***Attribute Type*** **:** ***value*** **가 소속 되있던 속성정보**
- ***VocType*** **:** ***value*****를 확장한 방식에 대한 정보**

5) 여기에서 질의어와 검색 대상 유사도를 평가하기 위한 비교대상은 질의어 ***Q***와 검색 대상 ***E***에서 최종 도출되는 모든 q_i와 모든 $value_k$ 간의 비교이다. 따라서 유사도 평가함수(*Related Score()*)를 다음과 같이 정의한다.

$$Related\ Score\ (Q,\ E) = \sum Related\ Score\ (q_i,\ value_k)$$

6) 또한 q_i는 ***VocType*** 정보를, $value_k$는 ***EntityType, AttType, Voc-Type***의 정보를 가지며 이는 표 1의 가중치상수로 대치된다. 즉, ***VocType***은 어떠한 어휘적 분해를 통해 도출되었는지를 나타내고 있는 하나의 어휘적 요소이므로 ***V, EntityType***은 어느 Entity에서 도출 되었는지를 나타내고 있는 것으로 의미적 연결관계를 대표할 수 있으므로 ***R, AttType***은 속성에 대한 것으로 ***A***의 가중치상수로 표현할 수 있다.

7) 그리고 valuek의 키워드 자체의 가중치 값, 즉 전자카탈로그를 대표하는 키워드로써의 의미를 VAL_{valuek}로 정의한다. 이는 TF/IDF 가중치 등 기존에 검증된 가중치로 재정의할 수 있으며, 이를 기반으로 불리언, 벡터, 확률, 추론망 모델 등 기존 정보 검색 모델에 기반의 정보 검색 프레임워크에 적용할 수 있다.

8) 이에 따라 유사도 평가함수는 다음의 Def. 7과 같이 정의된다.

Def. 7. Related Score Function

$$Related\ Score\ (Q,\ E) = \sum Related\ Score\ (q_i,\ value_k)$$
$$= \sum V_{qi} * (R\ value_k * (A\ value_k * (E\ value_k * VAL\ value_k)))$$

즉, 유사도 평가 함수는 확장된 질의어 단위요소인 q_i와 확장된 전자카탈로그의 최소 단위 정보인 $value_k$ 간의 유사도의 총합으로, 질의어 단위요소인 q_i에 어휘가중치 ***V***를, 전자카탈로그의 최소 단위 정보인 $value_k$에 키워드 가중치 ***VAL*** $value_k$, 속성 가중치 ***A****$value_k$*, Entity 가중치 ***E*** $value_k$를 적용한 유사도의 총합이다. 따라서 유사도를 키워드, 어휘, 속성 및 의미적 연결관계의 측면에서 세분화된 가중치를 통해 접근할 수 있도록 하였다.

9) 끝으로, 제시된 유사도 평가 함수에 따라 정보검색 모델에 적용할 수 있는 순위화 알고리즘을 다음의 Def. 8과 같이 정의한다. 이를 기반으로 의미적 연결관계에 기반한 전자 카탈로그 검색 시스템을 구현하였으며, 구현된 시스템에서 8)에서 제시한 각 가중치를 변용하는 실험을 통해 가중치가 지니는 의미를 검증해 보고자 하였다.

Def. 8. Ranking Algorithm

R(q,d) = *Related Score (q,d)*가 평가한 항목의 점수를 역정렬(desc.)한 순위

4.2 유사도 평가 함수와 기존 정보 검색 모델과의 관계

유사도 평가함수는 정보 검색 모델의 순위화 알고리즘에 적용됨으로써 기존 정보 검색 모델에 적용될 수 있다. 제시한 정의된 유사도 평가 함수는 어휘에 대한 정보 평가 척도인 ***VALvalue***$_k$으로 키워드를 기반으로 한 정보 검색 모델을 따르고 있다. 따라서 벡터모델을 중심으로 확률 모델, 나아가 추론망 모델에도 쉽게 반영 될 수 있다. 벡터 모델의 경우에는 ***VALvalue***$_k$의 대입으로 적용가능 하지만 확률 모델 및 추론 모델은 타 가중치와의 조정이 필요하다. 벡터 모델의 경우 0(불일치), 1(일치)의 값에서부터 다양한 어휘빈도에 기반한 값으로까지 단순히 값을 대입할 수 있지만, 확률 모델과 추론망 모델의 경우 확률 환경을 기반으로 하므로 각 가중치 설정 시 최종 확률 값 산출에 대한 사전 고려를 필요로 한다.

