지능형 정보검색을 위한 지식 기반 시소러스

정성호°, 김민구
아주대학교 컴퓨터공학과

A Knowledge Based Thesaurus for Intelligent Information Retrieval

Chung-ho Chung°, Min-koo Kim
Dept of Computer Engineering, Ajou University

요 약

지식구조로 시소러스를 이용하는 기존의 정보검색 시스템들이 사용자에게 만족할 만한 검색결과를 제시하지 못하고 있다. 이것은 기존의 정보검색 시스템들이 이용하고 있는 시소러스 구조가 사람의 지식구조과 다르고, 시소러스를 이용하는 검색 방법이 사람의 검색 방법과 차이가 있기 때문이다. 본 논문에서는 어떤 분야의 인간 전문가가 해당분야에 관한 전문지식이 없는 일반인이 필요로 하는 정보를 찾아주는 방법을 모델링한 지능형 정보검색 시스템을 개발하기 위하여 인간 전문가의 지식구조를 모방한 시소러스 구조를 설계하였고, 인간 전문가의 검색 방법을 모방한 검색 방법을 고안하였다. 설계된 시소러스 구조에는 인간 전문가의 지식구조 내에 표현되어 있는 여러 종류의 관계들이 포함되어 있고, 고안된 검색방법은 관련도를 사용자의 질의어와 확장된 색인어 사이의 관계의 종류를 추론한 결과와 거리 단계를 고려하여 평가한다.

1 서론

현대 사회를 '정보화 사회'라고 한다. 이 말은 정보가 우리의 생활에 매우 중요하다는 것을 의미한다. 많은 정보 중에서 필요한 정보를 빨리 찾는 것은 매우 중요하지만 쉬운 일이 아니다. 필요한 정보를 빨리 찾아서 활용하는 개인이나 단체가 경쟁에서 이길 수 있고 이득을 얻을 수 있기 때문에, 사람들은 필요한 정보를 찾기 위해서 많은 시간과 노력을 투자하고 있다. 컴퓨터 관련 기술의 발전으로 많은 정보가 컴퓨터에 저장되어 있고, 컴퓨터의 대중화로 컴퓨터는 일상생활에 널리 사용되고 있다. 그래서, 컴퓨터에 저장되어 있는 많은 정보 중에서 사용자가 필요로 하는 정보를 찾아주는 정보검색 시스템들이 개발되었다. 그러나, 개발된 정보검색 시스템들은 사용자에게 만족할 만한 검색결과를 제시하지 못하고 있다. 개발된 정보검색 시스템들은 사용자가 필요로 하는 정보를 의미하는 색인어를 얻고 있을 경우에만 정보를 찾아주는 단순한 정보검색 시스템들이거나, 사람의 지식구조 및 검색방법과 차이가 있는 지식구조 및 검색방법을 이용하는 정보검색 시스템들이기 때문이다. 본 논문에서는, 어떤 분야의 인간 전문가가 해당분야에 관한 전문지식이 없는 일반인이 필요로 하는 정보를 찾아주는 방법을 모델링한 지능형 정보검색 시스템을 개발하기 위하여 인간 전문가의 지식구조를 모방한 시소러스 구조를 설계하였고, 인간 전문가의 검색 방법을 모방한 검색 방법을 고안하였다. 설계된 시소러스 구조에는 인간 전문가의 지식구조 내에 표현되어 있는 여러 종류의 관계들이 포함되어 있고, 고안된 검색방법은 관련도를 사용자의 질의어와 확장된 색인어 사이의 관계의 종류를 추론한 결과와 거리 단계를 고려하여 평가한다.

본 논문의 구성은 다음과 같다. 2 장에서는 지식구조로 시소러스를 이용하는 기존의 정보검색 시스템들에 관해서 살펴본다. 3 장에서는 인간 전문가의 지식구조를 모방한 시소러스 구조에 관해서 자세히 기술한다. 4 장에서는 인간 전문가의 검색 방법을 모방한 검색방법에 관해서 기술한다. 5 장에서는 결론과 앞으로의 연구과제에 관해서 기술한다.

