전문가 검색 엔진에서 개념 그래프를 이용한 Web 정보 획득

박사준*, 김상경*, 황수철**, 김기태*
*중앙대학교 컴퓨터공학과
**인하공업전문대학

Information acquision of WEB using the conceptual graph in expert search engine

Sa-Joon Park *, Sang-Kyong Kim*, Su-Chul Hwang**, Ki-Tae Kim*
*Dept. of Computer Science & Engineering, Chung-Ang University
**Inha Technical Junior College

요 약

전문가 검색 엔진은 전문가 시스템과 같은 목적에서 특정 전문 분야에 대한 특수한 정보를 수집 검색하기 위한 검색 엔진을 내용이다. 본 논문은 전문가 검색 엔진을 만드는 과정에서 초기 작업에 해당하는 웹 정보 수집에 대한 논문이다. 본 논문은 웹 페이지에서 하이퍼링크와 참조되는 웹 페이지에 대한 표면 지식을 이용하여, 홈페이지 그래프를 작성한다. 그리고 나서 홈페이지 그래프와 사전에 미리 준비된 개념 그래프를 이용하여, 웹 페이지 수집 중 특정 전문 분야에 해당하는 웹 페이지인지를 판별하여 사용자가 수집하고자 하는 분야에 대한 웹 페이지만을 수집한다. 본 논문은 이에 대한 개념, 실제 및 구현과 앞으로의 개선 상황을 제안한다.

1. 서 론

정보 혁명기를 맞이하여 인터넷은 엄청난 속도로 거대해 지고 있다. 특히, 월드 와이드 웹은 매일 2억 페이지 이상 증가하고 있다.[5] 그런데 현재의 검색 엔진은 인터넷상에 존재하는 웹의 약 20% 미만의 정보만을 가지고 있다. 이로 인하여, 사용자가 검색 엔진을 통하여 원하는 정보를 얻고자 하여도 세대로 정보를 얻을 수 없다. 이러한 문제를 해결하기 위해 정보를 특정 영역이나 관심 분야로 분류(classification)하고 이에 맞게 요약하고 정형화 데이터 베이스로 만드는 추출(extraction)함으로써 이러한 문제를 해결하고 있다.[4]

인터넷상에는 여러 분야에 대한 전문 웹사이트가 존재한다. 그러나 아직 이런 전문 분야의 웹사이트에 대한 전용 검색 엔진이 없는 형편이다. 본 논문은 전문가 시스템과 같은 취지의 전문가 검색 엔진을 만들려고 한다. 여기서 말하는 전문가 검색 엔진이란, 모든 영역을 목표로 검색 서비스를 제공하는 것이 아니라, 특정 전문 영역에 대한 전문적인 검색 엔진을 말한다. 본 논문은 '인터넷상의 하이퍼링크를 이용한 개념 그래프 기반 검색 시스템'[3]을 기반으로 한 전문가 검색 엔진 구현의 첫 단계인 웹 정보 수집단계에 관한 논문이다.

본 논문에서는 2장에서 관련 연구를 살펴보고, 3장에서 시스템 구성을 설명하고, 4장에서 개념 그래프와 홈페이지 그래프를 이용한 적합성 판정에 대해 설명하며, 5장에서는 본 논문의 구현 결과를 제시와 추후 개선 방향을 제안한다.

2. 관련 연구

본 논문에서 구현한 프로그램은 이미 Web Wanderers, Crawlers, Spiders라는 이름으로 많은 연구가 되어 왔다.[8]

2.1. 정보의 분류

정보 분류는 해당 정보 미리 정해져 있는 어떤 부류에 속하는지를 여부를 판단하는 것이다. 또, 분석이나 추출, 데이터 마이닝 등의 전단계로, 이들 작업에 들어가기 위해 컴퓨터가 이해하기 쉬운 형태로 정보를 구성하는 역할로도 사용한다.[6]

웹 문서 분류에는 (1) 통계적(statistical) 방법과 (2) 제 1 순위 논리(First-Order logic) 표현법이 있으며, (3) URL 휴리스틱(heuristics)에 의한 방법이 있다.[6]