5. 구현 및 실험

5.1 데이타 모델의 구현

3장에서 정의된 Semantic Data Model 상의 연결 관계 정보는 그림 4의 각 단계를 거쳐 데이타베이스 모델에 적용 될 수 있다. 또한, 데이타의 확장은 그림 5와 같이 각 단계별로 데이타 추출 및 어휘확장을 통해 이루어질 수 있다. 또한 이러한 과정을 통해 추출된 키워드 정보는 검색을 위한 색인정보로써 그림 6의 색인정보와 같이 구현할 수 있으며, 이를 기반으로 의미적 연결관계를 이용한 전자 카탈로그 정보 검색 모델은 그림 7의 기본모델과 같이 구현할 수 있다. 또한, 그림 8 같이 각각의 분류체계 정보를 추가함으로써 상품 정보 키워드를 이용한 다중의 분류체계에 대한 검색을 구현할 수 있다.

5.2 키워드 가중치의 적용

본 연구에서의 실험은 유사도 평가 함수의 적용은 벡터 모델에 의거하여 진행 되었다. 유사도 평가 함수의 가중치(***V, R, A***)는 개별적인 정수로 정의하고 키워드 가중치인 ***VALvalue***$_k$는 어휘 빈도수를 다음과 같이 적용하였다.

키워드 가중치(***VALvalue***$_k$) = 1 / 전체 결과 내에서 해당 어휘가 발견 되는 횟수

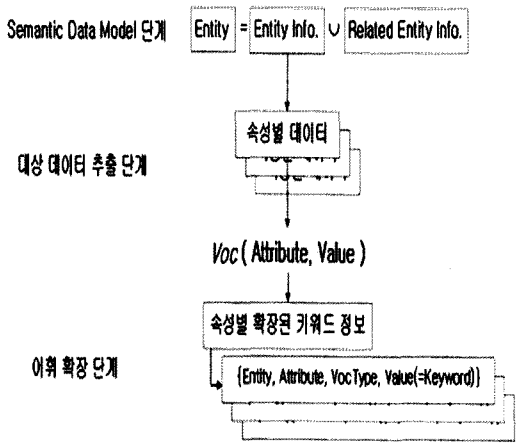


그림 4 Semantic Data Model의 구현 단계

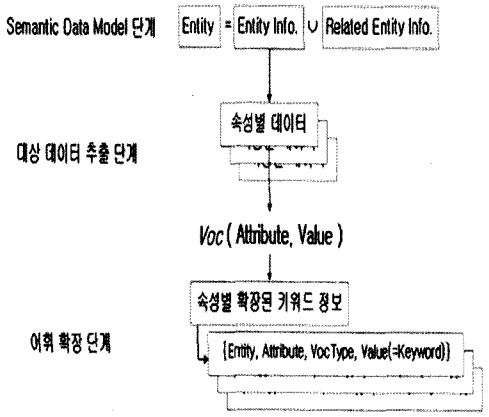


그림 5 데이타 구축 단계

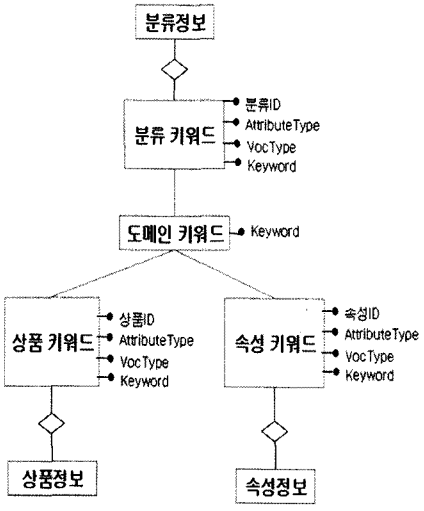


그림 6 색인정보

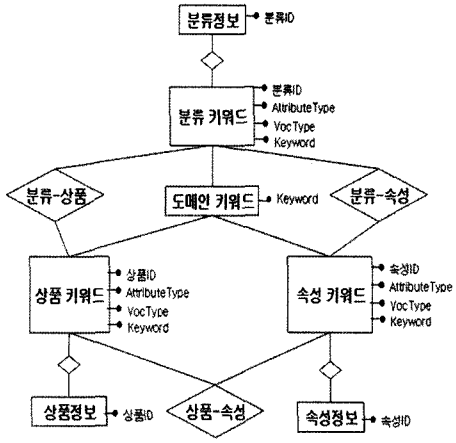


그림 7 기본 모델

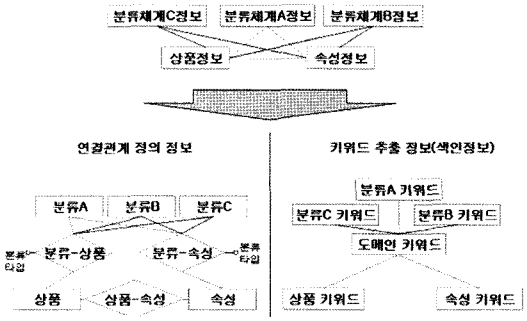


그림 8 다중 분류 체계 지원 모델

