의미적 언어자원을 활용한 과학기술정보 검색 서비스 개선

Improvement of Science and Technology Information Retrieval Service using Semantic Language Resource

조민희, 최성필, 최호섭, 윤화묵
한국과학기술정보연구원

Cho Min-Hee, Choi Sung-Pil, Choi Ho-Seop, Yoon Hwa-Mook.
Korea Institute of Science and Technology Information.

요약

현재 한국과학기술정보연구원의 과학기술정보 포털 서비스는 방대한 전문용어를 포함한 문서를 서비스하고 있으므로 포괄적인 질의어만으로는 사용자의 의도를 반영한 검색 결과를 얻을 수 없다. 따라서 본 연구에서는 의미적 언어자원으로 알려진 사용자 어휘지능망(U-WIN)의 동의어, 유의어, 관련어, 하위어, 상위어 관계 정보를 활용하여 검색어 자동 추천, 관련 단어 제시, 질의어 확장 등을 서비스에 반영하는 사용자 중심의 검색 서비스 요소를 제안한다. 이러한 어휘지능망의 의미 관계 정보를 활용한 서비스 요소를 통해 현재의 과학기술정보서비스의 검색 만족도를 향상시키는 동시에 사용자가 요구하는 정보를 빠르고 정확하게 검색할 수 있는 서비스 환경으로 개선시키고자 한다.

Abstract

KISTI portal service is currently presenting the documents with many terminologies, so users can't find the results having their intention by using an umbrella query. In this paper, we suggest user oriented retrieval service that reflects query auto-complete, related-word suggestion and query expansion that uses nouns and relationships of U-WIN which is known as a semantic language resource. We intend to advance the retrieval satisfaction of current science & technology information service by using U-WIN's semantic information and improve the service environment that user can retrieve what they want quickly and exactly.

I. 서 론

정보 검색 시스템은 사용자의 질의어와 일치하는 색인어를 가지고 있는 문서를 검색하기 때문에 질의어와 색인어가 일치하지 않는 경우 검색이 되지 않는다. 또한 대부분의 사용자들은 포괄적인 질의어를 통하여 요구하는 결과를 얻고자 하므로 정확한 검색을 수행하기 어렵다[2, 3]. 특히 과학기술정보 포털 서비스에서는 방대한 전문용어를 포함한 문서를 서비스 하고 있으므로 원시 질의어만으로는 사용자의 의도를 반영한 검색 결과를 쉽게 얻을 수 없다. 검색 결과를 얻더라도 일반 사용자들은 자신이 원하는 문서를 찾기 위해 질의어를 재형성하거나 관련 웹페이지들의 링크를 통하여 해당 문서를 찾는 등 시간과 노력이 많이 들게 된다.

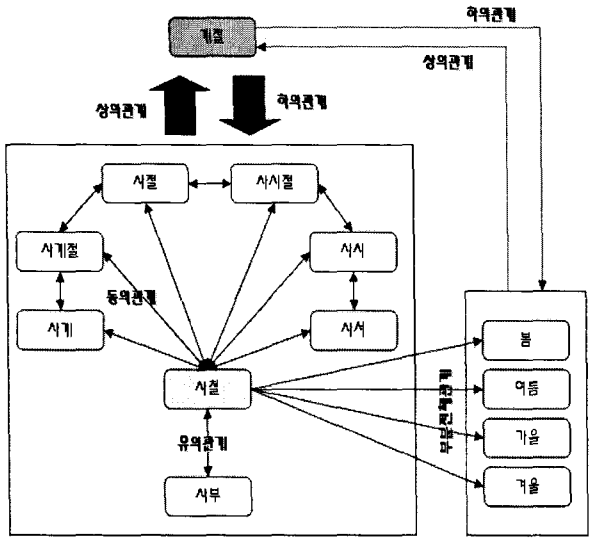
따라서 본 논문에서는 사용자의 검색 만족도를 향상 시키기 위하여 사용자 어휘 지능망(U-WIN: User-Word Intelligent Network)의 동의어, 대역어, 유의어, 관련어, 하위어, 상위어 관계 정보를 활용하고자 한다. 검색 효율을 저하시키는 색인어와 사용자 질의어의 불일치 문제를 U-WIN의 의미관계 정보를 활용하여 해결할 수 있는 방법을 제시하고, 전문지식이 없는 사용자들이 질의어를 쉽게 재조합하고 검색할 수 있도록 U-WIN의 정보들을 검색 인터페이스 환경에서 제공함으로써 사용자가 요구하는 적합 문서를 빠르고 정확하게 찾을 수 있도록 하고자 한다.

