온톨로지 기반 가중치 부여 논문 검색 모델

박현철
서울통신기술㈜

Weighting Assignments Paper Retrieval Model Based On Ontology

HyunChul Park
SEOUL COMMTECH CO.,

요 약

많은 연구원들이 자신의 연구 과제를 수행함에 있어 선행 연구 자료로 참고하는 것이 관련 주제에 관한 학술 자료이다. 현재 많은 학교와 기관 그리고 단체에서 관련 학술 자료를 발간하고 있으며 이를 참조하는 방식도 다양하다. 그러나 학술 자료를 참조함에 있어 단어 기반 검색이 사용 되고, 발간된 자료의 양이 방대해짐에 따라 사용자가 원하는 정보를 참조하는 데 많은 어려움이 따른다. 본 논문은 이러한 기존 학술 자료 검색 방법을 보완하기 위하여 온톨로지를 기반으로 하는 가중치 부여 논문 검색 모델을 제안한다. 제안한 모델은 논문 관련 정보를 온톨로지로 구축하고, 검색 문서에 가중치를 부여하는 순위화 알고리즘을 적용한 것이다. 이는 기존 유사도 적용 기법에 시멘틱 개념을 적용한 것으로 효율적이고 정확한 논문 검색을 보장한다.

1. 서 론

시멘틱 웹(Semantic Web)은 정보의 의미를 개념으로 정의 하고 개념간의 관계성을 명시화하여 웹 문서에 의미 정보를 덧붙이고, 에이전트가 이 의미정보를 자동으로 검색하여 정보를 제공하는 것을 의미한다[1]. 이러한 개념은 현재의 정보 검색 시스템이 단어의 빈도수나 링크 정보를 바탕으로 문서의 유사도를 측정하고, 순위를 부여하는 방식을 사용하고 있기에 사용자의 검색 욕구를 만족시키지 못한 것에서 기인한다. 따라서 본 논문에서는 기존의 문서 유사도 측정 방식에 시멘틱 웹 기반의 서비스 제공을 정형화하는 온톨로지[2]를 이용하여 사용자에게 의미 있는 검색 모델을 제공한다.

본 논문에서 제안하는 검색 모델은 논문 분야를 온톨로지로 구축한 후, 사용자 입력을 통해 문서의 유사도를 측정하고, 측정된 유사도를 바탕으로 온톨로지 검색을 수행한다. 이는 기존의 유사도 적용 방법에 온톨로지 검색의 시멘틱 기법을 적용한 것으로 사용자는 보다 의미 있는 정보 검색 결과를 보장 받을 수 있다.

본 논문의 구성은 다음과 같다. 제 2장에서는 논문 검색에 사용되는 기존 시스템을 분석하고, 제 3장에서는 온톨로지 구축과 순위 측정 알고리즘을 기술하고, 제 4장에서는 본 연구에서 제안하는 정보 검색 모델을 기술한다. 마지막으로 제 5장에서는 결론 및 향후 연구 과제를 언급한다.

2. 기존 시스템 분석

본 장에서는 학술 정보를 검색함에 있어 대표적으로 이용되는 국회 전자 도서관과 KSI 학술 논문 정보 시스템을 분석한다.

2.1 국회 전자 도서관

(그림 1) 국회 전자 도서관 메인 화면

국회 전자 도서관[3]은 일반 검색, 상세 검색, 고급 검색의 3가지 서비스를 제공한다. 일반 검색은 전체, 단행본, 학위논문, 학술지 중 검색 대상 자료를 선정하여 자료를 구분하고 사용자가 검색 항목을 선택하여 검색어를 입력하는 방식이다. 상세 검색은 일반 검색과 마찬가지로 사용자가 검색 대상 자료를 선택하여 자료를 구

분하지만 검색 항목을 설정하는 부분에서 항목별로 절단 검색과 불리언 연산을 사용할 수 있다. 고급 검색은 일반 검색, 상세 검색의 빈칸 채우기 방식과는 달리 직접 검색 질의어를 명령어 형식으로 입력하여 검색하는 방식이다. 따라서 국회 전자 도서관에서 제공하는 검색 서비스는 사용자가 자료를 구분하고 항목을 선택하는 등 효율적인 검색이 이루어질 수는 있으나 단순한 키워드 기반 검색이기 때문에 사용자의 의도와는 전혀 무관한 정보를 제공해 주는 경우가 발생한다.

