다중 개체 중심적 통합 방식의 버티컬 검색

- 학술 연구 정보 분석 서비스에의 적용 사례를 중심으로 -

Vertical Search Based on Multiple Entity-centric Unification

정한민, Hanmin Jung*, 이미경, Mi-Kyoung Lee**, 성원경, Won-Kyung Sung***, 류범종, Beom-Jong You****

요약 본 논문은 기존의 단일 분야를 대상으로 서비스되고 있는 버티컬 검색의 한계를 지적하고 사용자의 검색 욕구를 보다 충실히 만족시키기 위해, 여러 분야 (개체 유형)들을 포함하는 실체 (개체)들이나 단일 분야 내의 실체들을 포함하는 질의어를 처리할 수 있는 다중 개체 중심적 통합 방식의 버티컬 검색을 제시한다. 이를 위해, 질의어를 분석하여 개체 유형 간 결합이 필요한 지를 판단한 후 동적으로 상황에 맞는 서비스 컴포넌트들을 결합하는 기술과 개체 유형 별 필드들을 구축하고 필드 별 검색을 수행하는 기술을 도입하였다. 버티컬 검색 서비스 분야로서 학술 연구 정보를 대상으로 하여 약 453,000 편의 해외 학술 저널 논문을 메타데이터 기반으로 등록하였으며, 개체 유형으로는 연구 주제와 연구자를 다루고 있다.

Abstract This paper describes a vertical search system based on multiple entity-centric unification, which enables to deal with the search queries including multiple domains. To implement the system, we introduced two search technologies; one is for merging service components dynamically according to the entities in the search keywords, and the other is for searching fields with appropriate entities. Our current system includes about 453,000 overseas journal papers for article information search and two entity types; research topic and researcher.

핵심어: Vertical Search, Unified Search, Semantic Web, Multiple Entity, OntoFrame

*정한민: 한국과학기술정보연구원 정보기술연구실 책임연구원 e-mail: jhm@kisti.re.kr

**이미경: 한국과학기술정보연구원 정보기술연구실 연구원 e-mail: jerryis@kisti.re.kr

***성원경: 한국과학기술정보연구원 정책연구실 책임연구원 e-mail: wksung@kisti.re.kr

****류범종: 한국과학기술정보연구원 정보기술연구실 실장 e-mail: ybj@kisti.re.kr

1. 서론

Google (http://www.google.com/)로 대표되는 웹 검색과 대치되는 개념으로 버티컬 검색 (Vertical Search)이 있다. 버티컬 검색은 특정한 이해를 가진 사용자들을 위한 분야가 한정된 검색으로 정의할 수 있다 [4]. 특정한 분야에 대해 검색을 제공하기 때문에 사용자들의 기대 수준 또한 높을 수 밖에 없다. [1]과 [5]의 조사에 따르면, 버티컬 검색의 페이지 뷰 (PV; Page Views)는 매년 30 ~ 60%씩 증가하고 있고, 54%의 사용자가 그들의 특정 비즈니스 또는 작업에 특화된 검색 엔진을 매우 사용하고 싶어한다고 알려져 있다. 이는 버티컬 검색이 단순 검색뿐만 아니라 응용과 문제 해결 기능까지 포함할 수 있어야 한다는 것을 의미한다. 검색 행태에 있어서도 사용자들은 시스템이 올바른 결과를 전달해주거나 올바른 결과를 발견할 수 있도록 도와줄 것으로 강하게 믿고 있으며 목표 중심적 (Goal-oriented)이고 정교하게 검색을 시도하며, 여러 검색어들을 AND 나 OR 없이 나열하는 방식으로 검색 질의를 구성한다고 한다. 검색어로는 대부분 해당 분야 용어들을 사용하기 때문에 버티컬 검색은 검색어 통제에도 신경을 써야 한다.

이러한 특정 분야 검색의 필요성과 요구 사항은 인터넷 상에서의 검색 방식에도 변화를 가져 오게 했는데, 최근 국내 포털들을 중심으로 서비스되기 시작한 버티컬 검색이 대표적인 예이다. 네이버 (http://www.naver.com/)의 영화, 인물, 자동차 검색, 다음 (http://www.daum.net/)의 도서 검색, 파란 (http://www.paran.com/)의 게임, 취업, 재테크 검색 등이 특정 분야에서의 집중적 검색 서비스를 통해 사용자 검색 욕구를 만족시키기 위해 노력하고 있다. 기존 단순 통합 검색과의 차이점은 분야[1]에 따라 특화된 서비스 컴포넌트들을 조합하여 제공한다는 데 있다. 예를 들어, 네이버의 인물 검색에서는 인물 정보, 인터뷰, 팬 커뮤니티, 뉴스 등, 영화 검색에서는 영화, 이미지, 동영상, 뉴스 등의 서비스 컴포넌트[2]들을 집중적으로 제공함으로써 각 분야에서 차별화된 검색 서비스를 제공하고 있다. 2005 년부터 한국과학기술정보연구원 (KISTI)이 개발하고 있는 학술 연구 정보 분석 서비스에서도 2007 년부터 버티컬 검색 개념이 도입되었는데 [2], 연구 주제, 연구자, 학술 행사에 대해 시맨틱 웹 서비스 플랫폼인 OntoFrame 을 이용하여 페이지 유형에 따라 특화된 서비스를 제공하고 있다.