기존의 정보 검색의 경우 전체 시스템 상의 문서에 포함된 어휘 수로 어휘 수를 세는 기준이 제시 되었지만, 본 논문에서 제시된 검색 데이타 모델은 요소 간 관계에 따라 유연하게 확장 되므로 상황에 따라 그 기준이 바뀌게 되어 기존 정보검색 모델에서 연구된 키워드 가중치를 그대로 적용하는데 문제가 있다. 즉, 동일 상품의 어휘에 대해서 상품 검색의 경우에는 상품정보가 기준이 되지만, 상품과 연결된 분류에 대한 검색일 경우, 그 기준이 분류정보가 되어야 하므로 상황에 따라 그 기준이 가변적이다. 이러한 가변성을 극복하고, 실험을 단순화하기 위해, 본 실험에서는 한 어휘가 도출한 결과가 많을수록 그 어휘가 Entity를 특징지우는 정도가 상대적으로 낮다고 가정하고, 해당 어휘가 현재의 검색 대상을 도출하는데 사용되는 빈도수의 역수로 하였다. 즉, 해당 어휘가 도출하는 최종 검색 대상 정보 및 의미적 연결 관계를 지니는 정보에서 그 어휘가 발견되는 빈도수의 역수를 가중치로 적용하였다.

따라서 적용된 유사도 평가 함수는 정수로 표현된 각 가중치와 빈도수의 역수인 키워드 가중치를 곱한 값을

모두 합한 것을 검색 대상 요소의 유사도 점수로 보고, 점수가 가장 높은 것이 가장 질의와 인접한 결과로 보았다.

5.3 실험 데이타

실험은 국가 조달청에서 사용하고 있는 조달목록정보 카탈로그 중 일부를 이용하여 진행되었다. 이 전자카탈로그는 UNSPSC를 변용한 G2B 분류 체계를 기준으로 하여 8자리 분류코드로 구분된 G2B 분류체계 정보, 8자리 식별번호로 구분된 조달 표준 상품 정보 및 사용되는 속성에 대한 속성 정의 정보로 구성되어 있다. 또한, 상품 정보에는 국제 표준 분류체계 중의 하나인 UNSPSC와 관세 기준 표준 분류체계인 HS, 기존 조달청 분류 체계인 군급 분류체계 등 추가적인 타 분류 체계 정보가 포함되어 있다. 상품 정보는 조달 표준 상품이 정의된 정보로 상품명에 '모델명, 제조사, 대표규격'의 정보를 포함한 경우가 많아 상품명이 비교적 정련도가 높고 데이타 양이 풍부한 편이다. 또한 모든 상품이 사용하는 공통 속성과 상품별로 달리 사용되는 개별 속성으로 나뉘어져 관리되고 있다.

