인터넷 웹에서의 특정 분야의 전문 지식 획득[1)]

김상경°*, 박사준*, 김재호**, 김기태*
*중앙대학교 컴퓨터공학과
**국립원주대학 행정전산과

Expertise aquisition of special session in internet web

Sang-Kyong Kim°*, Sa-Joon Park*, Jae-Ho Kim**, Ki-Tae Kim*
*Dept. of Computer Science & Engineering, Chung-Ang University
**Dept. of Administration & Computer Science, Wonju National College

요 약

전문가 검색 엔진은 전문가 시스템과 같은 목적에서 특정 전문 분야에 대한 특수한 정보를 수집 검색하기 위한 검색 엔진을 지칭한다. 특정 전문 분야를 위한 검색 엔진을 제작하기 위해서는 해당 분야만으로 구성된 웹 문서가 필요하다. 본 논문에서 제안한 전문가 검색 엔진은, 특정 분야의 웹 문서만 수집하기 위해서 개념 지식을 사용하여 웹 문서의 특정 분야 귀속 여부를 판단하였다. 개념 지식을 사용하여 웹 문서의 특정 분야 귀속 여부를 판단하기 위해서는, 개념 지식이 특정 분야에 대해 충분히 수집이 되어야 하며, 다른 분야와 충돌하지 않아야 한다. 이러한 개념 지식을 구축하는 것은 사람의 손으로 하는 것은 매우 어려운 일이므로, 본 논문에서는 학습을 통하여 개념 지식을 확장하고, 이를 전문가가 개입하여 학습 과정을 확인하였다. 본 논문은 개념 지식의 학습과 학습의 효용성에 대한 실험 및 결과에 대한 논문이다.

1. 서 론

인터넷에서 원하는 정보를 검색하기 위해서 포탈 사이트의 인터넷 검색엔진을 사용하여 정보를 검색한다. 인터넷은 정보통신 기술의 혁신적인 발전에 힘입어 매우 빠른 속도로 성장하고 있다. 그러나 인터넷 검색 엔진 기술은 인터넷의 발전 속도를 따라잡지 못하고 있다. 그 결과 현재의 검색 엔진으론 원하는 정보를 적절히 얻는데 한계에 이르렀으며, 이러한 문제를 해결하기 위하여 새로운 형태의 검색 엔진들이 계속 제안되고 있다. 새로 제안된 검색 엔진 중에는 전문가 검색 엔진[1][2]이란 방법이 있다.

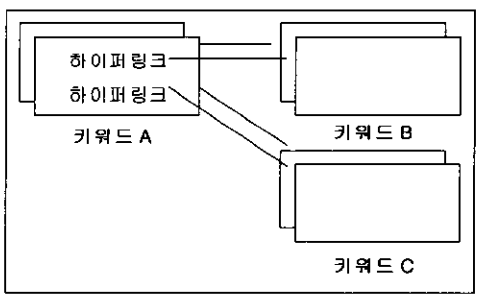
전문가 검색 엔진은 전문가 시스템처럼 특정 전문 분야의 검색 엔진을 제작하기 위하여, 해당 분야의 전문가의 지식을 사용한다. 이러한 전문가의 지식을 개념 지식이라 하는데, 개념 지식을 수작업을 통해 일일이 구축하기에는 상당한 한계가 있다. 그러므로 기초가 되는 개념 지식으로부터 학습을 통하여 개념 지식의 확장이 필요하다. 본 논문은 '전문가 검색 엔진에서 개념 그래프를 이용한 Web 정보 획득'[1]에 기초하여, 웹 문서 수집 로봇[3]을 사용하여 특정 전문 분야의 웹 문서를 수집할 때, 초기 개념 지식과 학습된 개념 지식의 웹 문서 수집 능력을 비교 실험하고 이를 평가한 논문이다.

본 논문에서는 2장에서 관련 연구를 살펴보고, 3장에서 시스템 구성에 대해 설명하며, 4장에서는 개념 지식의 학습에 대한 실험 및 결과를 설명하고, 마지막 5장에서는 결론 및 향후 과제를 제안한다.

1) 본 논문은 2000학년도 중앙대학교 학술연구비 지원에 의한 것임.