2.2 KSI 학술 논문 정보 시스템

(그림 2) KSI 학술 논문 정보 시스템

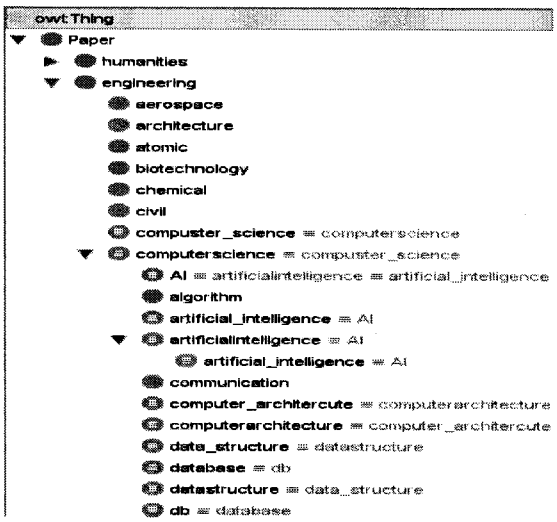
KSI 학술 논문 정보 시스템[4]은 빠른 검색, 상세 검색, 간행물 검색, 분야별 검색 서비스를 제공하고 있다. 빠른 검색은 사용자가 분류 항목을 설정한 후 검색어를 입력하는 방식이고, 상세 검색은 검색 항목을 다양하게 설정하여 사용자가 항목별로 절단 검색과 불리언 연산을 사용할 수 있는 방식이다. 간행물 검색은 발행 기관과 간행물 이름으로 관련 학회나 기관을 검색한 후, 해당 기관에서 발행한 간행물의 내용을 재 검색하는 방식으로 이루어진다. 분야별 검색은 사용자가 원하는 학회를 쉽게 찾을 수 있도록 총 9개의 분야로 나누어 놓았으며 그 하위에 관련 분야별로 다시 분류해 놓았다. 검색을 원하는 분야를 선택한 후 검색어를 입력하면 '제목'에 검색어가 포함되어 있는 모든 원문을 검색할 수 있다. KSI 학술 논문 정보 시스템은 사용자의 편의를 위해 단어 기반 검색 서비스와 분류 시스템을 제공하고 있지만, 의미 기반이 아닌 검색 범위를 제한 하는 방식을 제공하고 있기 때문에 사용자가 원하는 결과를 찾기 위해서는 여러 번 검색을 시도해야 하는 단점이 있다.

3. 온톨로지 기반 가중치 부여 알고리즘

3 장에서는 논문 검색 모델에서 사용되는 온톨로지를 구축하고 본 연구에서 제안하는 가중치 부여 알고리즘을 제안한다.

3.1 온톨로지 구축

온톨로지는 W3C에서 정의한 시멘틱 웹 온톨로지 기술 표준 언어인 OWL(Web Ontology Language)로 기술된다. OWL과 같은 표준언어로 기술함으로써 구축된 온톨로지의 활용성을 극대화 한다. 논문 온톨로지 구축은 학술 진흥 재단의 연구 분야 분류표를 바탕으로 구축한다. 연구 분야 분류표에서와 같이 대분류, 중분류, 소분류를 통해 논문이 포함되는 분야를 구분하였으며 대분류를 구성한 후 중분류에서는 공학 분야를 선택하여 이를 확장 구성하였다. 347개의 클래스로 구성되고 최상위 Thing클래스부터 논문의 정보를 담고 있는 Title클래스까지 총 5개의 계층으로 분류된다. Protégé 3.2.1[5]을 사용하여 구축하였으며 최상위 클래스인 owl:Thing 클래스는 Paper클래스를 subClass로 가지고 Paper 클래스는 공학, 자연과학, 사회과학, 인문학, 복합학, 의약학을 subClass로 가진다. 해당 클래스는 학술 진흥 재단 분류를 바탕으로 다시 세분화 된다. 또한 사용자 질의를 바탕으로 검색을 수행할 경우에 유사어를 처리 하기 위해 클래스 속성에 유사어를 적용하였다.



