온톨로지 기반의 자연어 검색 시스템 설계 및 구현

강래구* · 임동일* · 정채영**

*,**조선대학교 전산통계학과

Design and Implementation of Ontology-Based Natural Language Search System

Rae-Goo Kang* · Dong-Il Lim* · Chai-Yeoung Jung**

*,**Dept of Computer Science & Statistics. Chosun University

E-mail : kangrg@hanmail.net

본 연구는 문화관광부 및 한국문화콘텐츠진흥원의 문화콘텐츠기술연구소(CT)육성사업의 연구결과로 수행되었음.

요 약

지금까지의 상품 검색 방법으로는 찾고자하는 정보를 검색할 때 주로 단어의 빈도수나 어휘 정보를 이용하는 키워드 기반의 검색이 주로 쓰이고 있었다. 키워드 기반의 검색에서는 사용자의 질의와 관련이 없는 문서들까지도 같은 결과로 나타내 주고 이로 인해 사용자는 제시된 결과를 한번 더 수동적으로 검색해야하는 부담을 않게 되었다. 이러한 문제점을 해결하기 위해 온톨로지가 대두되었다. 본 논문에서는 온톨로지를 이용한 상품 검색 시스템을 직접 구축하여 분류별 검색을 통해 얼마나 정확한 검색을 하는지 실험하였다. 실험을 위해 전국적으로 On/Off라인 할인점을 운영 중에 있는 A할인점의 상품 데이터 약 40,000여개를 데이터베이스로 구축하였고 User Interface 개발환경은 JSP와 PowerBuilder9.0을 사용하여 검색 시스템을 개발하여 실험하였다. 그 결과 본 논문에서 제안하고 설계한 상품 도메인 온톨로지를 이용한 검색 방법이 기존의 키워드 기반의 검색 방법보다 우수한 결과를 나타내고 있음을 입증하였다.

ABSTRACT

Up until now, when a user search product information, the keyword-based search that mainly uses frequency of words or vocabulary information has been utilized in large. In the keyword-based research, the user should have to bear additional burden in order to search the displayed results manually once again because it shows those files that have no connection at all with the inquiries made by the user. To resolve such a problem, ontology has been emerged. In this paper, product search system using ontology was constructed directly and also tested how accurate search it does perform through the searching according to classification. To test this, about 40,000 product data of A discount store, which was operating on/off line discount stores, were constructed as database, and developmental environment for User Interface was tested by having developed the search system using JSP and PowerBuilder 9.0. Results from the test proved that the search method using Domain Ontology for product presented and designed in this paper was superior to the existing keyword-based search method.

키워드

Ontology, 지능형 검색, 분류별 검색, 상품 도메인

Ⅰ. 서 론

현재의 검색 사이트에서는 데이터베이스를 이용하여 자료를 저장하고 스키마 구조를 통해 자료를 찾아 사용자들에게 단순하게 제공하고 있다.

**교신저자 · cyjung@chosun.ac.kr

자연어로 된 검색문장이 입력되었을 때, 주로 단어의 사용 빈도수나 어휘정보를 이용하여 문서의 유사도를 측정하고 순위를 부여하기 때문에 자연어로 입력된 검색 문장의 정보와 관계가 없는 동형이의어나 동음이의어와 같은 단어들까지

도 동일한 결과로 나타내게 되고 이로 인해 사용자는 제시된 결과를 한번 더 검색을 해야 하는 부담을 안게 되었다. 이러한 문제점을 해결하기 위해서 1999년 W3C 에서는 시맨틱 웹(Semantic Web)을 제안하게 되었고 그 중심에 온톨로지(Ontology)가 있다.[1][2]

본 논문에서는 자연어로 입력된 검색문장을 불용어 리스트를 통해 재구성하고, 정확한 상품 검색을 위해 상품 도메인 온톨로지를 이용하여 사용자가 입력한 질의를 의미적인 해석을 통해 원하는 결과를 보다 쉽고 정확하게 검색할 수 있도록 하였다.

