시소러스를 이용한 정보검색 지원 시스템

Information Retrieval Support System Using Thesaurus

신성혁, 심빈구, 이승준, 최영주
오롬정보

Shin Sung-Hyuk, Shim Bin-Gu, Lee Seoung-Jun,
Choi Young-Ju, Kwon Rae-Nam
OromInfo.

요약

본 논문에서는 빠른 속도로 증가하고 있는 인터넷상의 정보와 서비스를 검색함에 있어서 정보의 과부화(information overload)로 인한 검색의 비효율성을 극복하기 위하여 시소러스를 이용하여 정보검색을 효율적으로 수행하기 위한 오딘(Odin)시스템을 제안하고자 한다. 오딘시스템을 이용하여 원하는 키워드의 추천어를 이용하여 효율 적으로 검색할 수 있다.

Abstract

The amounts of information and services are rapidly increasing. This causes inefficiency of retrievals of information overload problem. To overcome this kind of problems, This article suggests Odin system, keyword suggestion system based on Thesaurus. Odin system will suggest appropriate search-words and this leads to the user's satisfied result.

I. 서 론

1900년대 중반 처음 컴퓨터가 등장한 이후로 수십 년 동안 놀라운 속도로 발전에 발전을 계속해 왔다. 급속한 기술적인 발전과 더불어 대량 생산으로 인한 가격의 하락은 일반인들 누구나 컴퓨터를 접하게 되었고 나아가 누구에게나 필요한 생활필수품이 되고 말았다. 초창기에는 복잡한 수식으로 활용되었던 컴퓨터가 이제는 다양한 종류의 정보를 저장하는 용도로도 많이 활용되고 있다. 컴퓨터에 저장되는 정보들은 문서정보 외에 음성 정보, 영상 정보 등 수많은 형태가 존재하지만 여전히 문서 정보가 월등히 많은 것은 사실이다. 뿐만 아니라 인터넷상의 정보와 다양한 서비스가 급속한 속도로 증가하고 있다. 이런 정보를 검색하는 작업이 점점 어려운 문제가 되는 것은 피할 수 없게 되었다. 이러한 문제점을 해결하기 위하여 다양한 정보 검색 시스템이 개발되어서 사용자에게 필요한 정보를 검색하는 일을 돕고 있다. 하지만 정확한 용어를 모르거나 혹은 일치하지 않는 단어로 인하여 정보를 상당부분을 놓치는 경우가 발생하기도 한다.

본 논문은 시소러스를 이용하여 사용자가 입력한 키워드뿐만 아니라 추천어를 제공하여 사용자들에게 효율적인 검색을 지원하는 오딘(Odin)시스템을 제안하고자 한다.

본 논문의 구성은 2장에서는 관련연구를 살펴보고 3장에서는 오딘(Odin)시스템의 개요를 살펴보고 4장에서는 시스템 구성 및 적용사례를 살펴보고 5장에서는 향후 연구과제에 대하여 살펴보기로 한다.

II. 관련연구

시소러스(Thesaurus) 어원(語源)은 그리스어로 '지식의 보고(寶庫)'라는 뜻인데, 로제가 영의의 어휘를 내용상으로 분류하여 관련어(關聯語)를 표시한 사전을 만들어 시소러스라는 이름을 붙인 이래, 그러한 사전을 시소러스라고 일컫게 되었다. 이러한 시소러스는 분류와 사전의 결합으로 상위 및 하위개념 사이의 전후관계를 명확히 하기 위해서 공식적으로 조직, 통제된 색인어의 어휘로 실물적, 추상적 세계의 개념대상의 대응 상징체계인 용어에 대한 상호간의 관계를 표현한 지식베이스로써 인간의 학습, 탐구활동 등 제반 지식활동의 대상이 되는 개념(용어)간의 관계를 표현한 지식구조 맵이다.[1][2] 정보학 사전(사공철 외 2001, 197)에서는 시소러스에 대하여 다음과 같이 설명하고 있다. 즉, "개념사이의 관계를 명확하게 나타내기 위하여 일정한 형식으로 조직되고 통제된 색인어의 어휘집이다. 문헌에 대한 주제 분석을 통해 얻어진 주제 개념들을 그 시스템이 사용하고 있는 색인표목으로 변환시키기 위한 표준용어를 제공하며, 색인 작성 시 적절한 색인표목의 선택과 색인어의 통제를 위해서 사용된다. 정보검색 시 시소러스의 활용은 사용자의 검색요구가 있을 경우, 이용자가 요청한 내용(질의어)에 대한 개념(주제)분석을 통하여 주제용어를 선택하고 선택한 용어를 그 시스템이 사용하고 있는 표준용어로 바꾸기 위하여 사용된다." 이어서 현대적인 시소러스의 주요목적을 다음과 같이 열거하고 있다.