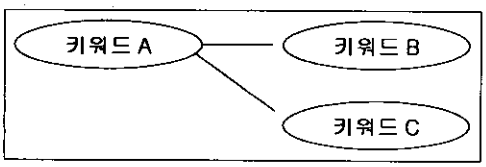
2. 관련 연구

2.1. 하이퍼링크 정보를 이용한 개념 관계 추출 방법



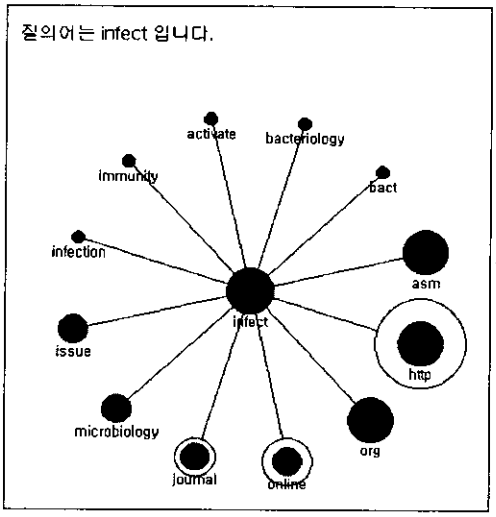
[그림 1] 하이퍼링크로 생성된 웹 문서간의 관계도

웹 문서는 하나 이상의 핵심어를 가지게 된다. 또, 웹 문서는 하이퍼링크를 사용하여 서로 연결되어 있다. 각 문서의 핵심어와 하이퍼링크인 링크를 이용하여 개념을 생성한다. 링크는 계층적 혹은 내용와 참조를 나타내는데, 이를 이용하여 웹 문서의 계층적 구조와 참조가 핵심어간의 계층적 구조와 참조로 추상화되며, 이것이 개념으로 나타난다.[4]



[그림 2] 추상화된 키워드간의 관계도

2.2. 전문가 검색 엔진



[그림 3] 전문가 검색 엔진 실행 화면

전문가 검색 엔진은 전문가 시스템과 같은 목적에서 특정 전문 분야에 대한 특별한 정보를 모아 특정 정보를 검색하기 위한 검색엔진이다. 전문가 검색 엔진은 인터넷 웹 환경에서 사전에 준비된 특정 분야에 대한 개념 지식을 가지고 해당 분야에 대한 웹 문서를 선별적으로 수집한다. 그 후, 수집된 웹 문서의 하이퍼링크 정보를 이용하여 데이터 마이닝을 수행하여 개념 관계를 추출한다. [2][5]

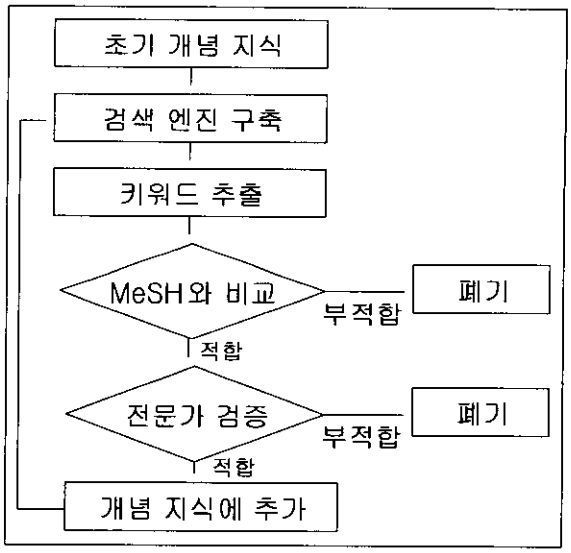
2.3. MeSH

MeSH는 Medical Subject Headings(의학 주제명표목집)의 약자이다. MeSH는 단순한 표목표와는 달리 Subject Tree & Tree Numbers, Exploded Term(Subheadings), Permuted Index로 구성되어 있다. Tree에는 용어의 정의 및 상하개념과 용어 분류가 들어있으며, Exploded에는 용어의 기능(Subheadings)이 들어 있고, Permuted에는 용어의 알파벳순서가 들어있다. 가장 중요한 것은 Permuted Index로서 일종의 전문용어 시소러스(Thesaurus), 알파벳순 주제명전거파일과 같은 역할을 한다. [6][7]

3. 시스템 구성

본 논문에서 제안한 개념 지식 학습 방법은 다음과 같다.