그렇지만, 지금까지 소개한 버티컬 검색 서비스들은 특정 분야에 속한 실체 (인물 분야에서의 "원더걸스", 영화 분야에서의 "쿵푸팬더" 등과 같이 고유 명사 또는 해당 분야에 속하는 개념어가 그 예임) 하나에 대해서만 검색 페이지를 차별적으로 구성한다는 한계가 있다. 예를 들어, 네이버 버티컬 검색 서비스를 이용하여 "쿵푸팬더"에서 주연 성우를 맡았던 "잭블랙"을 동시에 알기 위해 "쿵푸팬더 잭블랙"과 같이 질의어를 입력하면, '인물 + 영화' 페이지가 아닌 인물 페이지나 영화 페이지에서 두 질의어들을 같이 가지고 있는 단순 검색 결과만을 얻을 수 있다 (그림 1 참조). 또한, 동일 분야일지라도 두 개 이상의 실체들로 검색하게 되면 (예. "원더걸스 박진영") 버티컬 검색 결과를 제대로 얻지 못한다. 이러한 한계는 각 분야의 버티컬 검색을 위해 재구성된 방대한 데이터베이스를 동적으로 병합하고 다중 분야에 맞는 서비스 컴포넌트들을 제시할 수 있는 기재의 부족함에 기인한다.

그림 1. 네이버 버티컬 검색에서의 다중 질의어 처리 실패 예 (상단 화면: "쿵푸팬더"에 대한 영화 검색 결과 예, 중단 화면: "잭블랙"에 대한 영화 검색 결과 예, 하단 화면: "쿵푸팬더 잭블랙"에 대한 영화 검색 결과 예)

[1] 본 논문에서는 분야를 특정 영역 내의 실체들을 가지고 있는 범주로 정의하고, 분야를 개체 유형 (Entity Type)으로, 분야 내의 실체를 개체로 대응시켜 기술한다. 이는 서비스 관점에서와 시스템 관점에서 달리 바라볼 수 있게 때문이며, 관점에 따라 용어를 달리 사용하는 것이다.

[2] 검색 결과 페이지를 구성하는 단위 검색 결과 화면

2. 다중 개체 중심적 통합 검색

[5]의 연구에서와 같이 다중 질의어 처리는 버티컬 검색에서 흔히 접하게 되는 현상이기 때문에 수작업에 의존한 버티컬 검색 방식이나 단일 질의어 처리 방식을 그대로 이용하기에는 한계가 있다. 이러한 기존 버티컬 검색의 한계를 극복하기 위해, 본 연구는 두 가지 해결 방안을 제시하고자 한다. 첫째는 질의어를 분석하여 분야 간 결합이 필요한 지를 판단한 후 동적으로 상황에 맞는 서비스 컴포넌트들을 결합하는 것이며, 둘째는 개체 유형 별 필드들을 메타데이터 방식으로 구축하고 필드 별 검색을 수행하여 검색 결과의 정확도를 높이는 것이다. 전자를 위해 표 1 과 같이 개체 유형(분야)들의 조합에 대해 제공할 수 있는 서비스 컴포넌트들을 미리 정의하고, 질의어 분석 결과에 따른 최적 조합을 선정하여 검색 결과를 구성한다. 이 때 최적 조합은 아니지만 애매성이 있어 또 다른 조합들이 발생하는 경우에는 검색 결과 탭을 이용하여 이들을 제시한다 (그림 4 와 5 의 상단 탭 참조). 후자를 위해서는 검색 API 와 추론 API 의 다중 키워드 검색 방식을 필드 내 AND 검색 방식에서 필드 별 AND 검색 방식으로 변경한다. 예를 들어, 네이버의 인물 검색에서는 "쿵푸팬더 잭 블랙"의 다중 키워드가 제목 필드, 본문 필드 각각에서 동시에 검색되어 검색 결과가 혼재되어 나타나지만, 본 연구에서는 "neural network Jinde Cao"의 다중 키워드에 대해 "neural network"는 제목 필드와 연구 주제 필드에서, "Jinde Cao"는 저자 필드에서 검색되어 사용자 의도에 좀더 부합하는 검색 결과를 생성할 수 있게 해준다. 이를 위한 전제 조건은 검색 대상이 메타데이터로 구축되어야 하며, 검색 키워드를 분석하여 개체 유형을 파악할 수 있어야 한다는 것이다.