1) 특정 주제 분야의 지식구조를 보여준다.
2) 특정 주제 분야의 표준어휘를 제공하여 색인 작업의 일관성을 유지한다.
3) 용어간의 참조체계를 제공한다.
4) 사용자가 유사한 어휘 중에서 정확한 검색어를 선택할 수 있도록 도와준다.
5) 새로운 개념을 기존 개념들 간의 관계 체계에 맞추어 제 자리를 잡게 한다.
6) 계층적 분류를 제공하여 탐색을 체계적으로 확장하거나 축소할 수 있게 한다.
7) 장기적으로는 특정 분야의 어휘사용을 표준화하는 수단이 될 수도 있다.[3]

이러한 목적과 특성을 이용하여서 사용자가 키워드를 이용하여 검색 시 시소러스는 정확한 검색을 위하여 효율적으로 지원할 수 있는 장점을 가진다.

III. 오딘(Odin)시스템 개요

1. 오딘(Odin)이란?

오딘 시스템은 추천검색시스템과 사용자 로그 시스템으로 구성된다. 추천검색시스템을 후긴 시스템, 로그지원시스템을 무닌 시스템이라고 이름을 붙였다.

2. 시소러스 용어 관계

오딘시스템에 사용되고 있는 시소러스의 관계의 종류 및 관계지시기호는 기본적으로 ISO 2788:1986(E)를 따른다. 다만 우리나라 특성을 고려하여 KDC, DDC, NK, SK, SNN, 각종 외국어 코드를 추가하였다.

[표 1] ISO 2778 용어관계 정의

용어관계정의	설명
BT(broader term)	상위개념어
BTG(broader term/generic)	상위개념어/屬
BTI(broader term/instance)	상위개념어/사례
BTP(broader term/partial)	상위개념어/부분
NT(narrower term)	하위개념어
NTG(narrower term/generic)	하위개념어/屬
NTI(narrower term/instance)	하위개념어/사례
NTP(narrower term/partial)	하위개념어/부
RT(related term)	관련어
SN(scope note)	범위주기
TT(top term)	최상위개념어
UF(used for 혹은 use for)	비우선어, ~대신 사용하라
USE(use)	우선어, ~를 사용하라

[표 2] 오딘에서 사용되는 시소러스 추가 코드

용어관계정의	설명
KDC	한국십진분류기호(4판)
DDC	듀이십진분류기호(21판)
NK(North Korea)	북한어
SK(South Korea)	북한어에 대응되는 국어
SNN(Scientific Name)	학명
ENG(영어), ESP(스페인어), FRA(프랑스어), GER(독일어), GRE(그리스어), ITA(이탈리아어), JPN(일본어), LAT(라틴어), MON(몽골어), RUS(러시아어), SAN(범어), CHN(중국어)	각 외국어에 대응되는 한글관계도 정의

