의미적 연결 관계에 기반한 전자 카탈로그에서의 확장된 어휘 인덱스 구축 및 이를 이용한 검색 성능 향상 기법

이동주°, 이태희, 이상구
서울대학교 지능형 데이터 베이스 시스템 연구실
{therocks°, thlee, sglee}@europa.snu.ac.kr

Construction of Keyword Index and Improved Search Methods for e-Catalogs Based on Semantic Relationship

Dongjoo Lee°, Taehee Lee, Sang-goo Lee
Intelligent Database System Lab. Of SNU

요 약

본 논문에서는 기 구축된 전자 카탈로그를 의미적 연결 관계에 기초한 확장된 전자 카탈로그로 변환하는 방법을 제안한다. 이를 통해 구축된 확장된 전자 카탈로그에서 의미적 태깅에 의한 확장된 어휘 인덱스 구축 방안과, 이를 이용한 검색 성능 향상 기법을 제안한다. 기존의 전자 카탈로그는 상품 정보가 분류별로 생성된 테이블에 저장되고 저장된 테이블로부터 생성된 키워드 인덱스로부터 검색이 이루어 졌다. 이러한 검색은 상품이 가지는 정보를 데이터베이스에 구축된 테이블에만 한정하게 되어, 전자 카탈로그에 포함된 상품이나 분류간의 의미적 연결 관계들을 충분히 이용하지 못하였다. 전자 카탈로그에 내재된 의미적 요소를 충분히 활용하기 위해서는 전자 카탈로그를 의미적 연결 관계에 기초한 모델로 구성할 필요가 있다. 본 논문에서는 의미적 모델 기반 전자 카탈로그 시스템으로의 전환 과정을 XML형태의 명세를 이용해 반자동적으로 전환할 수 있는 툴을 구현하며, 단순 키워드 어휘 인덱스 구축이 아닌, 어휘 인덱스의 의미적 확장을 제안하고, 이를 위한 태그 요소로써 어휘에 대한 형태소 분석 결과, 수치 환산 및 확장 요소, 속성 값의 도메인 정보 등을 제시하였다. 이를 기반으로 최적의 검색 결과를 얻어 내도록 하는 인접도 평가 함수에 적용하는 방법을 제시한다.

1. 서 론

전자 카탈로그는 e-Business에서 상품 정보 공유를 위해 매우 중요한 요소이며, 전자 카탈로그의 구축 및 이를 이용한 정보 검색은 전자 카탈로그 이용에서 핵심이 되는 부분 중 하나이다. 전자 카탈로그에는 분류를 기반으로 각 상품의 속성 값에 대한 정보가 담겨 있으며, 이를 데이터베이스에 효율적으로 구축하는 기법은 이전부터 매우 다양하게 연구되어 왔다[1]. 전자 카탈로그 데이터베이스로부터 간단한 키워드를 이용한 상품 검색을 위한 방안 또한 다양하게 연구되어 왔으며, 최근에는 전자 카탈로그를 의미적 연결 관계에 기초하여 구축하기 위한 모델이 제시되었으며[2], 이를 통한 상품 검색 기법에 관한 연구도 진행되고 있다[3].

상품은 계속적으로 증가하고 있으며, 이를 전자 카탈로그로 구축하는 작업 또한 계속적으로 수행되어 전자 카탈로그로 구축된 상품의 수는 수십만 건 이상으로 많아지고 있다[4]. 현실에는 이보다 더 많은 상품이 존재 하고 있으며, 계속적으로 증가할 것이다.

