

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. 서론

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 서비스하고 있으므로 원시 질의어만으로는 사용자의 의도를 반영한 검색 결과를 쉽게 얻을 수 없다. 검색 결과를 언더라도 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 향상시키기 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 없

는 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 위한 지식 베이스(knowledge base)가 필요하다. 그러한 결과물로 온톨로지(ontology)와 시소러스(thesaurus)가 만들어지고 있다. 가장 많은 활용을 보이는 것이 Princeton 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