II. 한국어 어휘 지능망 U-WIN

사람의 언어를 이해하는 자연어처리 시스템을 개발하기 위해서는 의미처리를 위한지식 베이스(knowledge base)가 필요하다. 그러한 결과물로 온톨로지(ontology)와 시소러스(thesaurus)가 만들어지고 있다. 가장 많은 활용을 보이는 것이 Prinston 대학의 WordNet이다[5, 9]. 이것은 영어에 대한 어휘데이터베이스이므로 한국어 문서 적용엔 적합하지 않은 단점이 있다.

한국어에 대한 지식베이스 구축 작업도 기관과 학교 등에서 많이 현재 이루어지고 있다. 그 가운데 울산대 한국어처리연구실에서 개발하고 있는 사용자 어휘 지능망(U-WIN: User-Word Intelligent Network)은 한국어정보처리를 비롯한 정보

검색, 기계번역, 시맨틱 웹 등 다양한 분야에 이용될 수 있는 한국어에 대한 어휘 데이터베이스이다[4, 7, 8]. '인간이 가지는 여러 관념 속에서 공통적인 속성을 기반으로, 인간의 보편적인 인지 체계와 개념 관계를 파악하여 이것을 표현한 언어를 대상으로 한 형식적이고 명세적인 어휘의 의미적 · 개념적 네트워크로서, 한국어 WordNet을 비롯한 기타 시소러스, 어휘 분류, 온톨로지를 연결시킨 전체적인 어휘 네트워크'라고 말한다. 현재 60만 이상의 일반 어휘와 전문용어로 구성되어 있다[4]. U-WIN은 [그림 1]과 같이 단어간의 상하관계, 동의관계, 유의관계, 부분-전체관계, 반의관계, 관련어 등의 의미 관계 정보를 가지고 있다. 본 논문에서 제시한 서비스에서 활용되는 의미 관계는 다음과 같다.



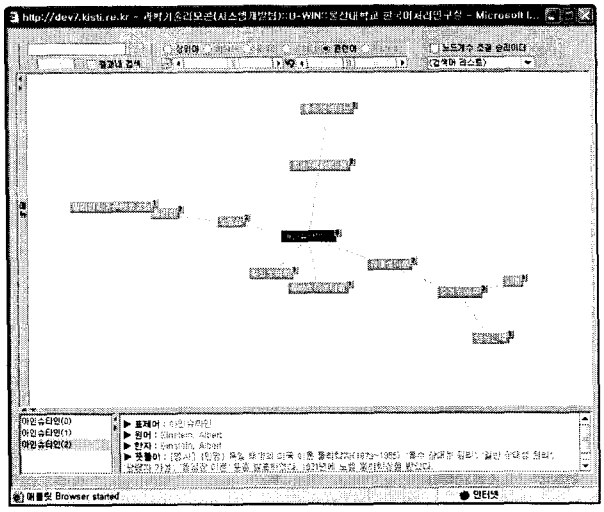
▶▶ 그림 1. U-WIN의 의미적 관계 예

◈ 동의어 관계와 유의어 관계

의미상으로 동일한 뜻을 가지지만, 동의어 관계는 두 단어가 문맥상으로 완전한 대체가 가능하고, 유의어 관계는 단어를 완전 대체할 수 없다. [그림 1]을 통하여 '사철'과 동의어 및 유의어 관계를 맺는 단어를 볼 수 있다.