이 같은 대량의 전자 카탈로그에서 원하는 정보를 효율적으로 검색하기 위해서는 단순히 데이터베이스에 질의를 하는 것만으로는 불가능하다. 따라서 키워드나 문장 등의 검색어를 이용한 상품 검색 시스템이 필요하며, 검색 어휘에 대한 인접도 평가함수[3]를 이용한 검색 결과의 순위화가 필요하다. [3]에서는 검색 어휘를 확장하고 의미적 연결 관계를 이용한 검색 성능 향상을 보였지만, 검색 어휘 추출 및 확장이 사전을 기반으로 하여 불필요한 값들이 많이 추출 되는 단점이 있었다. 본 논문에서는 어휘 인덱스 구축을 위해 형태소 분석을 적용하고, 전자 카탈로그로 구축된 정보를 최적으로 이용하기 위해, 전자 카탈로그를 의미적 연결 관계에 기반한 모델을 이용해 구축한다. 이 같은 과정을 XML명세로부터 반자동으로 변환하기 위한 방법을 제시하고, 이를 구현한다. 검색 어휘에 대해, 엔터티의 속성별로 다른 인덱스 추출 기법을 이용하여 어휘 인덱스를 구축하고, 이를 통해 검색 어휘의 상품에 대한 인접도를 구할 수 있는 함수를 기존의 벡터 기반의 함수[3]로부터 확장하여 제시한다.

본 논문은 다음과 같이 구성된다. 2. 관련 연구는 전자 카탈로그 구축 기법과 인접도 평가 함수에 관한 연구를 알아보고, 3. 확장된 전자 카탈로그 구축에서는 확장된 전자 카탈로그 구축 모델과 이를 위한 자동 변환 시스템에 대해서 논하고, 4. 인접도 평가 함수 에서는 인접도 평가 함수의 요소와 함수의 구성에 대해서 알아본다. 끝으로 5. 결론 및 향후 과제에서 본 연구에 대한 결론을 맺고 본 연구에서 수행하지 못한 실험 및 평가에 대한 향후 과제를 제시하고 끝맺는다.

2. 관련 연구

e-Business가 활성화되기 시작함에 따라서 전자 카탈로그에 관한 연구 역시 다양하게 수행되었다. 전자 카탈로그에 대한 연구는 주로 전자 카탈로그를 데이터베이스에 구축하는 방안에 관한 연구, 상품의 분류를 위한 분류 체계 모델 에 관한 연구 등이 많이 수행되었고, 최근에는 전자 카탈로그에 구축된 정보를 효율적으로 검색하기 위한 기법에 대한 연구가 수행기도 하였다. [1]에서는 전자 카탈로그의 기능과 데이터베이스 구축 기법에 대한 소개를 하고 있다. 상품 분류 체계는 UNSPSC[5], eCl@ss[6]와 같이 코드를 기반으로 한 상하 관계를 표현하는 분류 체계가 일반적이데, 이에는 상품 정보를 표현하는데 있어 많은 한계가 있다. [2]에서는 이러한 한계를 지적하고, 상품의 의미적 관계 및 분류 체계와의 관계까지도 표현할 수 있는 의미적 분류 모델을 제시한다. 또한, 다양한 사용자의 각기 다른 관점을 만족시키기 위한 분류 체계의 의미 표현 기법에 대해서

본 연구는 정보 통신부의 대학 IT 연구센터 ITRC (Information Technology Research Center)의 지원을 받아 수행되었음

논하며, 이를 위해서는 상품의 저장 구조를 하나의 분류 체계에 국한 시켜서 구축할 수 없음을 논하고 있다. [7]에서는 현실적으로 가장 많이 사용되고 있는 관계형 데이터베이스에서의 전자 카탈로그 구축을 위한 스키마 디자인 및 각 기법들의 성능 평가를 보여주고 있다.

최근에는 [8]에서 전자 카탈로그를 온톨로지적 관점에서 접근하여 상품 온톨로지 시스템 구축에 있어서의 현실적 문제들에 대해서 논한다. [3]은 의미적 연결 관계에서 검색 결과를 효율적으로 표현하기 위한 순위 결정 알고리즘에 대해서 연구하였고, 이에 대한 실험 결과를 보여준다.