(그림 3) 논문 온톨로지

3.2 가중치 부여 알고리즘

본 연구에서 제안하는 가중치 부여 반영치는 [정의 1]과 같이 구성된다. 문서의 가중치는 사용자 입력 키워드와 문서의 유사도를 나타내는 코사인 유사도 값과 사용자 입력 카테고리와 연관 관계가 있는 온톨로지 상위 클래스의 수직 노드 거리에 따른 근접도로 구성된다. 코사인 유사도는 사용자가 입력한 검색어를 바탕으로 각 논문에 대한 제목, 키워드, 초록에 대한 유사도를 계

산한 값이다. 이는 벡터 공간 모델에서 사용되는 코사인 유사계수[7] 방법을 사용한 것으로 색인어에 의해 정의되는 벡터 공간에 문서와 질의어를 벡터 공간상에 한 점으로 취급하고 문서간의 코사인 값을 이용하여 그 각이 작을수록 유사도가 높다는 점을 이용한 것이다. 수직 근접도는 사용자가 입력한 카테고리 입력 값이 존재할 경우, 카테고리 검색어를 기반으로 온톨로지 검색을 수행한다. (그림 4)와 같이 코사인 유사도를 적용한 문서 각각이 포함하는 상위클래스를 탐색하고, 사용자가 입력한 카테고리의 키워드와 매칭되는 상위 클래스가 존재할 경우 상위 클래스 노드와의 수직 노드 거리를 계산하여 가중치를 부여한다. 다수의 매칭 클래스가 존재할 경우 이를 재귀적으로 반영하여 각 문서에 대한 수직 근접도를 계산한다.

[정의 1] 가중치 부여 반영치

$$\mathrm{Sim}(d_j,q) = P_j * \frac{\vec{d_j} * \vec{q}}{|\vec{d_j}| * |\vec{q}|} = P_j * T_j$$