2.2. 로봇의 사용 용도와 Robots Exclusion의 필요성

로봇은 검색 엔진들의 정보의 보유 측면에서 본다면 대략 두 가지 형태로 나누어 볼 수가 있다. 하나는, 로봇 에이전트(Robot agents)를 통해 자료를 수집하거나 사용자가 정보를 등록하게 함으로써 자기 자신의 데이터베이스를 구축하고 있는 경우이고, 다른 하나는, 다른 검색 엔진에서 보유하고 있는 데이터베이스를 이용하여 사용자에게 서비스를 하는 형태이다. 이런 로봇들은 여러 가지 이유로 웹 서

버를 혼란스럽게 하거나 적당하지 않은 웹 서버의 부분들을 검색하거나 일시적인 정보 또는 부작용(side-effect)을 가질 수 있는 cgi-script를 검색하는 일도 발생해 왔다. 이런 이유로 Robots Exclusion이 필요하게 되었다.[1]

2.3. Robots Exclusion

2.2.1. The Robots Exclusion Protocol
웹사이트 관리자가 로봇은 방문하지 말라고 프로토콜에 미리 표시하는 방법이다.[8]

2.2.2. The Robots META tag
웹 문서 작성자가 HTML의 META Tag에 해당 웹 문서에 대한 정보와 검색 엔진에 인덱싱되어야 할지 말아야 할지를 정의해 놓는 방법이다.[8]

3. 시스템 구성도

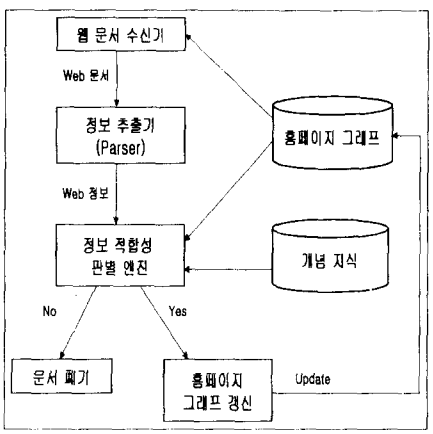

그림 1 전체 시스템 구성도

본 시스템은 웹 문서 수신기, 정보 추출기, 정보 적합성 판별 엔진에 해당하는 3개의 실행 모듈과 개념 그래프와 홈페이지 그래프에 해당하는 2개의 데이터베이스로 구성된다.

3.1. 웹 문서 수신기.

홈페이지 그래프를 참조하여 전송 받아야 할 웹 문서를 할당받고, 웹 문서를 실제로 전송 받는 모듈이다. 웹 문서 수신기는 웹 문서를 상대방 사이트로부터 전송 받아 일체의 가공 과정을 거치지 않고, 정보 추출기로 웹 문서를 넘긴다.

3.2. 정보 추출기.

웹 문서의 HTML을 파스(Parse)하여 웹 문서에서 필요한 정보를 얻는다. 본 시스템에서는 하이퍼링크와 하이퍼링크의 제목, 웹 문서 제목, 본문 내용 등의 4가지 정보를 추출한다.
추출된 정보 중 하이퍼링크와 하이퍼링크 제목은 정보 적합성 판별 엔진이 해당 웹 문서를 적합성 판정하면 홈페이지 그래프를 갱신하는데 사용된다.
추출된 정보 중 웹 문서 제목과 본문 내용은 정보 적합성 판별 엔진에서 해당 웹 문서의 적합성 판정 여부를 위해 사용된다.

3.3. 정보 적합성 판별 엔진.

정보 추출기로부터 넘겨받은 제목과 본문 내용을 가지고 해당 웹 문서의 적합성 여부를 판정한다. 이 때, 이미 구성되어 있는 개념 그래프와 홈페이지 그래프를 이용하여 정보의 적합성을 판정한다. 자세한 알고리즘은 4장이 기술한다.

3.4. 시스템의 작동 원리

표 1은 시스템의 작동 원리를 나타내며, 그림 2는 시스템의 작동 알고리즘을 나타낸다.

표 1. 작동 원리

1. 홈페이지 그래프로부터 전송 받을 웹 문서를 할당받아 문서를 전송 받는다.
2. 웹 문서에서 정보를 추출한다.
3. 추출된 정보를 가지고 정보의 적합성을 판정한다.
3.1. 적합 판정을 받으면, 홈페이지 그래프를 갱신한다.
3.2. 부적합 판정을 받으면, 해당 웹 문서에서 수집된 정보를 폐기하고, 해당 웹 문서가 전송 보류되었음을 홈페이지 그래프에 기록한다.