3. 확장된 전자 카탈로그 구축

3.1. 의미적 연결 관계 기반 전자 카탈로그

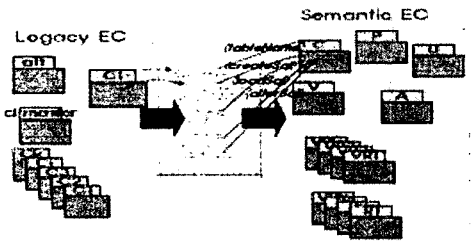
의미적 연결 관계를 기반으로 한 확장된 전자 카탈로그 모델에서는 전자 카탈로그의 구성 요소를 엔터티와 엔터티들 간의 관계로서 정의한다. 엔터티는 상품(Product), 분류 스킴(Classification Scheme), 속성(Attribute), 단위(UOM:Unit Of Measure)로 정의 되며, 관계는 엔터티들 간의 관계로 정의된다. [8]에서는 관계를 일반적 관계와 상품 온톨로지에서 나타나는 특화된 관계로 구분하여 제시하고 있다. 이와 같은 의미적 분류 모델을 이용하여 전자 카탈로그를 다음과 같이 일반화 할 수 있다.

$EC = \{E, R\}$, $E = \{P, C, A, U\}$
$ME \in \{C, A, U\}$, $MA = \{\alpha_1, \alpha_2, ..., \alpha_m\}$
$P = \{ (a, v) \mid a \in A, v \in VALUE\}$
$ME = \{(\alpha, v) \mid \alpha \in MA, v \in VALUE\}$
$R = \{ (e_1, e_2, r) \mid e_1 \in E_1, e_2 \in E_2, E_1 \in E, E_2 \in E, r \in DR\}$
EC : 전자 카탈로그, E : 엔터티, R : 관계, DR : 관계 정의
ME : 메타 엔터티, MA : 메타 속성
P : 상품, C : 분류 체계, A : 속성, U : 단위,

상품 정보는 메타 정보를 통해 표현되고, *C, A, U*는 이러한 메타 정보라 할 수 있다. 위에서 정의된 엔터티와 관계는 기 구축된 전자 카탈로그로부터 추출되어 의미적 분류 모델로 확장 구축될 수 있다.

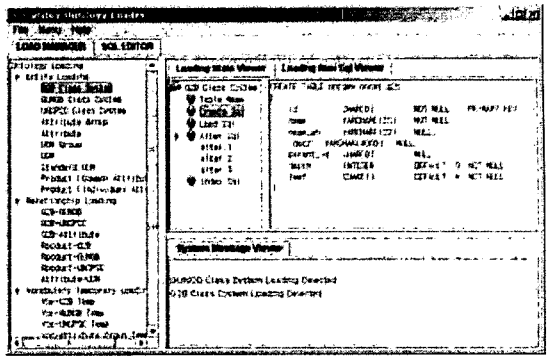
3.2. XML명세를 통한 반자동적 추출

확장된 전자 카탈로그 구축은 기 구축 전자 카탈로그로부터 ① 엔터티와 관계를 추출하는 과정, ② 어휘 인덱스를 구축하는 과정으로 나누어진다. 엔터티와 관계를 추출하기 위해서 기 구축 전자 카탈로그의 데이터베이스로부터 확장 모델로 추출하는 과정을 XML로 명세하고 명세된 XML 파일을 읽어 들여 엔터티와 관계를 추출하는 반자동화된 과정을 통해서 수행될 수 있다. 어휘 인덱스 또한, XML을 이용하여 어휘 추출 대상 엔터티와 추출 대상 속성을 명세하여 자동적으로 추출 할 수 있다. 이와 같은 방법은 추출 과정에서 겪게 되는 다양한 오류를 XML 명세를 수정하는 것으로 간단히 해결할 수 있게 하고, 추출 과정의 변경에 관해 용이하게 확장할 수 있게 한다.



(그림 1) XML명세를 통한 추출

(그림1)에서 <tableName>, <createSql>, <loadSql>, <alterSql> 등의 태그를 이용해서 추출하는 과정을 도식화 하여 볼 수 있고, 이 같은 과정은 GUI를 제공하는 툴로 (그림2)와 같이 구현된다.