이중 실험에 사용된 데이타는 G2B 분류코드 상 표 3에 해당하는 상품 정보, 해당 분류의 정의 정보, 해당 분류에서 사용되고 있는 속성에 대한 정의 정보이며 이는 총 380개의 분류 정의 정보, 7510개의 상품 카탈로그, 1628 개의 속성 정의 정보로 이루어져 있으며 구체적인 예는 그림 9와 같다.

표 3 실험에 사용된 데이타의 G2B 분류코드 정보

4316XXXX	소프트웨어
4317XXXX	통신 및 컴퓨터하드웨어
4318XXXX	통신 및 컴퓨터소모품
4410XXXX	사무용기기 및 보조용품
4411XXXX	사무용 및 탁상용 부품
4412XXXX	사무용 소모품

(단, 상위 2단계까지만 표시함)

5.4 실험 설계

실험은 의미적 연결 관계에 따라 다른 Entity 정보를 얼마나 정확하게 검색 가능한가에 초점을 두고 상품의 일부 정보를 질의어로 해당 상품이 실제 속해 있는 분류를 첫 페이지 안에 제시 할 수 있는가를 기준으로 결과를 평가하였다. 예를 들면 'LG XCANVAS 42인치' 라는 질의어를 가지고 해당 상품이 실제 속한 'TV' 분류를 몇 번째 순위 결과로 제시하는 가를 측정하였다. 여기서 하나의 상품은 하나의 분류체계에서 하나의 분류에 속하므로 Recall의 의미는 적다고 판단되어 평가에서 배제하였다.

실험 데이타는 상품 정보를 균등하게 4분할 한 다음, 1개의 집합(1/4)을 질의어로 이용하고 3개의 집합(3/4)을 검색 대상으로 이용하였으며, 실험 결과는 각 집합을 한 번씩 질의어로 이용하여 총 4회 반복 실험한 평균으로 구하였다. 이때 질의어는 실험에 따라 제품명을 사용하는 경우와 상품의 전체 정보를 하나의 어휘열로 사용하는 경우가 있었다. 또한 검색 대상은 상품 정보 및 해당 상품에 해당하는 분류 및 속성 정의 정보를 이용하여 데이타 모델에 따라 구축하였다. 즉, 실험 데이타의 3/4에 해당하는 해당 상품 정보를 및 상품들이 소속된 분류 및 속성에 대한 정의 정보를 어휘확장함수를 통해 키워드화하여 각각의 키워드 테이블을 구축하고, 이를 모아 도메인 키워드 테이블을 구축하였다. 또한 상품, 속성, 분류 간의 연결관계를 매핑 테이블로 구성하였다.