온톨로지는 넓은 의미로는 데이터베이스라고 할 수 있지만 데이터베이스보다 복잡한 형태의 지식과 관련되어 있다는 의미에서 지식베이스라고 부르기도 한다. 지식베이스 구축을 위한 하나의 방안으로 활발히 연구되고 있는 온톨로지는 정보검색은 물론 지식을 수집하고 표현할 수 있는 용어의 집합이라고 정의하고 있다.[1][3]

온톨로지는 특정분야의 용어 그리고 용어의 관계정의는 물론 용어의 조합규칙과 용어의 확장에 대한 관계도 정의한다. 즉, 온톨로지는 영역 지식 표현을 위한 형식 체계인 것이다.[4][5]

II. 온톨로지 시스템 설계

대부분의 인터넷 쇼핑몰은 상품 검색시 주로 키워드를 기반으로 검색을 함으로써 사용자가 입력한 검색어가 데이터베이스의 상품 목록과 일치해야만 정확한 검색이 이루어지며 또한, 유사한 상품을 검색함에 있어서 키워드가 서로 다를 경우엔 검색할 수 없는 단점이 있었다. 이는 기존 웹에서 주로 사용되고 있는 HTML이 의미 정보를 제공하지 못하기 때문이다.[6]

2.1 온톨로지의 개념 설계

개념 설계는 구축하고자 하는 온톨로지에 필요한 클래스를 결정하는 것으로 시작한다. 하지만 클래스를 결정하고 설계한다는 것은 매우 어렵고 불명확하다.

표 1. 온톨로지 클래스 사전의 일부

클래스	상위 클래스	하위 클래스
유음료	Daily식품	우유, 두유, 유산균음료
디저트	Daily식품	호상, 젤리, 푸딩, 무스
유제품	Daily식품	버터, 마아가린, 치즈
탄산음료	음료	콜라, 사이다, 기타 탄산음료
과즙음료	음료	천연과즙, 기타 과즙음료
국산주류	주류	소주, 맥주, 전통주, 탁주
수입주류	주류	위스키, 브랜디, 와인

이 과정은 온톨로지 개발자와 관련 분야 전문가의 경험과 직관에 강하게 의존하게 된다.[7] 표 1은 본 논문에서 구축한 온톨로지 클래스 사전의 일부를 나타내고 있으며 클래스는 다음과 같이 세 단계 트리구조로 분류할 수 있다.

2.2 온톨로지의 구현

온톨로지를 구축함에 있어서 최상위 클래스는 부모를 갖고 있지 않기 때문에 제일 상위에 위치하게 된다. 핵심 클래스란 부모를 가지고 있는 클래스로서 각 분류의 조상 클래스가 된다. 보편적으로 온톨로지를 구축함에 있어서 핵심 클래스는 2단계 혹은 3단계 수준에 위치하게 된다. 마지막으로 행위 클래스는 트리구조에서 가장 하위에 위치해 있고 자손으로 비유된다. 최상위 클래스와 핵심 클래스를 제외한 모든 것을 행위 클래스로 간주한다.[8]

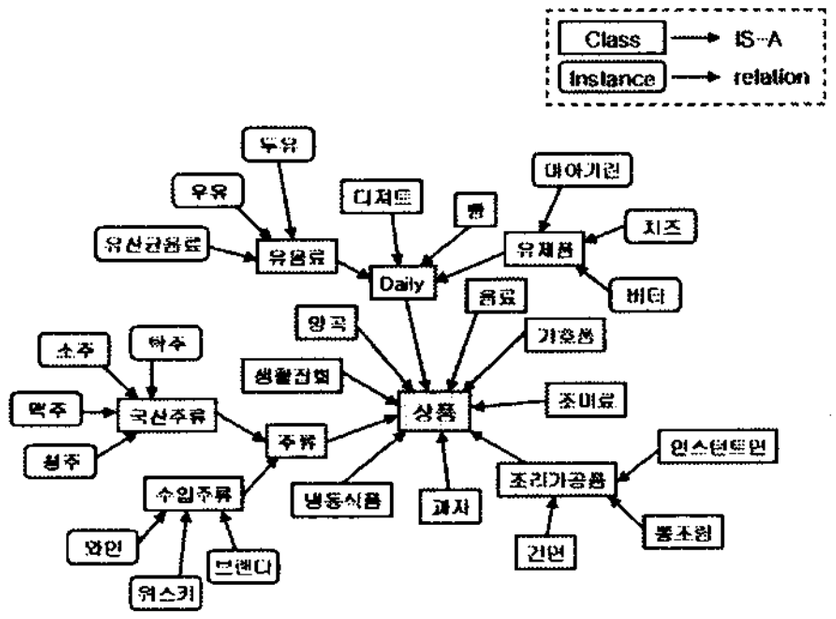

그림 1. 온톨로지 클래스의 개념과 관계 일부

그림 1은 본 논문에서 구축한 상품 도메인 온톨로지 클래스의 개념과 관계를 축약하여 보여주고 있다.