(그림 2) 의미적 분류 모델에 기반한 전자 카탈로그 추출 도구

3.3. 어휘 인덱스 구축

의미적 분류 모델을 통한 엔터티의 검색을 위해서는 각 엔터티와 검색어를 연결하는 어휘 인덱스 구축이 중요하다. 어휘 인덱스 구축을 위해서는 각 엔터티의 속성 값으로부터 적절한 어휘를 추출하여 이를 확장한 형태로 저장할 필요가 있다.

일반적인 문서의 어휘 인덱스 구축과는 달리 전자 카탈로그에서의 어휘 인덱스 구축은 의미적 확장을 지원하기 위해 구축 대상 엔터티와 구축 대상 속성(또는 메타 속성)에 따라서 다른 방법으로 수행된다. 분류의 경우 주요한 추출 대상은 분류 명과 분류 설명이 된다. 분류 명의 경우 분류를 나타내기 위한 적절한 어휘로 표현되며, 이 같은 경우 분류를 나타내기 위한 명사가 주요 추출 대상이 된다. 분류 설명의 경우 해당 분류에 대해 사람이 이해할 수 있는 형식으로 분류에 대해 서술되어 있는데, 이 같은 경우 서술된 어휘의 형태소 분석[9]을 기초로 한 각 어휘의 정제 및 확장을 통한 어휘 인덱스 구축이 필요하다. 그러나 형태소 분석을 이용한 어휘 인덱스 추출은 형태소 분석 결과가 얼마나 정확 하느냐에 따라서 어휘 인덱스의 품질이 결정되는데, 종종 관련 없는 어휘가 추출되므로 이를 보완할 필요가 있다. 본 어휘 추출 모듈에서는 형태소 분석 후에 형태소 분석 후보들에 대한 경험적 선별과 명사 사전을 통한 어휘 확장을 통해 전자 카탈로그를 위한 적합한 어휘 인덱스를 구축하도록 하였다. 형태소 분석에는 [9]에서 배포한 한글 형태소 분석 모듈(HAM)을 이용하였다. 형태소 분석 외에도 크기나 길이와 같이 단위를 가지는 수치를 값으로 하는 속성 값의 경우 수치에 대한 단위 및 크기의 변환 및 확장을 위해 표준화를 통한 어휘 인덱스 추출을 수행하도록 하였다. (표1)은 메타 속성 및 속성에 따른 어휘 인덱스 추출 기법에 대한 예를 보여준다.

(표 1) 메타 속성 및 속성별 어휘 인덱스 추출 기법

인덱스 추출 기법	해당 메타 속성 및 속성
형태소 분석 확장	자연어 형태의 서술인 메타 속성 및 속성
명사 추출	분류명과 같이 대상에 대한 명칭인 메타 속성 및 속성
수치 확장 및 표준화	크기, 길이와 같이 수치가 단위에 따라서 변하거나 확장될 수 있는 속성
단순 추출	분류 코드와 같이 원래 값만을 대상으로 하는 속성

3.4. 어휘 인덱스의 의미적 태깅

추출된 어휘의 의미적 활용도를 높이기 위해서는 추출된 어휘의 의미를 나타내는 태그(Tag)를 부여할 필요가 있다. 태그에는 추출된 속성, 어휘의 품사, 어휘에 추가된 조사, 도메인(Domain), 어휘 확장 유형, 어휘의 출현 빈도 등과 같은 것들이 있다. 태깅 값은 검색 시에 검색어와 엔터티와의 인접도를 구하기 위한 요소로서 적용되어 검색 성능을 향상시키는 요소로 작용할 수 있다. 그러나 이러한 요소는 검색에 대한 과부하를 초래할 수 있기 때문에, 이를 효율적으로 제한할 필요가 있다.

(표2)는 실험을 위하여 작성된 어휘 인덱스 추출 모듈을 이용한 분류 명과 분류 설명에 대한 인덱스 추출의 예이다. 추출된 인덱스에 대해서, 품사, 조사, 어미, 확장 유형, 빈도 등이 태깅된다.