3. 오딘(Odin)시스템 특징

현재 수많은 문서들을 검색하기 위하여 많은 검색 시스템이 개발되었고 현재 서비스 되고 있다. 이들 시스템의 경우 대용량 자료 검색 및 문서의 우선 순의 정책으로 사용자들의 정보검색을 지원하고 있다. 하지만 사용자가 입력한 키워드로만 검색을 하기 때문에 키워드와 매칭 되는 단어들에 한해서만 검색이 지원되고 있는 것은 사실이다. 이러한 단점을 해결하고자 하는 것이 오딘시스템이다. 사용자가 비타민이라는 단어로 정보검색을 했을 경우 현재의 검색시스템은 비타민에 한정된 정보들을 제공한다. 질의한 용어에 대하여 정확률은 높을 수 있으나 다양한 검색결과를 요구하는 사용자들에게 부족함이 있다. 따라서 정확률(Precision)뿐만 아니라 재현율(Recall) 까지 향상 시킬 수 있는 시스템이 오딘이다. 오딘의 경우 50만여 용어 시소러스에 기술되어 있는 관련 단어들을 이용하여 검색어를 추천하게 된다. 사용자는 비타민이라고 입력하였으나 영어로 된 Vitamin, 바이타민 등 기존 정보 검색 시스템에서 지원되지 않는 음차어나 동의어를 추천 검색어로 제공하여 사용자에게 더 많은 정보를 제공하고 많은 정보가 터무니없는 정보가 아닌 요구하는 정보들로 제공하기 때문에 정확률뿐만 아니라 재현율까지 향상 시킬 수 있다. 또한 사용자 수준별에 따른 검색 서비스도 제공할 수 있다. 현재의 정보 검색 시스템은 사용자 수준과 상관없이 같은 검색 결과를 제공한다. 비록 사용자가 비타민을 검색 했을 때 단순 비타민에 관한 정보를 원하는 경우도 있겠지만 전문적이고 학술적인 내용을 검색하고자 하는 경우 기존은 정보 검색 시스템의 경우는 사용자가 별도의 노력을 들여서 용어를 검색하여야 한다. 하지만 50만여 용어 시소러스에는 전문용어도 구축되어 있어 별도의 노력이 없이도 전문용어를 검색할 수 있다. 앞에서 사용한 비타민이라는 키워드를 이용하여 정보 검색을 했을 때 프로비타민과 같은 전

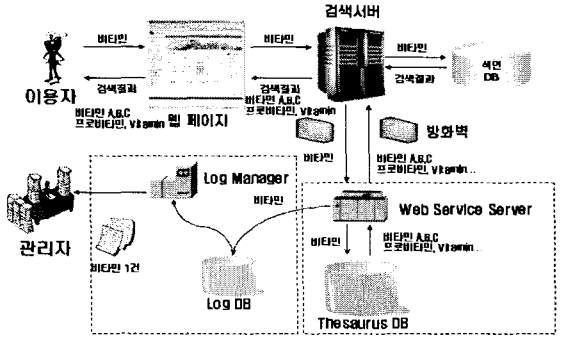
문용어까지 추천해 줌으로써 사용자가 수준에 맞추어 정보검색을 진행 할 수 있다.

이러한 추천 검색어는 50만여 용어 시소러스를 이용하여 검색어를 제공하는데 입력 받은 질의어를 시소러스 DB에 적용하여 USE/UF, NT, BT, RT, ENG, JPN 등 용어 키워드로 제공하게 되고 해당 시스템과 밀결합 되어 즉각적인 링크정보를 생성하여 사용자에게 제공한다. USE/UF, ENG, JPN을 우선적으로 제공하여 사전적 정보를 제공하고, NT 정보를 제공하여 검색의 초점을 제공하게 되며, BT 정보를 제공하여 검색의 범위를 확대할 수 있다. 이와 같은 추천검색어 시스템은 앞에서 언급한 후긴 시스템이라고 불리어 진다. 하지만 인간의 용어 사용은 무궁무진하며 이 시스템을 이용하는 사용자의 관심도에 따른 다른 용어들의 편차가 있을 수 있다. 이런 문제점을 해결하고자 하는 시스템이 무닌 시스템이다. 무닌 시스템은 자동으로 용어 시소러스에서 검색하지 못한 용어들을 정제하여 보관하게 되며 오딘 시스템을 사용하는 기관에 따른 키워드 및 사용빈도 등을 자동으로 분석하여 사용자에게 제공하고 검색하지 못한 용어들은 앞으로 구축될 용어 시소러스에 반영되어 구축된다. 따라서 사용자의 정보 검색 패턴에 따른 추천 검색어 지원도 가능하게 된다.

IV. 오딘(Odin)시스템 구성 및 적용사례

1. 오딘(Odin) 시스템 구성

오딘시스템은 13년간 전문가로 이루어진 구축팀 에서 구축한 50만여 용어 시소러스를 이용한다. 본 서비스는 정보 검색 추천 서비스로 기존의 검색서비스의 최소 수정으로 효율적인 검색의 향상을 가져 올 수 있다. SOAP을 이용한 XML Web Service로 OS와 상관없이 모든 플랫폼을 지원한고 ASP (Application Service Provider)기반 서비스이다. 시스템의 구성은 다음 그림과 같다.