표 1. 개체 유형들의 조합에 따른 서비스 컴포넌트 매트릭스 (본 연구에서는 연구 주제와 연구자의 두 개체 유형을 다루고 있다.)

서비스 컴포넌트 / 개체 유형 조합	Topic Trends	Social Network	Search Results	…
연구 주제	O	O	O	…
연구자		O	O	…
연구 주제들	O	O	O	…
연구 주제(들) + 연구자(들)	O	O	O	…
연구자들		O	O	…
개체 유형 검색 실패			O	…

본 연구는 시맨틱 웹 서비스 플랫폼인 OntoFrame 을 활용하되, 서비스 분야는 학술 연구 정보로 하여 약 453,000 편의 해외 학술 저널 논문을 메타데이터 기반으로 등록하였다. OntoFrame 은 문서 및 메타데이터를 온톨로지를 참조하여 수집하고 변환하는 URI 서버, 검색 서비스를 제공하는 검색 엔진 (Mariner 2[3]), RDF Triple 형태의 의미 지식을 탑재하고 추론하는 방식으로 추론 서비스를 제공하는 추론 엔진 (OntoReasoner)로 구성된다. 그림 2 와 같이 검색 키워드가 입력되면, 질의 분석기 (Query Analyzer)와 개체 검색기 (Entity Finder)가 검색 키워드 내에 포함된 개체들을 찾아내고, 최적 개체 조합기 (Optimal Entity Mixture Selector)가 여러 개체 조합들 중 최장 일치 방식으로 최적 조합을 선택한다 (예. 검색 키워드가 "semantic web ontology language"인 경우 'semantic web + ontology language'와 'semantic web + ontology + language'의 두 조합으로 분석됨). 이 때 부분 매칭된 결과는 조합 구성에서 제외한다. 예를 들어, 상기 예에서 "web ontology language"가 존재하더라도 "semantic"이 단독 개체로서 존재하지 않으므로 'semantic + web ontology language'는 조합에서 제외된다. 동명 이인의 경우 (그림 4 의 상단 탭 참조, 여기서는 3 명의 "Jinde Cao"가 존재하므로 3 개의 탭이 구성되었다.)에는 이들을 각기 다른 개체로 간주하여 별도의 조합으로 구성한다. OntoFrame 은 URI (uniform Resource Identifier) 식별 체계에 따라 각 개체에 식별자를 부여하기 때문에 이러한 구분이 가능하다. 최적 조합과 기타 후보 조합들에 대해 표 1 의 개체 조합에 따른 서비스 컴포넌트 매트릭스에 따라 버티컬 검색으로서의 개체 페이지를 구성한다 [3].

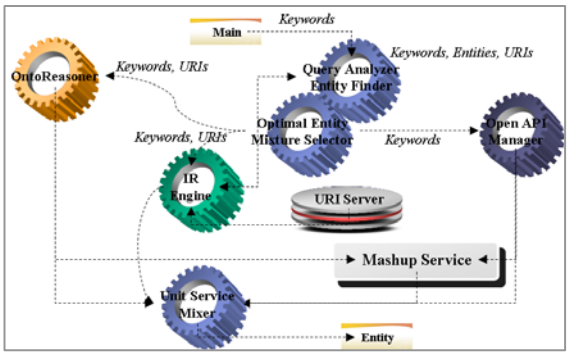

그림 2. 최적 개체 조합 방식에 의한 개체 페이지 구성 흐름도

3. 시스템 구현 및 실험

OntoFrame 이 제공하는 9 가지의 추론 서비스 컴포넌트들을 지원하기 위해서 16 개의 추론 서비스 API 들이 구성되었다. 추론 서비스 API 와 추론 엔진과의 통신은 웹 서비스를 통해 이루어지며 추론의 결과는 XML 형식으로 변환되어 웹 서비스를 통해 통합 검색 서비스로 반환된다. 검색 서비스 역시 동일한 방식으로 구성된다 (그림 3 참조) [6].

[3] http://www.diquest.com/solution/solution_mariner2_1.jsp?menu=1

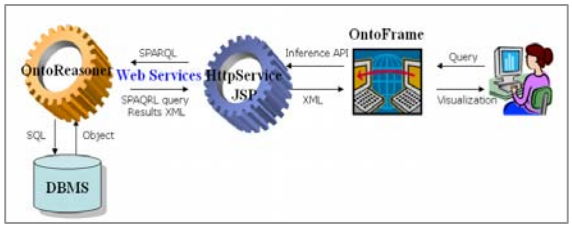

그림 3. 다중 개체 중심적 통합 검색 서비스를 제공하는 OntoFrame 과 추론 엔진 간의 통신 예 (검색 엔진의 경우에는 SPARQL (SPARQL Protocol and RDF Query Language) 처리 부분은 생략되지만, 검색 API 와 XML 형식은 그대로 유지된다.)