2 관련 연구

전자 도서관, 전자 출판 등의 기본이 되는 대용량 텍스트 데이터베이스의 비정형화된 정보를 검색의 대상으로 하는 정보검색 시스템들의 중요성이 증대되고 있다. 초기의 정보검색 시스템들은 사용자가 필요로 하는 정보를 의미하는 색인어를 얻고 있을 경우에만 정보를 찾아 주었다. 초기의 정보검색 시스템을 이용하는 사용자는 해당분야에 관한 전문지식을 가지고 있거나 해당분야 전문가의 도움을 받아야 했다. 그래서, 해당분야에 관한 전문지식이 없는 사용자가 필요로 하는 정보와 관련된 단어나 문장을 입력할 경우에도 정보를 찾아주는 정보검색 시스템들이 개발되었다. 이러한 정보검색 시스템들의 대부분은 지식구조로 시소러스를 이용하여 사용자의 질의어를 확장해서 사용자의 질의어와 관련도가 높은 색인어들을 찾아낸다. 시소러스는 용어들 사이의 관계를 기술하고 있는 사전이다. 용어들 사이의 관계에는 동의어 관계, 상위어 관계, 하위어 관계, 관련

이 연구는 아주대학교 대학원 연구비 지원(과제번호 971-0901-007-2)을 받아 수행되었음.

어 관계 등이 있다. 개발된 정보검색 시스템들에는 용어들 사이의 개념적인 관계만을 이용하는 정보검색 시스템들[1] [10], 용어들 사이의 이러 종류의 관계들을 영역에 독립된 지식과 영역에 종속된 지식으로 분류한 정보검색 시스템[4] 용어들 사이의 관련도를 수치로 표현한 정보검색 시스템들[6][7] 등이 있다. 이러한 정보검색 시스템들은 사용자에게 만족할 만한 검색결과를 제시하지 못하고 있다. 이것은 개발된 정보검색 시스템들이 이용하고 있는 시소러스 구조가 사람의 지식구조에 저장되어 있는 용어들 사이의 여러 종류의 관계들을 포함하고 있지 않고, 용어들 사이의 관련도를 사람의 관련도 평가 방법과는 다르게 수치로 표현하고 있기 때문이다. 또한, 시소러스를 이용하여 사용자의 질의어와 관련도가 밀접한 색인어들을 검색하는 방법이 사람의 검색방법과는 차이가 있기 때문이다. 인간 전문가의 지식구조를 모방한 시소러스 구조를 설계하고 인간 전문가의 검색방법을 모방한 검색방법을 고안하기 위하여 이러 의미에 관한 인어심리학적인 이론들을 고려하여 영어 단어들을 체계적으로 분류한 워드넷[9]을 참고했다.

3 시소러스 구조

인간 전문가의 지식구조에 저장되어 있는 용어들 사이의 여러 종류의 관계들을 포함하고 있는 지식기반 시소러스 구조는 <그림 1>과 같다.

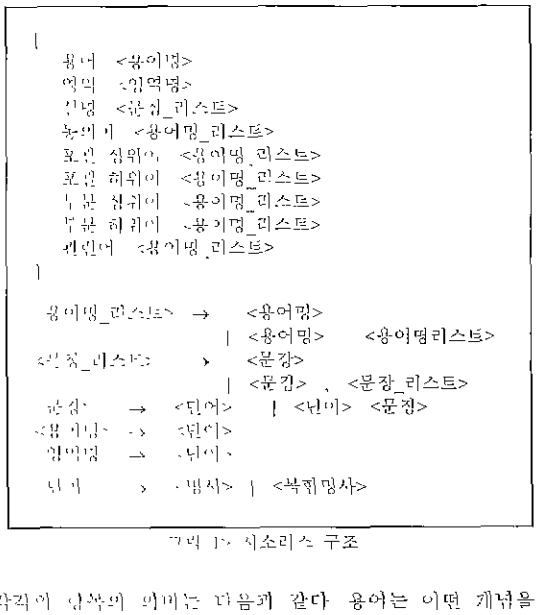

그림 1> 시소러스 구조

각각의 기호의 의미는 다음과 같다. 용어는 어떤 개념을 나타내는 단어와 의미하고, 영역은 용어가 속해 있는 분야를 의미하고, 설명은 용어의 의미를 나타내기 위한 문장들을 의미한다. 동의어는 용어와 같은 의미를 나타내는 용어들을 의미한다. 포괄 상위어는 포괄적인 상위어를 나타내는 속관계에 있는 용어들을 의미하고, 포괄 하위어는 포괄적인 하위어들 나타내는 속관계에 있는 용어들을 의미한다. 부분 상위어는 부분 상위어를 나타내는 용어들을 의미하고, 부분 하위어는 부분 하위어를 나타내는 용어들을 의미한다. 관련어는 관련은 있으나 동의어, 상위어, 하위어 등의 관계에 있지 않은 용어들을 의미한다. '정렬', '내부 정렬' 등의 용어들을 <그림 1>의 시소러스 구조로 나타내면 다음과 같다.