$$P_j = \frac{1}{F^p}$$

P_j : 온톨로지 노드간의 수직 근접도

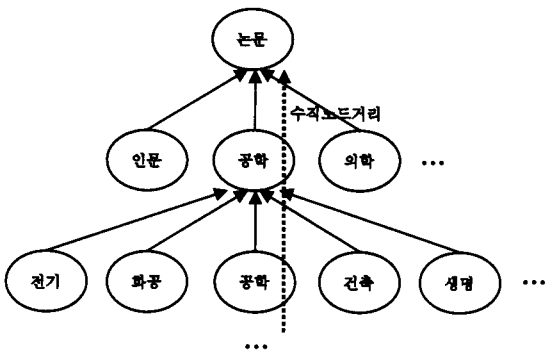
F : 수직 근접도 결정 인자

p : 문서 클래스와 온톨로지 클래스간 수직 노드 거리

T_j : 코사인 유사도

$\vec{d_j}$: 문헌 벡터

$\vec{q}$: 질의 벡터

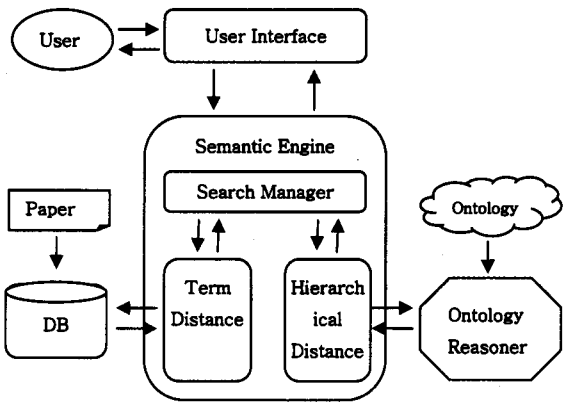


(그림 4) 온톨로지 수직·노드 거리

4. 논문 검색 모델

4.1 시스템 구성도

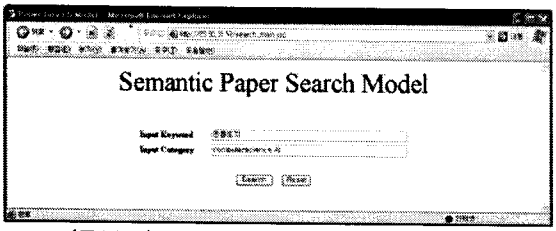
전체적인 시스템 구조는 (그림 5)에서 보는 바와 같이 User Interface Module과 Semantic Engine 그리고 논문 정보를 저장하는 DB와 Ontology를 포함하는 Ontology Reasoner로 구성된다. User Interface Module은 사용자의 입력을 받아 Semantic Engine으로 전송하는 역할을 수행하며 또한 Semantic Engine으로부터 순위화된 결과를 전송 받아 사용자에게 제공한다. Semantic Engine은 Search Manager Module, Term Distance Module 그리고 Hierarchical Distance Module로 이루어진다. Search Manager Module은 Term Distance Module과 Hierarchical Distance Module로부터 수신한 정보를 이용하여 검색된 문서의 최종 순위를 결정하는 Module이다. Term distance Module은 User Interface로부터 전송받은 키워드를 기반으로 DB에 저장된 논문 정보의 유사도를 계산하는 Module이다. Term Distance Module에 의해 검색된 문서 리스트는 Serarch Manager Module에게 전송되고 이는 수직 근접도를 계산하기 위한 검색 대상 리스트가 된다. Hierarchical Distance Module은 Search Manager Module로부터 전송 받은 리스트를 대상으로 상위 클래스를 탐색하기 위해 Ontology Reasoner에게 상위클래스 리스트를 요청한다. 리스트를 전송 받은 Hierarchical Distance Module은 사용자가 입력한 카테고리 키워드와 Ontology Reasoner를 통해 추출된 상위 클래스 리스트를 비교하여 일치하는 클래스에 대하여 수직 근접도를 계산한다. 계산된 수직 근접도는 Search Manager Module로 전송되어 유사도와 함께 문서 순위화에 사용된다. DB는 논문에 대한 정보와 문서 자체를 포함하고 있는 저장 장치이며 각 논문에 대한 제목, 저자, 논문 등록 날짜, 초록, 키워드, 발행처 등의 정보가 저장된다. Ontology Reasoner는 Protégé 3.2.1[5]에 의해 구축된 온톨로지가 로드된 추론기로 유사도 검색 결과 리스트에 포함된 문서의 상위 클래스를 탐색하는 Module이다.