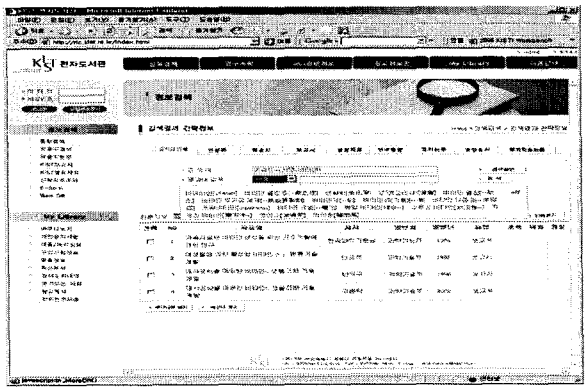


▶▶ 그림 1. 오딘 시스템 구성도

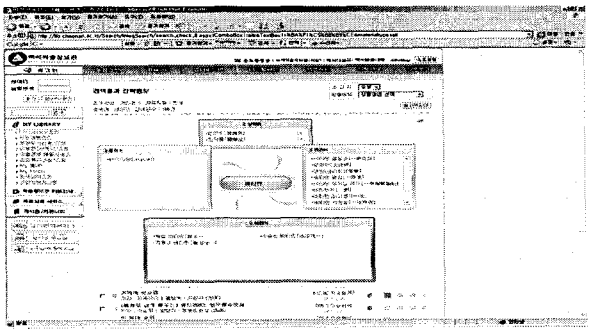
그림 1에서 볼 수 있듯이 오른쪽 시소러스 DB쪽의 추천 검색 시스템을 후긴 시스템이고 왼쪽 Log DB가 있는 쪽이 무닌 시스템이 되겠다.

2. 오딘 시스템 적용 사례

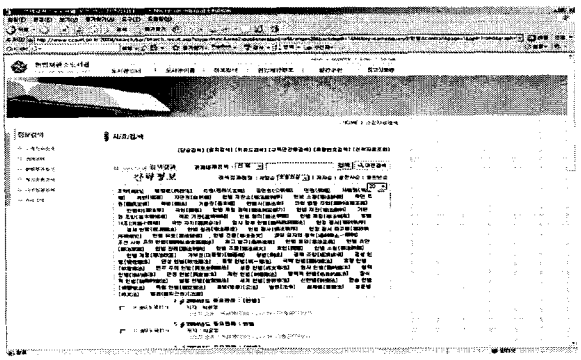
서지데이터를 검색하기 위한 추천 키워드를 제공하고 있으며 다년간의 전자도서관 개발 경험으로 기존 시스템과 완전 밀결합되어 전자도서관 이용자 통계 및 검색쿼리 통계를 제공하고 있다. 현재 KIST(http://ric.kist.re.kr)한국과학기술원, 헌법재판소(http://www.ccourt.go.kr/library/index.asp), 백석대학교(http://lib.cheonan.ac.kr)에서 적용되어 사용되고 있다.



▶▶ 그림 2. 적용사례 : KIST(http://ric.kist.re.kr)



▶▶ 그림 3. 적용사례 : 백석대학교(http://lib.cheonan.ac.kr/)



▶▶ 그림 4. 적용사례 : 헌법재판소(http://www.ccourt.go.kr/library/index.asp),

V. 향후 연구 과제

오딘 시스템을 이용한 추천어 검색 서비스를 제공함에 있어 향후 연구 과제로는 현재 문장으로 입력되는 키워드에 대하여 검색의 효율성이 많이 떨어지고 있다. 문장의 경우 사용자마다 강조하는 용어가 다르기 때문에 추천 검색어를 제공하기 싶지 않은 것이 사실이다. 또한 추출된 검색 실패한 용어들에 대하여 자동으로 구축하는 방안에 대하여 연구가 필요를 느낀다.

▌참 고 문 헌▐

[1] 최석두, 한상길 "시소러스 標準開發에 대한 硏究", 지식처리연구, 제1권, 제2호, 2000.
[2] 최석두 "한글 시소러스 구축 기준에 관한 연구, 연세대학교 박사학위논문, 2002.
[3] 유명희, "형용사 시소러스에 관한 연구", 이화여자대학교 석사학위 논문, 2003.