◈ 상 · 하위어 관계

상 · 하위어 관계는 어휘에 대한 의미 계층을 나타내는 것으로, 상위어는 개념적으로 포괄적인 의미를 가지고 있고 하위어는 구체적인 개념을 가지고 있다. 예를 들어 '줄기세포'의 상위어는 '세포'이고, 하위어는 '배아줄기세포' 등이다. 질의어가 포괄적이거나 구체적인 경우 상 · 하위어 관계의 어휘를 이용하여 질의어를 재구성함으로써 검색 결과의 범위를 조절할 수 있다 .



▶▶ 그림 2. U-WIN의 Visualization 서비스를 통한 검색 결과 화면

◈ 관련어

특정 어휘와 밀접한 관계를 가지는 어휘를 관련(related term)라 한다. [그림 2]를 통하여 '아인슈타인'은 '상대성이론', '특수상대성이론', '노벨상', 'theory of relativity' 등과 의미적 관련을 맺고 있음을 알 수 있다.

◈ 기타

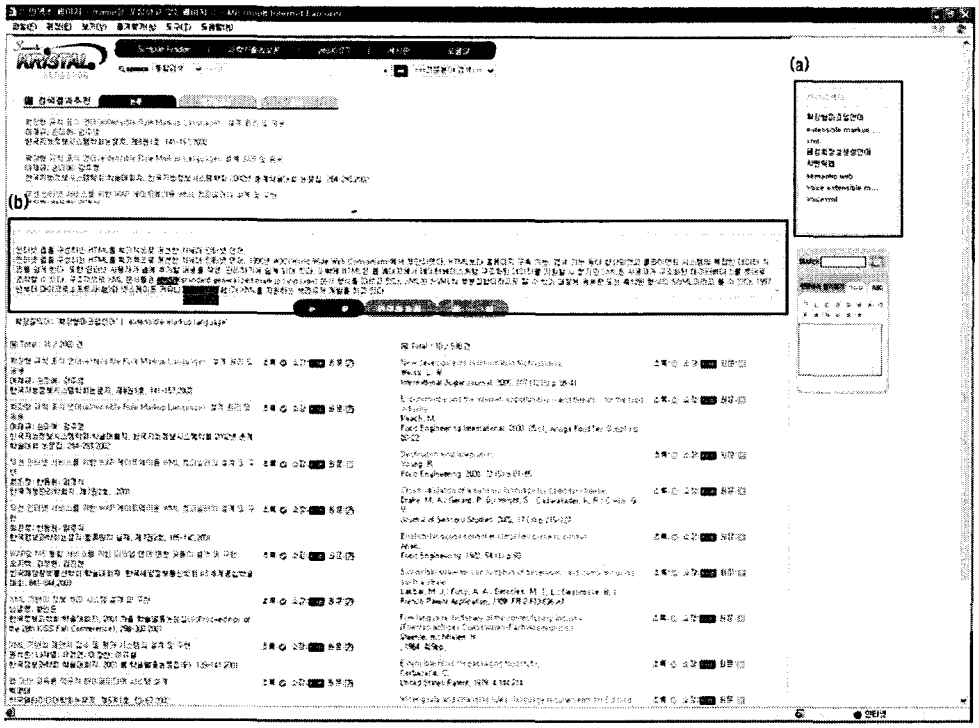
부모 노드가 같은 형제 노드 어휘(Sibling Term)를 통하여 관련어 제시 등을 할 수 있다. 예를 들어 '아인슈타인'의 부모노드가 '물리학자'이므로 같은 부모 노드를 갖는 '가보르', '가이거', '리프만', '볼츠만' 등을 의미적 관련어로 제시할 수 있다.

또한 U-WIN은 한글 단어에 대한 영어 표제어 정보, 영어 단어에 대한 한글 표제어 정보를 가지고 있다. 이러한 대역어 정보는 한영 교차 검색 서비스에 응용될 수 있다.