검색 엔진은 상품 정보 질의어를 기반으로 해당 분류

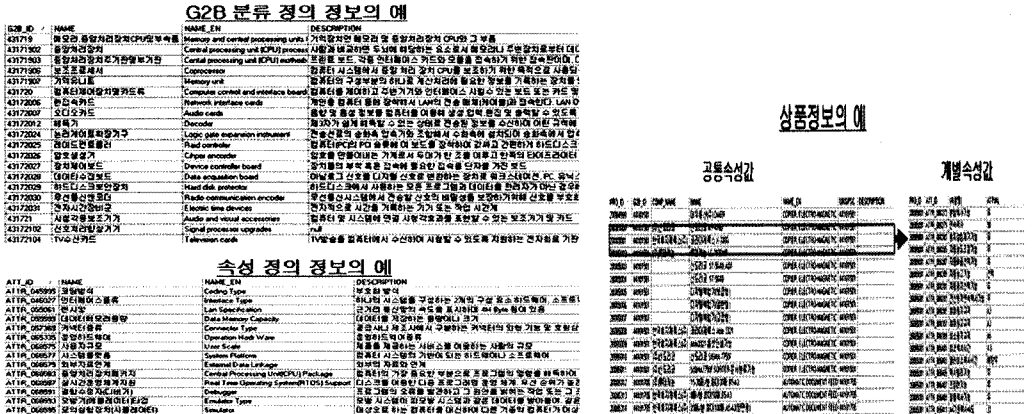


그림 9 실험에 사용된 데이타의 예

를 찾은 것으로 구현하였다. 이를 위해 질의어가 입력되면 어휘확장함수를 통해 분해하였으며, 이렇게 분해된 어휘는 두 가지 방식으로 목표 분류를 찾아 나갔다. 첫 번째 방식은 분류의 정의에 사용된 어휘를 찾아서 분류를 직접적으로 찾는 방식으로 기존의 일반적인 정보검색 시스템과 동일한 방식이다. 두 번째 방식은 본 논문에서 제시한 의미적 연결 관계에 따른 검색으로 상품 및 속성에 사용된 어휘를 찾아 상품 및 속성 정보를 추출한 다음 해당 상품 및 속성과 의미적 연결관계를 지니는 분류를 찾는 방식이다. 검색엔진은 이렇게 추출되는 분류를 앞서 제시한 유사도 평가함수에 따라 서열화하여 사용자에게 제시하도록 구현 되었다.

실험은 유사도 평가 함수에 사용된 각 평가 요소인 가중치를 변경하였을 때 나타나는 서열화된 결과에 대한 정확도를 측정하는 것으로 진행하였으며, 질의어를 구성하는 정보 또한 실험상 비교 기준으로 사용하였다. 각 비교 기준 정보는 표 4와 같다.

Entity 가중치는 분류, 속성, 상품 중 어느 Entity에서 도출 되었는가에 따라, 속성별 가중치는 명칭, 설명 혹은 상품일 경우 제조사, 기타로 구분하여 적용하였으며, 어휘 가중치는 공백 기준 추출인지, 명사사전에 일치하는 명사 추출인지에 따라 구분하여 적용하였다. 그리고 키워드 가중치는 5.2 절에 제시된 바와 같으며 적용유무를 구분하여 실험하였다. 또한 질의어 구성정보는 질의어가 물품명인지 혹은 전체 상품정보를 하나의 문자열로 구성한 것인지에 대한 구분이다. 또한, Entity 가중치를 0으로 설정하여 Related Entity에 대한 접근 경로를 차단함으로써 기존 일반 정보검색 환경을 재현하였다.

실험 평가는 사용자가 원하는 정보가 질의어에 해당하는 상품의 실제 분류라고 가정하고 해당 분류가 검색결과에서 몇 번째로 제시 되는가를 기준으로 하였다. 즉, 첫 번째 순위에 정확한 상품 분류정보를 낸 비율과 5 순위 내에 낸 비율, 10 순위 내에 낸 비율 및 25 순위, 50 순위 내에 낸 비율을 하나의 실험에 대한 성능 평가 척도로 삼았다.

6. 실험 결과 및 분석

6.1 실험 결과

표 5는 각 파라메터의 변화에 따른 실험 목록을, 그림 10은 표 5상의 각 실험 별 결과를 순위별 정확도로 나타낸 것이다.