(표 2) 어휘 인덱스 추출 예

값	확장된 어휘 인덱스 추출 값
농약보조제	(농약,N,N,N,,,3,2,1,1,1) (보조제,N,N,N,,,3,2,1,1,1) (보조,N,N,N,,,3,2,1,1,1) (농약보조제,N,N,N,,,3,1,1,1,1)
인류가 야생동물을 순치 개량한 것	(인류,NJ,N,N,OI,,3,1,1,5,1) (야생,N,N,N,을,,3,2,2,5,1) (동물,N,N,N,을,,3,2,2,5,1) (야생동물,NJ,N,C,을,,3,0,2,5,1) (순치,N,N,Z,,3,1,,3,5,1) (개량,N,N,Z,,3,,1,4,5,1) (개량하,VM,V,Z,,은,3,1,4,5,1)

3.5. 어휘 인덱스 저장 스키마 구조

엔터티로부터 추출된 어휘 인덱스는 관계형 데이터베이스에 저장된다. 이를 위해서 추출된 어휘의 값을 어휘 테이블에 저장하고, 각 엔터티에서 추출된 어휘는 어휘와 엔터티와의 관계로서 표현한다.

$VR_E = \{(e\_id, att, vc\_id, tag_1, tag_2, \ldots tag_t)\}$
VR_E : 엔터티-어휘간 관계
eid : 엔터티 식별 아이디
att : 속성명
vid : 어휘 식별 아이디
tag_i : 의미 확장 태그

이 같은 저장 구조는 어휘 인덱스에 대한 검색 질의를 어휘 테이블(Voc)에 국한하고, 이를 통해 다양한 엔터티에 대한 질의 확장 및 의미적 관계의 이용을 가능하게 한다.

4. 인접도 평가 함수

전자 카탈로그 검색용 인접도 평가 함수로는 [3]에서 벡터 모델을 기초로 한 함수가 제시되었다. 이를 확장된 어휘인덱스 구조에 적용시킬 수 있도록 확장할 수 있다.

전자 카탈로그에서 검색어 Q에 대한 대상 엔터티 e의 인접도는 다음과 같이 표현한다.

$Score(Q, e)$ ---------- (1)

Q는 인덱스 구축을 위해 사용된 어휘 인덱스 추출과 유사한 기법으로 태깅된 어휘 집합으로 확장될 수 있고,

$Q = \{q_1, q_2, \ldots, q_n\}$ ---------- (2)

$q_i = (voc, tag_1, tag_2, \ldots, tag_s)$

검색 대상 e는 전자 카탈로그를 구성하는 엔터티인데, 다음과 같이 e를 이루고 있는 속성(또는 메타 속성) a(또는 α)와 값 v의 쌍의 집합으로 표현된다.

$e = \{(a, v) \mid a \in ATT, v \in VALUE\}$ ---------- (3)

(a, v)는 어휘 인덱스 추출 과정을 통해서 태깅된 어휘 인덱스 $ivoc = (voc, att, tag1, \ldots, tag_t)$으로 치환되어 e는 다음과 같이 확장된 어휘 인덱스 집합으로 표현된다.

$e = \{ivoc_1, ivoc_2, \ldots, ivoc_i\}$ ---------- (4)

(1) 을 (2)와 (4)를 이용하여 표현하면 인접도 평가 함수는 다음과 같이 확장된다.

$Score(Q, e) = \sum_{i,j} Score(q_i, ivoc_j)$ ---------- (5)

e는 관계 R을 통해서 e'과 의미적 관계를 가질 수 있다. e'으로부터 얻어진 $Score(Q, e')$를 관계가 가지는 가중치를 적용하여 다음과 같은 일반식을 얻을 수 있다.