Ⅲ. 서비스 검색 요소 기술

방대한 웹 문서 속에서 기존의 단일 키워드 매칭의 검색 결과 방법은 검색 결과의 양도 클 뿐만 아니라 검색 결과 또한 사용자의 의도를 반영하기 어렵다[1]. 사용자가 요구하는 적합한 문서를 찾아내기 위해서는 정확한 검색어가 필요하다. 하지만 검색 결과 문서에 대한 전문 지식이 없는 대부분의 사용자들은 포괄적인 단일 질의어 검색으로 적합 문서를 찾기를 바란다. 따라서 시스템은 사용자가 원하는 검색 결과를 얻을 수 있는 환경을 제공해야 한다.

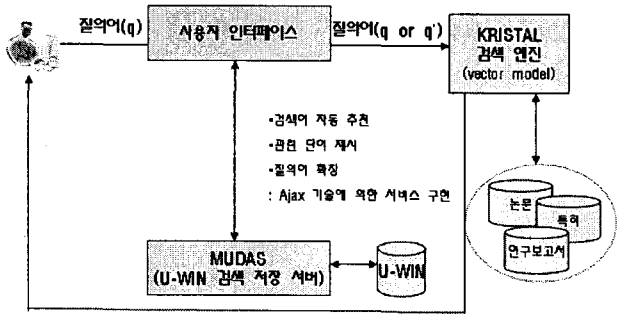
현재 www.yeskisti.net에서 서비스하고 있는 과학기술 문서는 웹문서와 달리 가공된 전문 데이터이므로 포괄적인 질의어로는 사용자가 요구하는 적합한 문서를 찾아내기 어렵다. 따라서 본 논문에서는 다음과 같은 검색 서비스를 제안한다. 사



▶▶ 그림 3. Semantic KRISTAL 화면

용자가 검색하고자 하는 대상어에 대한 검색어 자동 추천 기능, 질의에 사용된 단어와 연관된 단어를 제시하는 기능, 질의어와 의미 관계를 가진 단어를 통한 질의어 확장 검색 기능 등이 있다. 본 논문에서 제안한 서비스 기능 구현을 위한 사용자 인터페이스 시맨틱 크리스탈(Semantic KRISTAL)을

[그림 4]와 같이 구성하였다[11]. 사용자가 질의어(q)를 던질 경우에 사용자 인터페이스는 U-WIN의 관계정보 및 어휘정보를 저장하고 검색할 수 있는 다목적 사전 통합 액세스 시스템(Multi-purpose Dictionary Accessing System: MUDAS)을 통하여 질의어와 의미 관계를 가진 단어들에 대한 정보를 받게 된다[12]. 사용자의 선택에 의해 새로운 질의어(q')가 생성되거나 재검색이 이루어진다.

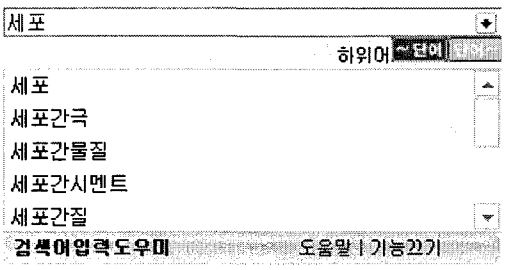


▶▶ 그림 4. Semantic KRISTAL 시스템 구성도

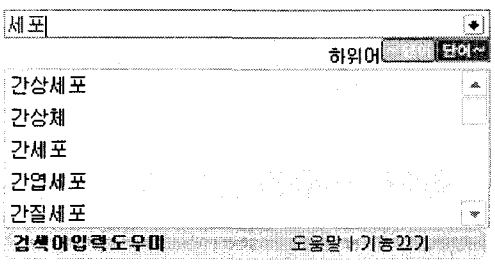
1. 검색어 추천 기능

현재 많은 웹 포털에서 사용자의 로그 정보를 활용하여 서비스되고 있는 기능이다. 본 논문에서는 U-WIN의 어휘정보를 활용하여 일차적으로 검색어 추천 기능을 제공하고, 이차적으로 U-WIN의 계층 구조를 활용하여 검색어의 하위어 단어를 추천하고 있다.