이와 같은 개념과 관계를 이용하여 상품 도메인 온톨로지를 구축하였다.

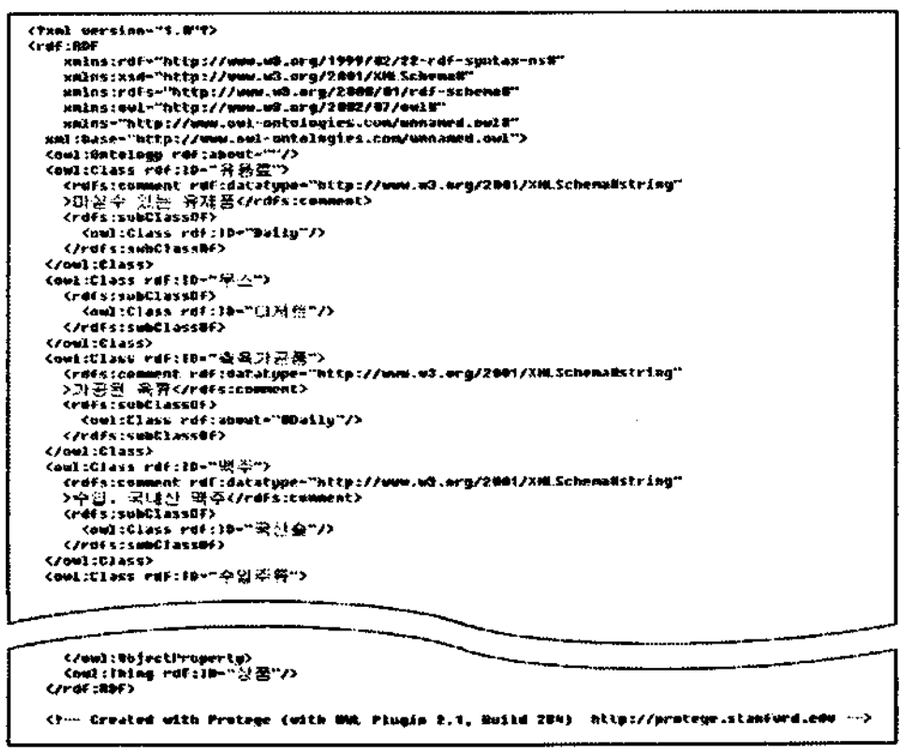

그림 2. OWL로 표현한 상품 도메인 온톨로지

온톨로지에 존재하는 개념과 그들의 관계는 Protégé 3.1을 이용하여 그림 2와 같이 OWL로 표현하였다. 여기서 한글로 표현된 클래스명과 속성 등은 이해를 돕기 위해 재구성한 것으로서 실제 표현함에 있어서는 모두 영문으로 작성하였다.

2.3 자연어 처리기능

사용자가 원하는 상품을 정확히 검색하기 위해서는 무엇보다 사용자가 입력한 검색어를 명확하게 가공해야 할 필요가 있다.

이러한 자연어 위주의 검색어에서 불용어를 제거함으로써 검색어를 보다 명확하게 하여 정확한 검색이 가능하도록 해야 한다.

표 2. 불용어 리스트 일부

불용어 리스트		
~은	~것	~만든
~는	~들	~짜리
~이	~의	~에서
~가	~에	~생산
~중	~인	~에는

표 2는 검색어에서 불필요한 단어를 제거하기 위해 본 논문에서 사용한 불용어 리스트의 일부를 보여주고 있다. 불용어 리스트는 자주 사용되는 조사와 동사를 기본적으로 포함하고, 온·오프라인 쇼핑몰 전문가의 자문을 통해 사용자들이 상품을 검색할 때 자주 사용하는 단어를 파악하여 작성하였다.

III. 시스템 구현 및 평가

이 장에서는 본 논문에서 구축한 상품 도메인 온톨로지를 이용하여 상품 검색을 했을 경우 얼마나 정확하게 원하는 결과를 검색해 내는지 직접 시스템을 구축하여 실험하였다.