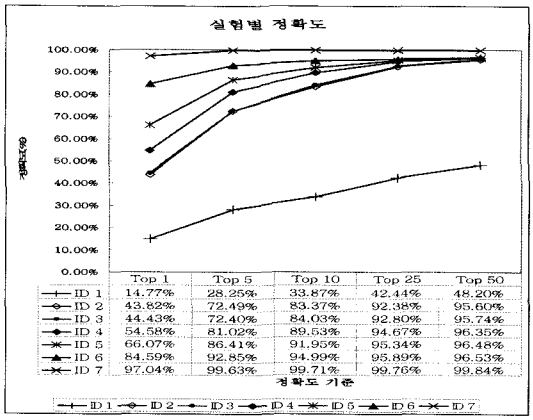
표 4 실험 파라메터

파라메터	파라메터 값
Entity 가중치	분류 점수, 속성 점수, 상품 점수
속성별 가중치	한글 명칭(상품명포함)에 해당하는 속성 점수 (**명칭***) 설명에 해당하는 속성 점수 (**설명***) 상품일 경우: 제조사(상품) 및 기타속성 점수를 추가 구분함
어휘(VocType) 가중치**	공백으로 구분하여 추출한 어휘에 대한 점수 (**일반***) 용어사전에 명시된 명사를 추출하여 확장한 어휘에 대한 점수 (**상세***)
키워드 가중치	적용 유무
질의어 구성정보	상품명(한글)인 경우 (**물품명***) 상품명(한글) 및 기타속성 포함한 전체 정보일 경우 (**전체***)

* 표 5 상의 명칭 ** 어휘 가중치는 질의어 어휘(사용자 입력 키워드)에 대한 것도 포함함

표 5 단계별 실험 목록

ID	비고	파라메터				
		Entity	속성	어휘	키워드가중치	질의어
1	의미적연결관계 미사용	분류=1 그외=0	모두 1	모두 1	사용	물품명
2	의미적 연결관계 적용	모두 1	(상동)	(상동)	미사용	(상동)
3	Entity가중치 적용	분류=10 속성=1 물품=2	(상동)	(상동)	(상동)	(상동)
4	속성별가중치적용	(상동)	명칭=10 제조사=5 기타=1	(상동)	(상동)	(상동)
5	VocType가중치적용	(상동)	(상동)	일반=10 상세=1	(상동)	(상동)
6	키워드가중치적용	(상동)	(상동)	(상동)	사용	(상동)
7	상세 질의어 정보	(상동)	(상동)	(상동)	(상동)	전체



	Top 1	Top 5	Top 10	Top 25	Top 50
ID 1	14.77%	28.25%	33.87%	42.44%	48.20%
ID 2	43.82%	72.49%	83.37%	92.38%	95.60%
ID 3	44.43%	72.40%	84.03%	92.80%	95.74%
ID 4	54.58%	81.02%	89.53%	94.67%	96.35%
ID 5	66.07%	86.41%	91.95%	95.34%	96.48%
ID 6	84.59%	92.85%	94.99%	95.89%	96.53%
ID 7	97.04%	99.63%	99.71%	99.76%	99.84%

그림 10 실험 결과

6.2 분석

의미적 연결 관계 확장과 데이타 모델에 의한 유사도 평가함수의 정확성뿐만 아니라 각 유사도 평가 요소가 유사도 평가에 유효한 영향을 끼치는 지를 판단하기 위하여 단계별로 각 유사도 평가 요소를 추가해 나가는 것으로 실험 결과를 정리하였다. 각 단계별 비교 초점은 다음과 같으며, 이를 통해 본 논문에서 제시하고 있는 데이타 모델 및 유사도 평가 함수의 유효함을 검증할 수 있었다.

첫째, 의미적 연결 관계에 의한 확장의 효과는 ID 1 실험과 ID 2 실험의 결과 차이로 그 효과의 유효성을 확인 할 수 있다. 이는 또한 일반적인 정보 검색 모델을 적용했을 때와 본 논문에서 제시한 데이타 모델(가중치 조정 전)을 적용했을 때의 기본적인 차이를 의미한다.

둘째, 각 요소 별로 적용된 가중치(Entity 가중치)의 효과는 ID 2 실험과 ID 3 실험의 결과 차이에서 확인 할 수 있다. 비록, 1% 이내의 차이를 보여 타 가중치에 비해 영향력이 가장 낮지만, 요소 별로 추출되는 데이타의 출처에 대한 정보이므로 다른 가중치와 완전히 독립적이지는 않다는 측면에서 의미가 있다.