이용자는 철자가 아리송한 전문용어들에 대하여 [그림5]와 같이 쉽게 단어를 찾을 수 있고, 검색 결과를 얻기 전에 [그림6]과 같이 검색어와 의미적으로 하위 관계를 가진 단어들을 통하여 검색 결과의 범위를 좁혀 나갈 수 있다.



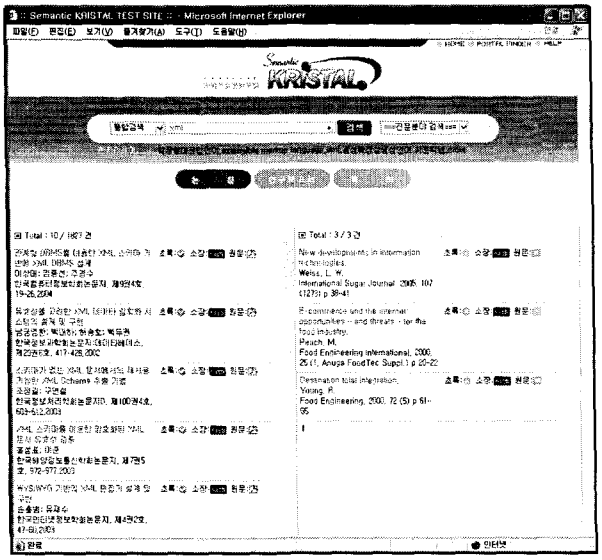
▶▶ 그림 5. 검색어 자동 추천 화면



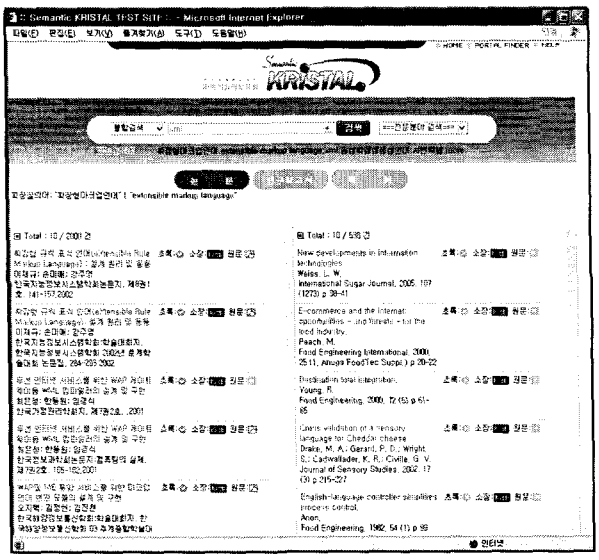
▶▶ 그림 6. 검색어에 대한 하위어 자동 추천 화면

2. 관련 단어 제시 기능

이용자가 검색한 질의어와 의미적으로 관련성이 있는 단어들을 [그림 3]의 (a)와 같이 실시간으로 화면에 제시하면서 이용자들의 이차 검색을 지원하고 있다. 이용자는 찾으려는 정보에 대해 쉽게 다가갈 수 있고 다양한 정보로의 접근이 가능하게 된다. 동의어, 관련어, 상위어, 하위어, 형제 노드의 어휘 순으로 관련어 정보를 출력한다.



▶▶ 그림 7. 'xml' 검색 결과 화면(질의어 확장 전)



▶▶ 그림 8. 'xml' 검색 결과 화면(질의어 확장 후)

3. 질의어 확장 검색

사용자가 입력한 원질의어와 관련이 있는 단어를 원질의어에 추가시킴으로써 관련 문서를 정확하게 더 많이 찾을 수 있다고 주장되어 왔고 이러한 연구는 검색 효율 향상을 위해 계속되고 있다[6, 9]. 질의어와 관련된 단어를 모두 확장시키게 되면 재현율은 높아질 수 있으나 정확성은 떨어지게 된다. 따라서 같은 의미를 가지는 동의어, 유의어, 한글대역어, 영어대역어를 통하여 질의어를 자동으로 확장한다. 이러한 질의어 확장을 통하여 색인어와 검색어의 불일치 문제를 해소시킬 뿐만 아니라 관련 문서를 정확하게 많이 찾을 수 있다.