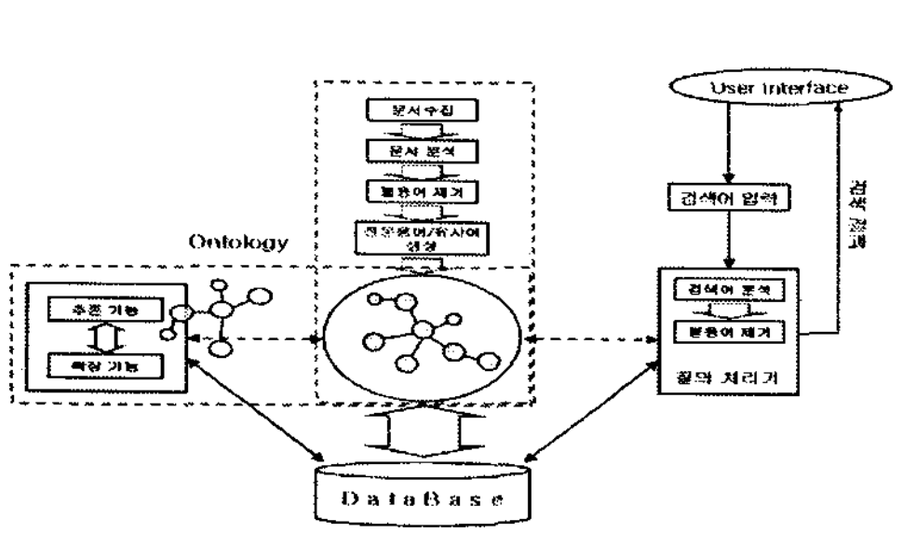

그림 3. 온톨로지 기반의 상품 검색 시스템

본 논문에서 구축하고자 하는 검색 시스템의 구조는 그림 3과 같다. 실험에 사용할 상품 데이터는 현재 전국적으로 온·오프라인 할인점을 운영중에 있는 A할인점의 상품 데이터 약 40,000여개의 품목을 Oracle 9i로 구축하여 사용하였다.

상품 검색 절차는 먼저 사용자 인터페이스를 통해 사용자가 입력한 검색어를 질의 처리기에 전송한다. 질의 처리기에 입력된 검색어는 불용어 제거를 통해 검색어를 보다 명확하게 재구성한다. 재구성된 검색어에 온톨로지를 통해 의미적인 내용을 부여함으로써 상품 검색 시 사용자가 원하는 상품을 정확하게 데이터베이스에서 검색할 수 있도록 하였다.

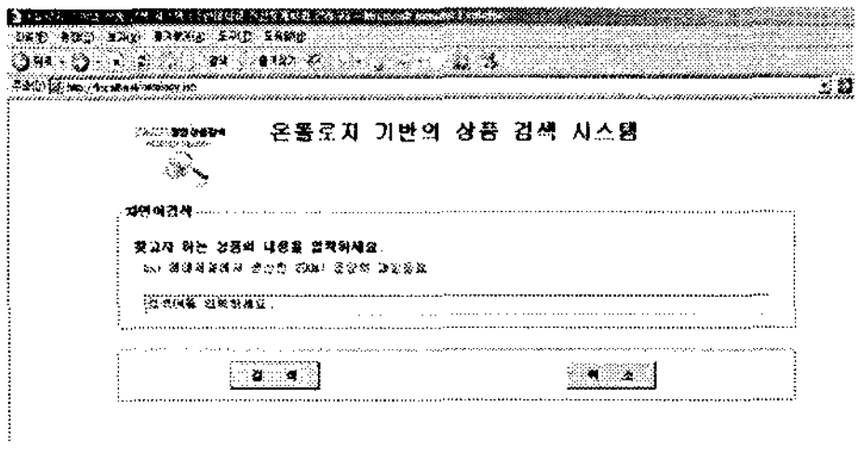

그림 4. 사용자 인터페이스(UI) 화면

그림 4는 본 논문에서 상품 검색을 위해 사용하게 될 사용자 인터페이스 화면과 소스 코드 일부를 보여주고 있다. 사용자 인터페이스는 JSP와 PowerBuilder9.0을 이용하여 웹상에서 사용 가능하도록 직접 구현하였다.