[
용어 정렬
영역 자료구조
설명 컴퓨터 기억장소 내에 저장되어 있는 자료를 키에 따라 원하는 순서로 배열함
동의어 sorting, sort 소트
포괄 상위어 : 자료구조
포괄 하위어 : 내부 정렬, 외부 정렬
부분 상위어
부분 하위어 :
관련어 오름차순, 내림차순
]

[
용어 : 내부 정렬
영역 자료구조
설명 주기억공간 내에서 정렬, 자료량은 적으나 빠르다
동의어 internal sorting
포괄 상위어 정렬
포괄 하위어 삽입 정렬, 버블 정렬, 선택 정렬, 힙 정렬
부분 상위어
부분 하위어
관련어 오름차순, 내림차순
]

4. 검색 방법

인간 전문가가 자신의 지식구조를 이용하여 해당분야에 관한 전문지식이 없는 일반인이 필요로 하는 정보를 의미하는 색인어들을 찾아내는 것을 모방한 검색방법은 다음과 같다. 검색방법은 질의어 확장 모듈, 색인어 추출 모듈, 관련어 평가 모듈 등으로 구성되어 있다.

4.1 질의어 확장 모듈

시소러스를 이용하여 사용자의 질의어를 확장하여 관련도가 밀접한 용어들을 찾아내는 모듈이다. 질의어 확장은 거리 3 단계 까지만 한다. 질의어 확장을 통해서 사용자의 질의어와 확장된 용어 사이의 관계의 종류를 추론하고 거리 등급을 계산한다.

4.1.1 관계의 종류를 추론하는 방법

관계의 종류를 추론하기 위하여 관계의 종류를 동의어 관계, 상위어 관계, 하위어 관계, 관련어 관계 등 4 종류로 분류한다. 포괄 상위어 관계, 부분 상위어 관계 등이 상위어 관계에 포함된다. 포괄 하위어 관계, 부분 하위어 관계 등이 하위어 관계에 포함된다. 사용자의 질의어와 확장되는 용어 사이의 관계의 종류를 추론하는 데에는 사용자의 질의어와 확장된 용어 사이의 추론된 관계, 확장된 용어와 확장되는 용어 사이의 관계 등이 고려된다. 관계의 종류가 4 종류이기 때문에 관계의 종류를 추

론하는 유형은 16 개 이다. 예를 들면, 사용자의 질의어와 확장된 용어 사이의 추론된 관계가 상위어 관계이고 확장된 용어와 확장되는 용어 사이의 관계가 관련어 관계이면, 사용자의 질의어와 확장되는 용어 사이의 관계의 종류는 관련어 관계로 추론된다.

4.1.2 거리 단계 계산

사용자의 질의어와 확장되는 용어 사이의 거리 단계는 시소러스를 그래프로 표현했을 경우에 간선들의 수와 일치한다. 그러나 사용자의 질의어와 확장되는 용어 사이의 관계의 종류를 추론하는 유형들 중에서 상위어 관계 – 하위어 관계와 하위어 관계 – 상위어 관계 등의 2 가지 유형은 거리 단계가 증가하지 않는다.

4.2 색인어 추출 모듈

질의어 확장을 통해서 찾아낸 용어들 중에서 색인어 정보를 저장하고 있는 역파일을 이용하여 색인어들만 추출해 내는 모듈이다.

4.3 관련도 평가 모듈

사용자의 질의어와 추출된 색인어들의 관련도는 추론된 관계의 종류와 거리 단계를 고려하는 <그림 2>의 관련도 평가 알고리즘을 이용하여 평가한다. 중복된 색인어들은 관련도가 가장 높은 것을 제외하고 삭제한다. 그리고 색인어들을 관련도가 높은 순에서 낮은 순으로 정렬한다.