초기의 개념 지식을 사용하여 우선 검색 엔진을 구축한다. 이 과정에서 추출된 키워드를 MeSH와 비교하여 키워드가 개념 지식으로 적합한지 판정한다. 적합하다고 판정되면 전문가가 최종적으로 검증을 한 후에 개념 지식에 추가한다.



[그림 4] 시스템 순서도

3.1. 초기 개념 지식

초기 개념 지식은 MeSH에 등록된 키워드들 중에서 전문가가 대표적으로 사용하는 적절한 키워드만 뽑아내어 구축한다.

3.2. 검색 엔진 구축 및 키워드 추출

검색 엔진 구축은 우선, 웹 로봇인 스파이더(spider) 프로그램을 사용하여 특정 전문 분야에 대한 웹 문서를 수집한다. 그 다음 수집된 문서에서 하이퍼링크를 이용한 문서들 사이의 개념 관계를 정의한 후, 웹 문서에서 키워드를 뽑아내어 키워드간의 개념 관계를 정의한다. 이러한 과정을 거쳐 키워드를 인덱싱(Indexing)한 후, 질의처리기를 사용하여 질의에 응답하게 한다. [2][4]

3.3. MeSH와 비교

검색 엔진 구축 과정에서 추출된 키워드들은 다음 개념 지식에 포함이 될 후보들이 된다. 이러한 키워드 중에는 보편적인 단어와 전문 용어가 섞여 있으므로, 우선 MeSH를 사용하여 전문 용어만 추출한다. 이 때, MeSH에서 주제어 영역과 용어 설명 영역을 추출된 키워드와 비교하여 적합성을 판정한다. 이 과정을 통하여 상당수의 일반 용어를 제외시킬 수 있다.[6]

3.4. 전문가 검증 및 개념 지식 추가

모든 전문 용어가 개념 지식이 될 수 없다. 전문 용어 중에는 상당수의 동음이의어가 포함되어 있어, 전문 용어임과 동시에 일반 용어인 경우가 많다.

ex) Back : 뒤로(범용 지식)
: 등(의학 용어)

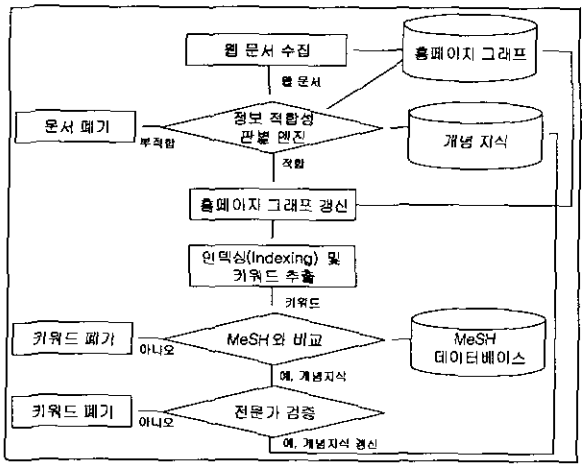
그러므로 선별된 키워드는 정보 검색 시스템과 전문 분야에 대한 지식을 갖춘 전문가의 검증을 거친 후에 기존의 개념 지식에 추가된다.[1]

4. 실험 및 결과

본 논문의 내용을 구현하기 위하여, 의학 분야에 대한 검색 엔진

을 구축하였으며, 그 중 전문 분야 웹 문서 수집에 대한 실험하였다.

4.1. 개념 지식 추출 알고리즘



[그림 5] 개념 지식 추출 알고리즘

웹 문서 수집기를 사용하여 웹 문서를 수집한 후, 개념 지식과 홈페이지 그래프를 사용하여 웹 문서가 수집하고자 하는 전문 분야의 문서인지 판별한다. 적합하다고 판정이 되면 홈페이지 그래프를 갱신한 후, 인덱싱(Indexing)과 키워드를 수집한다. 이 때, 추출된 키워드를 MeSH의 주제어와 용어 해석과 비교하여 용어가 존재하면, 전문가의 검증을 거친 후에 개념 지식에 추가시킨다.