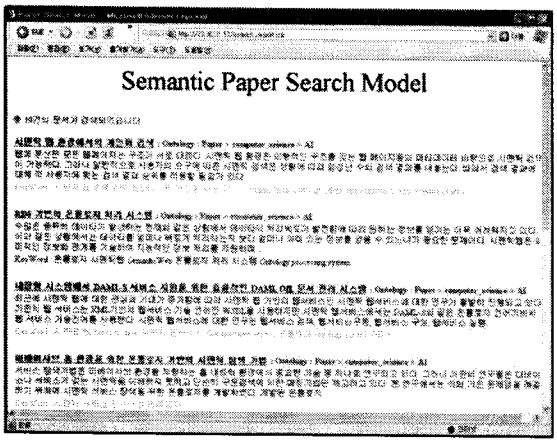
(그림 5) 시스템 구성도

4.2 실행 결과

(그림 6)과 (그림 7)은 사용자 입력화면과 결과 화면이다. (그림 6)과 같이 사용자는 Input Keyword란에 자신이 찾고자 하는 논문의 제목, 키워드, 초록에 포함되는 키워드를 입력한다. Input Category란에는 해당 논문이 포함된 분야를 입력한다. 예를 들어 온톨로지를 키워드로 가지며 해당 분야가 컴퓨터 혹은 컴퓨터 공학 그리고 인공지능에 관련된 분야를 검색할 경우에 키워드는 온톨로지 그리고 관련 분야는 컴퓨터공학 혹은 인공지능을 입력한다. (그림 7)은 사용자가 검색어를 입력하고 이를 Semantic Engine에서 처리하여 순위화 한 결과 화면이다. 사용자는 논문 제목 리스트를 통해 논문 파일을 열람할 수 있으며 해당 논문이 어떤 영역에 속하는지 온톨로지 계층 구조를 파악할 수 있다. 출력되는 화면은 논문의 제목, 계층 구조, 초록 그리고 키워드가 제시된다.



(그림 6) 사용자 인터페이스 화면(초기화면)



(그림 7) 사용자 인터페이스 화면(검색 결과 화면)

5. 결론 및 향후 연구 과제

본 논문에서는 논문 검색의 효율성을 증대시키기 위해 논문 검색 모델을 제안하였다. 기존의 검색 모델이 가지고 있는 문제점인 유사도 기반 검색, 분류를 통한 검색의 문제점을 해결하기 위해 시멘틱 웹 요소를 이용한 순위화 모델을 통해 사용자의 의도를 반영한 검색기법을 제안하였다. 사용자는 자신이 찾고자 하는 논문에 대한 기본적인 분류 지식만 가지고 있으면 해당 논문을 보다 쉽고 정확하게 검색할 수 있으며, 이는 연구자의 선행 연구에 많은 도움이 된다.

향후 연구 과제로는 본 논문에서 구성된 온톨로지에 대한 체계적이고 지속적인 관리가 필요하며 컴퓨터가 처리하는 비중을 높여 좀 더 지능화된 시스템의 설계가 필요하다. 아울러 유사도와 수직 근접도를 적용시킨 성능 측정 알고리즘을 개발하여 이를 이용한 실험 및 분석이 시도되어야 할 것이다.

참고문헌

[1] T.Berners-Lee, J.Hendler, O.Lassila "The Semantic Web," Scientific America, 2001.

[2] A. Gomez-Perez, O. Corcho, "Ontology Languages for the Semantic Web," IEEE Intelligent Systems, Vol.17, No.1, 2002.

[3] 국회전자도서관 Homepage.
http://www.dlibrary.go.kr/

[4] 한국논문정보시스템 Homepage.
http://www.papersearch.net/

[5] Protégé Homepage.
http://protege.stanford.edu/

[6] RacerPro Homepage,
http://www.racer-systems.com/

[7] M. W. Berry, Z. Dramac, and E. R. Jessup, Matrix, Vetor Space, and Information Retrieval, SIAM review, Vol. 41, No. 2, 335~362

[8] Lei Li and Ian Horrocks, "A software Framework For Matchmaking Based on Semantic Web technology", In Proceedings International WWW Conference, Budapest, Hungary. 2003.

[9] Decker, S.,; Melnik, S.: van Harmelen, F.,: Fensel, D.; Klein, M.,: Broekstra, J.: Erdmann, M.,: Horrocks, I., "The Semantic Web:the roles of XML and RDF", IEEE Internet Computing, Vol.4 Issue 5, pp.63~73, Sep.,-Oct., 2000.

[10] 최옥경, 한상용, "자동화된 통합 프레임워크를 위한 시맨틱 웹 기반의 정보 검색 시스템", 한국정보처리학회논문지 제 13-C권 제1호, pp.129~136, 2006. 2. 28.