그림 4 는 연구 주제와 연구자를 포함하는 검색 키워드에 대한 버티컬 검색 결과의 예이며, 그림 5 는 2 명의 연구자들을 포함하는 검색 키워드에 대한 예이다. 각 서비스 컴포넌트의 결과에서 보여주듯이 메타데이터와 추론을 사용하여 해당 개체에 대응하는 필드들에서 검색을 수행할 수 있게 한다.

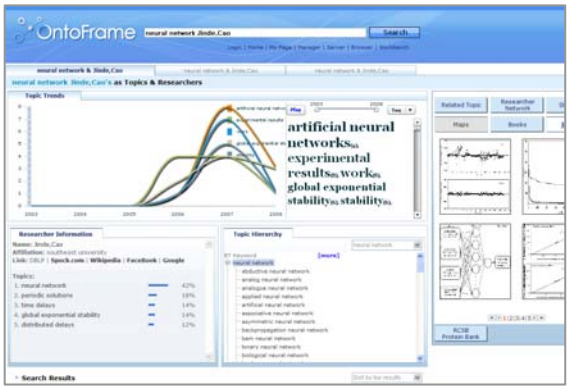

그림 4. 연구주제와 연구자를 포함하는 검색 키워드에 대한 버티컬 검색 예 (“neural network Jinde Cao”)

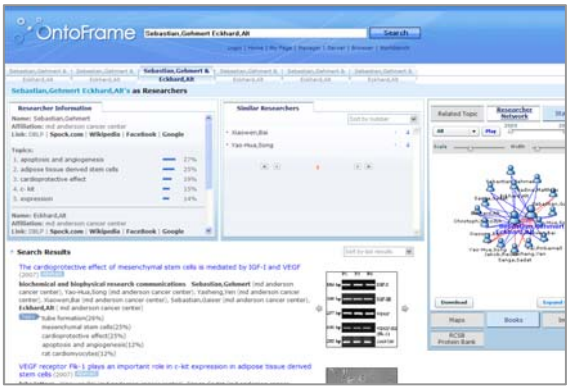

그림 5. 연구자들을 포함하는 검색 키워드에 대한 버티컬 검색 예 (“Sebastian Gehmert Eckhard Alt”)

4. 결론

본 연구는 질의어를 분석하여 분야 간 결합이 필요한 지를 판단한 후 동적으로 상황에 맞는 서비스 컴포넌트들을 결합하였으며, 개체 유형 별 필드들을 메타데이터 방식으로 구축하고 필드 별 검색을 수행하여 검색 결과의 정확도를 높이는 방법을 제시하였다. 이를 통해 버티컬 검색 대상이 되는 개체 유형 내에서의 자유로운 조합을 허용하고, 조합에 따라 서비스 컴포넌트들을 동적으로 배치하고 제시함으로써 사용자 의도에 부합할 수 있는 검색 결과를 생성할 수 있게 하였다. 현재 본 연구가 시스템 신뢰성에 미치는 영향을 분석하기 위한 사용성 평가를 진행 중에 있으며, 조만간 그 결과를 얻을 수 있을 것으로 기대한다.

참고문헌

[1] L. Gregoriadis, “The Changing Digital Environment and Vertical Search”, http://www.slideshare.net/AndyBlack/vertical-search-and-the-changing-digital-world, 2008.

[2] H. Jung, M. Lee, I. Kang, S. Lee, and W. Sung, “Finding Topic-Centric Identified Experts Based on Full Text Analysis,” In Proceedings of the 2nd International ExpertFinder Workshop at ISWC 2007 + ASWC 2007, 2007.

[3] H. Jung, “Multi-entity-centric Integrated Search System and Method,” U.S. Patent Application 12/174730 & PCT Patent Application PCT/KR2008/002270, 2008.

[4] D. McClure, “Top 10 Rules For Vertical Revolutionaries,” http://www.slideshare.net/dmc500hats/top-10-rules-for-vertical-revolutionaries, 2007.

[5] G. McCraken, “Vertical Search: Challenges & Opportunities”, http://www.slideshare.net/AndyBlack/vertical-search-challenges-and-opportunities, 2008.

[6] 이미경, 정한민, 성원경, “OntoFrame: 시맨틱 웹 기반의 추론 서비스”, 한국 IT 서비스학회 추계학술대회, 2008.