3.1 자연어 검색 시스템 실험 및 평가

대부분의 사용자들은 찾고자 하는 상품에 대한 기본적인 지식이 없는 경우가 많다. 이러한 경우 사용자들은 자신이 찾고자 하는 상품을 검색할 때 어떤 키워드로 상품을 검색해야 하는지 알지 못하기 때문에 정확한 검색을 하기 어려웠다.

반면 온톨로지를 이용한 자연어 검색을 활용하면 이와 같은 문제점을 쉽게 해결할 수 있다.

찾고자 하는 상품에 관한 검색 문장을 입력한 후 입력된 검색 문장을 불용어 리스트를 통해 불필요한 단어를 제거하여 검색어를 재구성한다. 이렇게 재구성된 검색어를 가지고 상품을 검색하게 된다.

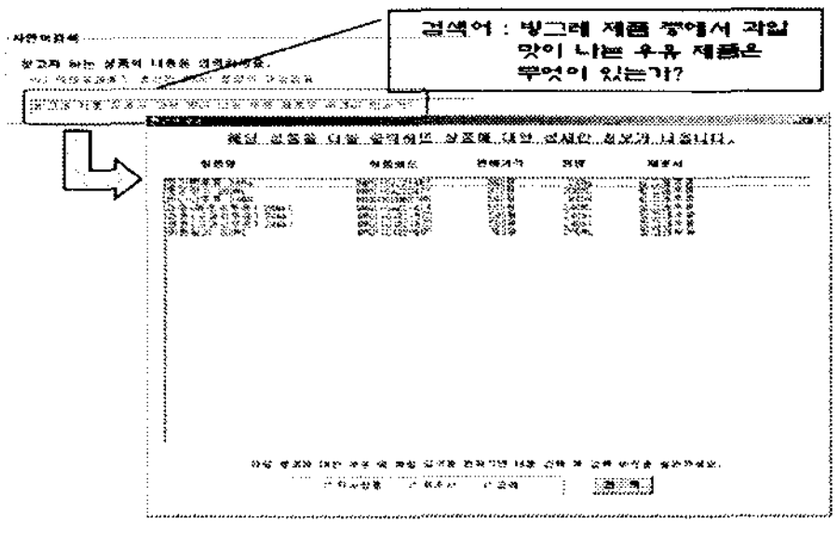

그림 5. 자연어 검색 결과 1

그림 5는 자연어 검색 문장을 온톨로지를 이용해 검색한 결과를 보여주고 있다. 그림에서 보면 "빙그레 제품 중에서 과일 맛이 나는 우유 제품은 무엇이 있는가?"라는 검색 문장으로 상품을 검색하였다.

위 검색 문장을 불용어 리스트를 통해 불필요한 단어를 삭제하게 되면 "빙그레 제품", "과일 맛", "우유 제품"으로 재구성이 되며 이렇게 재구성된 세 개의 단어를 검색어로 사용하게 된다. 이 중 "과일 맛"이라는 검색어는 표 3과 같이 온톨로지에 정의해 놓은 과일 관련 클래스 정보를 토대로 내용을 확장하여 검색을 하게 된다.

표 3. 과일 관련 클래스 일부

상위클래스	하위 클래스	
과 일	사 과	복숭아
	배	자 두
	딸 기	참 외
	포 도	바나나
	키 위	· · ·

결국 그림 5의 결과처럼 "과일"을 상위 클래스로 갖는 모든 하위 클래스에 속해 있는 상품들을 검색 대상으로 하고 재구성된 검색어인 "빙그레 제품"과 "우유 제품"을 공통적으로 포함하고 있는 상품들을 검색하게 된다.