셋째, 각 속성별로 적용된 가중치의 효과는 ID 3 실험과 ID 4 실험의 결과 차이로 확인 할 수 있다. 비록 다양한 속성정보에 대한 세분화된 가중치까지 적용하지는 못하였지만, 분류와 속성의 경우에는 명칭과 설명에 대한 가중치를, 상품에 대해서는 상품명, 제조사 및 그 외의 속성으로 단순 분류하여 적용한 본 실험의 경우에도 그 유효성을 보이고 있다.

넷째, 어휘 확장 함수에 의해 확장된 어휘의 어휘 확장 정보에 따른 어휘 가중치의 효과는 ID 4 실험과 ID 5 실험과의 결과 차이로 알 수 있다. 즉, 어휘 확장을 통해 해당 요소 정보에 대한 접근 경로는 확장 되지만, 그 만큼 정확한 접근에 대한 어려움이 생기는 문제점이 발생하는데, 어휘 가중치가 이를 보완해 주는 역할을 수행함을 보이고 있다.

다섯째, 5장 2절에서 제시한 키워드 가중치의 효과는 ID 5 실험과 ID 6 실험과의 결과가 입증하고 있다. 앞서 제시한 키워드 가중치는 본 모델에 맞추어 단순하게 구성되었지만, 정확도 향상에 많은 영향을 끼치고 있어 그 유효성이 검증됨을 보이고 있으며 이 단계까지 누적되어 검증된 정확도는, 선행 연구 [5]의 성과와 비교할 만하다.

끝으로, 질의어 정보의 정확도 향상에 따른 결과 정확도를 통해 데이타 모델 및 유사도 평가 함수가 가지는 유효성은 ID7 실험을 통해 알 수 있다. ID 7 실험은 ID 6 실험과 동일한 환경에서 입력하는 질의어 정보를 상세하게 확장한 경우이다. 즉, ID1 실험에서 ID6 실험은 질의어를 단순히 상품명(한글명칭)만을 가지고 입력하였지만 ID 7 실험에서는 기존의 전자 카탈로그가 가지고 있던, 상품 명 외의 기타 속성 정보 - 영문 상품명 및 개별 속성 정보 - 또한 질의어에 포함하여 수행한 실험으로써 높은 정확도를 보이고 있다. 이는 질의어가 가지는 정보가 높을수록 검색의 정확성이 높아진다 점과 상품이 가지는 개별 속성이 의미적 연결관계를 이용한 검색에서 유효한 정보로 작용한다 점을 보여 주고 있다. 또한, 이는 제시된 데이타 모델 및 유사도 평가 함수가 정확한 정보 검색을 지원하고 있음을 간접적으로 검증할 수 있는 결과이기도 하다.

7. 결 론

본 논문은 대용량 전자 카탈로그의 효율적인 검색 시스템 구축을 위한 기초 연구로써 의미적 연결관계에 기반한 전자 카탈로그 검색 시스템을 제시를 통해 효율적인 검색 시스템을 고안하였다. 해당 시스템을 구현하기 위해 필수적인 전자 카탈로그 데이타 모델 및 유사도 평가 함수를 제시하였으며 실험을 통해 그 유효성을 검증하였다. 본 논문에서 제안하고 있는 바는 다음과 같다.

첫째, 의미적 연결관계에 기반한 추상적인 전자 카탈로그 데이타 모델을 제시하였다.

둘째, 전자 카탈로그 데이타가 가지는 어휘적 특성을 반영하여 확장된 전자 카탈로그 데이타 모델을 제시하였다.

셋째, 제시된 데이타 모델에 의거하여 유사도 평가에 영향을 끼치는 요소들을 정리하였다.

넷째, 도출된 유사도 평가 요소를 중심으로 유사도 평가함수를 제안 하고 제시된 데이타 모델을 이용한 구체적인 구현 방안을 제안 하였으며, 실험을 통해 이를 검증하였다.