(1) 거리 단계에 관계없이 동의어 관계가 다른 관계들보다 관련도가 높고 상위어 관계와 하위어 관계는 관련도가 같다.
(2) 거리 단계가 같을 경우에는
  a. 상위어 관계와 하위어 관계는 관련도가 같다.
  b. 상위어 관계와 하위어 관계가 관련어 관계보다 관련도가 높다.
(3) 거리 단계가 다를 경우에는
  a. 관계의 종류가 같을 경우에는
    거리 단계가 가장 적은 관련도는 높다.
  b. 관계의 종류가 다를 경우에는
    관련어 관계인 용어의 거리 단계가 상위어 관계 · 하위어 관계인 용어의 거리 단계보다 1 이 작으면 관련도는 같다.
    관련어 관계인 용어의 거리 단계가 상위어 관계 · 하위어 관계인 용어의 거리 단계보다 2 가 적으면 관련도가 높다.

그림 2. 관련도 평가 알고리즘

5. 결론 및 연구과제

본 논문에서는 여러 분야의 인간 전문가가 해당분야에 관한 전문지식이 없는 일반인이 필요로 하는 정보를 찾아주는 방법을 모델링한 지능형 정보검색 시스템을 개발하기 위하여 인간 전문가의 지식구조를 모델링한 시소러스 구조를 설계하였고, 인간 전문가의 검색 방법을 모델링한 검색 방법을 고안하였다. 설계된 시소러스 구조에는 인간 전문가의 지식구조 내에 표현되어 있는 여러 종류의 관계들이 포함되어 있고, 고안된 검색방법은 관련도를 사용자의 질의어와 확장된 색인어 사이의 관계의 종류를 추론한 결과와 거리 단계를 고려하여 평가한다.

앞으로의 연구과제는 지능형 정보검색 시스템의 전체적인 구성과 구성요소들을 설계하는 것이다. 또한, 검색결과의 정확도를 향상시키고 사용자의 특성에 맞는 검색결과를 제시하기 위하여 사용자의 피드백 방법에 관해 연구하는 것이다. 그런 다음, 지능형 정보검색 시스템을 개발해서 문서 데이터베이스들을 이용하여 기존의 정보검색 시스템들과 비교하고 분석하는 것이다.

6 참고문헌

[1] 김영희, "계층적 개념 그래프를 이용한 지식 기반 정보검색 모델", 한국과학기술원 박사학위논문, 1990

[2] 문유진, '의미론적 어휘개념에 기반한 한국어 명사 WordNet의 설계와 구축" 서울대학교 박사학위논문 1996

[3] 박영몽, "지적 정보 검색을 위한 인식론적인 시소러스 시스템의 설계 및 구현", 아주대학교 석사학위 논문, 1994

[4] 신동욱외 3 명, "도메인 독립 및 종속지식을 이용한 효율적 정보검색", 한국정보과학회 논문지 pp. 511-519, 1994. 03

[5] 정영미, '정보검색론", 구미무역 (주) 출판부, 1993

[6] 정재훈, 이상구, '정보 검색을 위한 효율적인 시소러스 구조에 관한 연구" 한국정보과학회 봄 학술발표 논문집, Vol 22, No 1, pp 949-952, 1995

[7] 최재훈, 박종진, 한종진, 임해동, "지능형 정보검색을 위한 객체 기반 시소러스", 한국정보과학회 가을 학술발표 논문집, Vol 22 No 2, pp. 227-230, 1995

[8] Aitchison, J., A. Gilchrist, ' Thesaurus construction - A Practical Manual', London ASLIB, 1972

[9] George A. Miller 외 4 명, "Introduction to WordNet" PRINCETON UNIVERSITY, 1993

[10] R. Rada, H. Mili, E. Bicknell, and M. Blettner "Development and application of a metric on semantic nets", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 19, No. 1, pp. 17-30, January/February 1989

[11] Sung H. Myaeng, Christopher Khoo "Linguistic Processing of Text for a Large-Scale Conceptual Information Retrieval System", Proc. Of 2nd International Conference on Conceptual Structures, pp. 69-83, 1994

[12] William B. Frakes, ' Information Retrieval ' Prentice Hall, 1992