[그림 7]과 [그림 8]을 비교해 볼 때, 질의어 확장시 더 많은 관련문서가 찾아짐이 보여진다. 또한 한영 대역어를 이용한 질의어 확장을 통하여 한글문서와 영어문서를 동시에 검색해 이용자들에게 출력해줌으로써 이용자들의 재검색 횟수를 줄이게 한다.

하지만 질의어 확장을 통한 검색 결과가 이용자에게 유효하지 않을 수 있으므로 이용자가 원질의어의 관련 단어 정보를 통하여 직접 질의어를 재구성할 수 있는 환경을 제공해야 한다. [그림 3]의 (b)는 검색어에 대한 사전 뜻풀이를 제시하고, 내부적으로 이차 검색과 검색어 추가를 실시간으로 할 수 있는 인터페이스를 마련하여 이용자의 재검색이 쉽게 이루어지도록 하였다. 뜻풀이를 구성하는 단어들은 의미를 나타내는 중요한 단어이므로 질의어 재구성시에 월등한 역할을 함이 보여졌다.

Ⅳ. 결 론

본 논문에서는 U-WIN의 의미 관계 정보를 이용한 검색어 추천, 관련 단어 제시, 질의어 확장 기능 등을 인터페이스에서 제공함으로써 이용자의 오퍼레이션을 최소화하여 재검색을 쉽게 하고 검색 내내 제공되는 의미 정보를 통하여 사용자가 요구하는 적합 문서를 빠르고 정확하게 찾을 수 있게 되었다.

동일한 단어일지라도 분야 정보가 다른 경우 다른 의미로 해석되는 경우가 많다. 어휘망은 의미 관계에 의해 어휘들이 연결되어 있으므로 다양한 의미를 가지는 어휘의 경우 이용자가 원하는 의미가 선택되지 않으면 동의어, 유의어, 상하위어, 관련어 등의 정보가 사용자에게 유효하지 않을 수 있다. 이러한 문제를 개선하기 위한 연구가 앞으로 지속적으로 필요할 것이다.

▌참 고 문 헌▐

[1] 김영민 "시맨틱을 이용한 연구 논문 검색 시스템", 한국 인터넷 정보학회, 제4권, 제3호, pp.15-22, 2003.
[2] 김형일 "워드넷을 이용한 검색 질의어의 모호성 해결", 한국정보 과학회, 제27권, 제 2호, pp.75-77, 2000.
[3] 박수현 "한국어 정보 검색 시스템에서 시소러스를 이용한 검색 효율 향상, 동서대학교 논문집, pp.335-344, 1999.
[4] 옥철영 "사용자 어휘지능망", KRISTAL 2006 발표 자료집, 2006.
[5] 임성신 "한국어 워드넷의 구축",한글언어인지학술대회, 제16권, 제 1호, pp.106-111,2004.
[6] 조원건 "시소러스를 활용한 정보검색시스템의 검색효율 향상

에 관한 연구",연세대학교 석사학위 논문, 2002.
[7] 최호섭 "한국어 의미망 구축과 활용-명사를 중심으로",한국어학 제17집, 2002.
[8] 최호섭 "사전을 기반으로 한 한국어 의미망 구축과 활용", 한국 정보 과학회, 제29권, 제2호, pp.448-450, 2002.
[9] George A. Miller, "Introduction to WordNet: An on-line lexical database",International Journal of Lexicography, 1990.
[10] Rijsbergen, Van C. J., "A new theoretical framework for information retrieval", ACM, pp.194-200,1986.
[11] http://www.kristalinfo.com:8080/sem_kristal
[12] http://www.kristalinfo.com/K-Lab/mudas/mudas.php