4.2. 실험 결과

실험은 초기 개념 지식을 이용하여 특정 전문분야의 웹 문서를 수집하였다. 이 과정에서 수집된 웹 문서를 가지고 검색 엔진을 제작 후, 추출된 키워드를 앞에서 언급한 개념 지식을 추출 방법을 통하여 추출하였다. 그 결과, 새로 추출된 5,473개의 개념 지식이 추출되었으며, 이를 기존의 개념 지식에 추가시켰다. 그런 다음 다시 웹 문서를 수집하여, 개념 지식의 학습 전과 학습 후의 부적합 판정을 받은 웹 문서를 비교 평가하였다. 부적합 판정을 받은 웹 문서는 수집하고자 하는 전문분야가 아니라고 판정을 받은 문서이다.

부적합 판정을 받은 웹 문서 중 판정이 잘못 내려진 비율은 학습 전이나 학습 후나 크게 다를 바가 없었다. 그러나 부적합 판정을 받은 웹 문서의 개수는 절반 이상 감소하였다. 또, 평가 대상 전체 웹 문서 판정 오류율은 학습 전 84.4%에서 학습 후 90.1%로 약 5.7% 개선되어, 학습 후 90% 이상의 정확도를 보였다.

5. 결론 및 향후 과제.

개념 지식의 학습을 통하여 전문 분야 적합성 판정 정확도가 약간 상승하였다. 개념 지식이 증가하여 부적합 판정을 받은 웹 문서의 개수가 감소하였기 때문에 이러한 결과가 나온 것으로 생각된다. 결과적으론 전체적인 정확도는 상승하였으나, 부적합 판정을 단순히 개념 지식의 존재 유무만으로 판정하기에는 상당한 한계가 있는 것으로 생각된다.

부적합 판정의 정확도를 높이기 위해서는 단순히 개념 지식의 존재 여부뿐만 아니라, TF-IDF와 같은 키워드의 존재 빈도 대비 중요도나 논리 연산, 점수부여(scoring) 방법에 의한 적합성 판정 방법이 요구된다.

[표 1] 논문 프로그램 수행 결과

전문 분야	의학 분야	
수집 시작 주소	http://www.medmark.org/main.html	
개념 지식 학습	학습 전	학습 후
개념 지식 키워드	15,939 개	21,412 개
전송 받은 웹 문서	10,153 개	10,007 개
수집된 웹 문서 주소	106,793 개	115,365 개
부적합 판정을 받은 웹 문서	2,625 개	1,059 개
상위 3,000개의 웹 문서 중 부적합 판정을 받은 웹 문서	698 개	223 개
상위 3,000개의 웹 문서 중 실제로 부적합 웹 문서	230 개	78 개
상위 3,000개의 웹 문서 중 적합성 판정오류가 난 웹 문서	468 개	145 개
부적합 판정을 받은 웹 문서 중 판정이 잘 못 내려진 비율	468 / 698 = 67%	145 / 223 = 65%
평가 대상 전체 웹 문서 판정 오류	84.4%	90.1%

[표 2] 전문 분야 웹 문서 판정 오류율 계산 공식

학습 전:
(평가 대상 - 적합성 판정 오류) / 평가 대상
= (3000 - 468) / 3000 = 84.4%
학습 후:
(평가 대상 - 적합성 판정 오류 - 부적합 문서 변동)
/ 평가 대상 = (3000 - 145 - (230 - 78)) / 3000 = 90.1%

참고문헌

[1] 박사준, 김상경, 황수철, 김기태, 전문가 검색 엔진에서 개념 그래프를 이용한 Web 정보 확득, 2000년 봄 학술발표논문집(B) 제 27권 1호 pp295-297, 한국정보과학회

[2] 박사준, 김상경, 최준영, 김기태, 인터넷 웹 환경에서의 전문가 검색 엔진 연구, 한국인터넷정보학회

[3] The Web Robots Pages,"http://info.webcrawler.com/mak/projects/robots/robots.html"

[4] 최준영, 인터넷상의 하이퍼링크를 이용한 개념 그래프 기반 검색 시스템. 1998,12.

[5] 이권국, 신일수, 이상준, 김기태, 전문가 검색 엔진에서 데이터 마이닝을 이용한 개념 관계 추출, 2000년 봄 학술발표논문집(B) 제 27권 1호 pp298-300, 한국정보과학회

[6] Henry J. Lowe and G. Octo Barnett. Understanding and using the medical subject headings(MeSH) vocabulary to perform literature searches. Journal of the American Medical Association, 271(4):1103-1108, 1994

[7] 김종은, http://delias.dongeui.ac.kr/mailing/messages/data/973.html