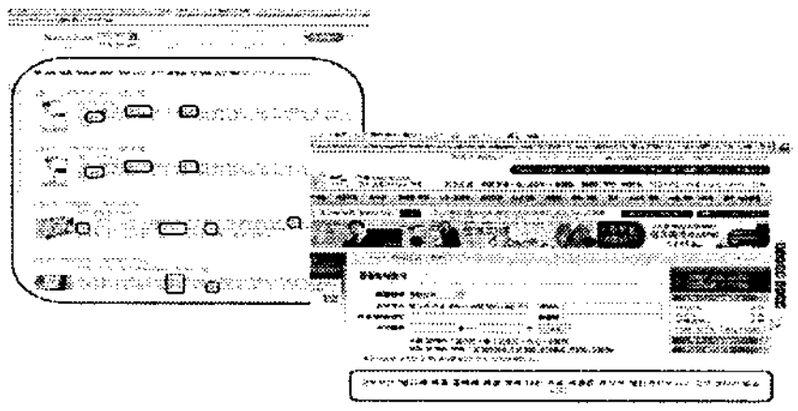

그림 6. 국내 온라인 할인점에서 검색한 결과

그림 6은 키워드를 기반으로 운영 중에 있는 국내 온라인 쇼핑몰 L할인점(뒷쪽 그림)과 E할인점에서 그림 5와 똑같은 검색문장으로 검색 한 결과이다. L할인점의 경우 "제품은", "제품", "맛", "맛이"와 같이 검색 조건과 전혀 무관한 단어를 검색어로 사용하여 검색을 진행함으로써 검색문장과 전혀 관련없는 상품들을 결과로 나타내었다. E할인점은 검색 조건에 맞는 상품을 단 한 가지도 검색해 내지 못했다.

국내 할인점의 검색 결과에서 나타나듯 키워드를 기반으로 하는 검색에서는 자연어로 입력된 검색문장을 통해 정확한 검색 결과를 얻어내기 힘들다는 것을 알 수 있다.

반면 온톨로지 기반의 자연어 검색을 이용할 경우, 불용어 리스트를 통해 검색문장을 재구성하고 검색어를 의미적으로 해석하여 검색을 진행하기 때문에 보다 정확한 결과를 얻을 수 있는 장점이 있다.

Ⅳ. 결론 및 향후 과제

본 논문에서는 온톨로지를 상품 검색에 활용하기 위해 상품 도메인 온톨로지를 직접 구축하고, 자연어검색을 통해 얼마나 정확한 검색을 하게 되는지 실험하였다.

그 결과, 온톨로지에 구축된 클래스 개념과 속성, 유의어 등 다양한 온톨로지 기능을 통해 사용자가 원하는 상품을 정확히 찾아내는 걸 확인하였다. 또한, 키워드를 기반으로 운영 중에 있는 국내 온라인 할인점과의 결과 비교를 통해 본 논문에서 구축한 온톨로지 기반의 자연어 검색 시스템이 보다 정확한 결과를 나타냄을 입증하였다.

이처럼 온톨로지를 이용하여 검색할 경우 검색어 자체에 의미를 부여함으로써 사용자가 원하는 상품을 보다 정확하게 검색할 수 있다는 장점이 있다. 매년 새로운 신상품이 다량으로 출시되고 있는 현 시점에서 초기에 구축해 놓은 온톨로지를 영구적으로 이용한다는 것은 불가능한 일이다. 정확한 상품 검색을 처리하기 위해서는 신상품과 새로운 분류의 상품들에 맞게 지속적으로 온톨로지를 수정 및 확장해 나가야 할 것이며, 상품 데이터베이스 설계에 있어서도 보다 간단하고 명확한 설계가 이루어져야 할 것이다.

참고문헌

[1] 김홍기, 김학래, 이강찬, 정지훈, 이재호 외, "월드 와이드 웹에서 시맨틱 웹으로," 『마이크로소프트웨어』, pp.242-301, 2002.

[2] Guarino, N., "Formal Ontology and Information Systems," Proceedings of FOIS'98, pp.3-15, 1998.

[3] Kuhanandha Mahalingam and Michael N. Huhns "An Ontology Tool for Query Formulation in an Agent-Based Context"

[4] Robert Neches et. al. "Enabling Technology for Knowledge Sharing" AI Magazine, Winter, 1991.

[5] Rudi Studer, V. Richard Benjamins, and Dieter Fensel. "Knowledge Engineering : Principles and Methods." Data & Knowledge Engineering. Vol.25, No.1-2, pp.184-185, 1998.

[6] 김영민, 이상춘, "시맨틱을 이용한 연구논문 검색 시스템", 『인터넷 정보학회 논문지』 Vol.4 No.3, pp.15-22, 2003.

[7] 한국전산원, "웹 온톨로지 개발 지침 연구", 2004.

[8] http://protege.stanford.edu.