$Score(Q, e) = \sum_{i,k} Score(qi, ivoc_i) + \sum_{r,k} w_r * Score(Q, e'_k)$

w_r : 관계 r의 가중치

r : e가 맺고 있는 관계

e'_k : e와 r을 통해 관계된 엔터티

$Score(Q, e')$은 $Score(Q, e)$와 마찬가지로 확장될 수 있고, 이렇게 생성된 결과를 $Score(Q, e)$에 따라 역정렬 하여 결과를 얻을 수 있다. $Score(q_i, ivoc_k)$는 q와 v에 태깅된 태그의 활용을 어떻게 하느냐에 따라서 다양하게 계산될 수 있다. 품사의 일치성, TF/IDF를 이용한 어휘 가중치, 속성, 어휘 확장 유형 등에 따라서 다양하게 변하는데 이에 대한 구체적인 구현과 실험은 향후 과제로 남겨둔다.

5. 결론 및 향후 과제

본 논문에서는 의미적 분류 모델에 기초한 확장된 전자 카탈로그 시스템을 구축하기 위해서 기존의 전자 카탈로그 시스템으로부터 XML형식의 명세를 이용해 반자동적으로 추출하는 도구를 구현하였다. [3]에서 제시한 전자 카탈로그 시스템에서의 정보 검색을 위한 알고리즘의 성능 향상을 위해, 형태소 분석 및 속성별 어휘 인덱스 구축을 제안하였다.

본 시스템은 기존 전자 카탈로그 시스템으로부터 엔터티와 관계의 추출, 어휘 인덱스의 추출 및 저장까지 구현 되었다. 검색 알고리즘은 아직 구현되지 않았고, 위에서 제시한 일반화된 모델에 Score(q, v)를 다양한 방법으로 적용하여 최적의 Score(q, v)를 결정하는 인자들을 얻는 실험을 향후 과제로 수행할 것이다. 엔터티, 관계 모델은 관계형 데이터베이스에서 일반화된 모델이기 때문에 전자 카탈로그뿐만 아니라, 이와 유사한 특징을 보이는 다른 데이터베이스로 확장될 수 있다. 향후 전자 카탈로그 이외의 도메인에 이를 적용하여 검색 성능 향상에 대한 실험을 수행할 것이다.

감사의 말

본 연구를 위해 한국어 형태소 분석 모듈의 사용을 허락해 주신 국민대학교 컴퓨터 공학부 강승식 교수님께 감사의 말씀을 드립니다.

참고 문헌

[1] Arie Segev, Dadong Wan and Caroline Beam, "Electronic catalogs: a technology overview and survey results" , Proceedings of the 4th International conference on information and knowledge management Baltimore, Maryland, USA, Nov 29-Dec 2, 1995, pp. 11-18.

[2] Dongkyu Kim, Sang-goo Lee, Jonghoon Chun, "A Semantic Classification Model for e-Catalogs", IEEE 6th Conference on E-Commerce Technology (CEC 2004), 2004

[3] 서광훈, "의미적 연결 관계에 기반한 전자 카탈로그 검색용 인접도 평가 함수에 관한 연구" , 서울대학교 전기.컴퓨터 공학부, 2005

[4] http://www.g2b.go.kr

[5] UNDP, "United Nations Standard Products and Service Code, White papr," available at http://www.unspsc.org/, 2001

[6] Cologne Institute for Business Research, "eCl@ss - New Standardized Material and Service Classification," available at http://www.eclass-online.com/, 2004

[7] Kiryoong Kim, Dongkyu Kim, Jeuk Kim, Sangwook Park, Ighoon Lee, Sang-goo Lee, Jong-hoon Chun, "An Evaluation of Dynamic Electronic Catalog Models in Relational Database Systems," Managing E-Commerce and Mobile Computing Technologies, IRM press, 73-90, 2003

[8] Ig-hoon Lee, Suekyung Lee, Taehee Lee, Sang-goo Lee, Dongkyu Kim, Jonghoon Chun, Hyunja Lee, Junho Shim, "Practical Issues for Building a Product Ontology System", DEEC 2005

[9] 강승식, "한국어 형태소 분석과 정보 검색" , 홍릉 과학 출판사, 2002, pp.302-332

[10] Andrey Balmin, "ObjectRank: Authority-Based Keyword Search in Databases" , Proceedings of the 30th VLDB Conference, 2004