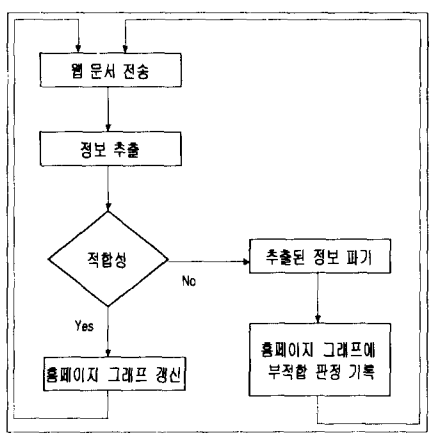

그림 2. 작동 알고리즘

4. 적합성 판정 알고리즘.

본 논문에서는 2.1.에서 설명한 정보의 분류 방법 중 통계적 방법을 이용하여 웹 문서의 적합성을 판정하였다.
본 논문이 구현한 프로그램에서 웹 문서의 적합성을 판정하기 위해서는 홈페이지 그래프와 사전에 미리 준비된 개념 그래프가 필요하다. 이 때, 사전에 미리 준비된 개념 그래프는 수집하려는 해당 전문 분야에 대한 전문적인 지식이다. 만약, 전문적인 지식과 범용 지식이

서로 중복이 되는 경우(즉, 다른 개념이 동일하게 표현되는 경우)에는 개념 그래프에서 제외되었다.

ex) Back : 뒤로 (범용 지식)
: 등 (의학 용어)

실제로 Back이라는 단어는 웹 문서상에 등(의학 용어)이라는 뜻보다는 뒤로(범용 지식)라는 뜻으로 훨씬 많이 쓰인다. 만약, Back이 개념 그래프에 포함되어 있을 경우, 의료 웹 문서가 아니라도, 의료 웹 문서로 간주 될 수 있다. 그러므로, 개념 그래프는 해당 전문분야에 속한 지식으로 아주 전문적인 개념으로 구성되어야 한다. 본 논문에서 사용한 알고리즘은 표2와 같다.

표 2. 적합성 판정 알고리즘

1. 전송 받은 웹 문서가 시작 URL(Seed Site)에서 거리가 몇인지 홈페이지 그래프에서 확인한다.
2-1. 거리가 일정 거리 미만이면 무조건 적합으로 판정한다.
2-2. 거리가 일정 거리 이상이면 추출된 정보와 사전에 준비된 개념 그래프를 서로 비교하여 서로 매치(Match)되는 정보의 개수를 통계를 낸다.
3. 2-2과정에서 매치(Match)되는 정보의 개수가 일정 개수 이상이면 적합으로 판정하고, 매치(Match)되는 정보의 개수가 일정 개수 미만이면 부적합으로 판정한다.

표 2의 알고리즘을 사용하기 위해서는 홈페이지 그래프를 사용해야 한다. 홈페이지 그래프는 시작 URL(Seed Site)로부터 계속 확장해 나간다. 시작 URL(Seed Site)은 수집하고자 하는 해당 전문 분야의 웹 문서로, 해당 전문 분야에 대한 많은 하이퍼링크를 가지고 있어야 한다.

시작 URL(Seed Site)에서부터 일정 거리 안에 있는 모든 웹 문서를 무조건 적합으로 판정하는 이유는 시작 URL이 해당 분야의 전문 사이트만으로 이루어져 있으며, 일정 거리 안에서는 모두 해당 분야에 대한 전문적인 웹 문서일 가능성이 매우 높기 때문이다. 만약에 그렇지 않을 경우 시작 URL로는 적합하지 않으므로, 그런 경우 시작 URL을 다른 웹 문서로 설정하여야 한다.

개념 그래프와 웹 문서로부터 추출된 정보의 매치(Match)는 문자열을 비교한다.

5. 결론 및 향후 과제.

본 논문의 내용을 구현하여, 의료 분야를 전문적으로 다루는 웹 문서에 적용했다.