또한, 제시된 데이타 모델에 대한 적합한 인덱스 구조

및 다양한 키워드 가중치의 적용, 벡터 기반 모델 외 타 모델에 대한 적용, 그리고 타 분야, 특히 온톨로지 환경에서의 검색에 대한 적용 등의 향후 연구과제가 남아 있다 하겠다.

참 고 문 헌

[1] Arthur Keller, Michael Genesereth, N. Singh, and M. Syed, "A smart catalog and brokering architecture for electronic commerce," Workshop on Electronic Commerce, 1994.

[2] Sherif Danish, "Building Database-driven Electronic Catalogs," SIGMOD Record, Vol.27, No.4, December, 1998.

[3] Ewa Callahan and Jurgen Koenemann, "A Comparative Usability Evaluation of User Interfaces for Online Product Catalogs," EC'00, October 17-20, 2000.

[4] 정지혜, 이상구, 우치수, "전자 상거래에서의 체계적인 상품 카탈로그 구축을 위한 분류체계 모델 및 구현," 한국 데이타베이스 학술대회 논문집, 15(1), pp. 343-349, 1999.

[5] 서광훈, 이경종, 김현철, 이태희, 이상구, "Naive-Bayesian Classifier를 이용한 전자 카탈로그 자동 분류 시스템", 한국정보과학회 제 31회 춘계학술발표회, pp. 91-93, 2004. 4.

[6] 김기륭, "전자 카탈로그 자동 분류기에 대한 연구", 서울대학교 전기·컴퓨터공학부, 2003.

[7] Y. Ding, M. Korotkiy, B. Omelayenko, B. Kartseva, V. Zykov, M. Klein, E. Schulten, and D. Fensel, "GoldenBullet: Automated Classification of Product Data in E-commerce," Withold Abramowicz (ed.), Business Information Systems, Proceedings of BIS 2002, Poznan, Poland, 2002.

[8] Domenico Beneventano and Stefania Magnani, "framework for the classification and the reclassification of electronic catalogs," SAC' 04, March 14-17, 2004, Nicosia, Cyprus.

[9] 김재범, 김동규, 이상구, "전자상거래 환경에서의 분류 체계 자동 통합 기법", 제27회 정보과학회 추계 학술대회 논문집, 2000.

[10] 김동규, "전자 카탈로그의 의미기반 모델 연구", 서울대학교 전기·컴퓨터공학부, 2004.

[11] Joerg Leukel, "Standardization of Product Ontologies in B2B Relationships - On the Role of ISO 13584," On Proc. of the Tenth Americas Conference on Information Systems, New York, New York, 2004.

[12] Peter Eklund, Richard Cole, and Nataliya Roberts, "Retrieving and Exploring Ontology-based sInformation," Handbook on Ontologies in Information Systems, Springer, 2003.12.

[13] Kemafor Anyanwu and Amit Sheth, "ρ-Queries: Enabling Querying for Semantic Associations on the Semantic Web," WWW2003, Budapest, Hungary, 2003.5.

서 광 훈

2002년 서울대학교 지리학(학사). 2005년 서울대학교 전기·컴퓨터공학(석사). 2001년~2002년 ㈜아이모바일테크놀러지 개발팀장. 2005년~2007년 ㈜KT 전임연구원. 2007년~현재 ㈜넴스 이사. 관심분야는 전자상거래 기술, 시멘틱 기술, 데이타베이스

이 상 구

1985년 서울대학교 계산통계학(학사). 1987년 Northwestern University 전산과학(석사). 1990년 Northwestern University 전산과학(박사). 1990년~1992년 EDS Research & Development 선임연구원 1999년~2000년 Georgetown University 객원교수. 1992년~현재 서울대학교 컴퓨터공학부 교수. 2002년~현재 서울대 e-비즈니스연구센터 센터장. 관심분야는 데이타베이스, 전자상거래 기술, 시멘틱 기술