개념 그래프는 이미 존재하는 의료 키워드를 사용하였으며, 사람이 의료 키워드에서 일부 범용 지식을 수작업했다. 시작 URL(Seed Site)은 "http://www.medmark.org/main.html"로 하였다.

약 8만 2천여 개의 웹 문서를 수집하였으며, 홈페이지 그래프에 등록된 수집된 웹 문서 URL은 약 32만 8천여 개였다. 8만 번째 웹 문서까지 수집한 결과 Seed Site로부터 거리가 4정도인 웹 문서까지 수집하였다.

전송 받는 8만 2천여 개의 웹 문서 중 약 30%정도인 2만 5천여 개의 문서가 부적합 판정을 받았다.

부적합 판정을 받은 웹 문서 중 앞의 3천 개를 확인해 본 결과 절반 정도의 웹 문서는 의료 분야와 상관이 없는 필요 없는 웹 문서였으나, 나머지 절반은 의료 분야와 관련이 없는 문서였다. 즉, 실제로 8만 2천여 개의 웹 문서 중 의료와 상관이 없는 문서는 1만 2천여 개이며, 이를 뺀 나머지 7만개는 의료와 관련이 있는 웹 문서이다. 이를 의료 문서만 놓고 보면, 7만개의 의료 관련 웹 문서 중 약 1만 3천 개의 웹 문서가 판정 보류를 받은 것이므로, 약 16% 정도의 적합성 판정 오류를 나타냈다.

표 3. 논문 프로그램 수행 결과

모으고자 한 전문 분야	의료 분야
홈페이지 그래프의 시작 주소	http://www.medmark.org/main.html
전송 받은 웹 문서	82,860 개
홈페이지 그래프에 등록된 웹 문서 (전송 받았거나 앞으로 받아야 할 웹 문서)	328,740 개
부적합 판정을 받은 웹 문서 개수	25,002 개
부적합 판정을 받는 웹 문서 중 판정이 잘못 내려진 비율	약 50 %
실제로 부적합 판정에 해당하는 웹 문서	약 1만 2천 개
전송 받은 웹 문서 중 의료 전문 웹 문서	약 7만 개
전체 문서 중 판정 오류가 난 웹 문서	25000 - 12000 = 13000 개
의료 문서 부적합 판정율	16 % (13000/82000)

본 논문은 개념 그래프를 컴퓨터가 자동으로 만드는 것이 아니라, 사람이 전문 분야에 대한 개념 그래프를 수동적으로 만드는 것이다. 이런 작업은 해당 분야에 대한 전문가이면서, 컴퓨터가 사용할 수 있는 적합한 형태로 만들 수 있는 지식 공학자가 만들어야 하므로 그리 쉽게 할 수 있는 작업이 아니다. 그러므로, 앞으로는 개념 그래프를 간단히 수작업으로 구축한 후, 개념 그래프를 컴퓨터가 학습을 통하여 자동으로 갱신하도록 개선해야 한다.

또, 부적합 판정을 받은 웹 문서중 상당수는 실제로는 해당 분야에 관련이 있는 웹 문서들이었다. 이런 웹 문서들에 대한 판정 오류를 줄이기 위해 본문의 텍스트 내용뿐만 아니라, 그림 인식이나 URL등 추정에 의한 웹 문서 판정 능력도 필요하다고 생각한다.

참고문헌

[1] 김홍주, Robot agents and Search Engine,"http://solgeo.dongguk.ac.kr/~k2/html/TS3-4.html"

[2] 조민재, 웹의 개념 지식을 이용한 자동 시소러스 생성법의 설계 및 구현. 1999.12.

[3] 최준영, 인터넷상의 하이퍼링크를 이용한 개념 그래프 기반 검색 시스템. 1998.12.

[4] Clare Cardie. Empirical Method in Information Extraction. AAAI Magazine. Winter. 1997

[5] Dayne Fritag. Information Extracting from HTML : Application of a General Machine Learning Approach. Information Extraction. AAAI. 1998

[6] Mark Craven, etc. Learning to Extract Symbolic Knowledge from the World Wide Web. Information Extraction. AAAI. 1998

[7] Quilan, J.R. Learning logical definitions from relations. Machine Learning 5(3): 239-266. 1990

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