이벤트 온톨로지 기반의 의미 정보 검색

(Semantic Search based on Event Ontology)

한 용 진 † 박 세 영 ††
(Yong-Jin Han) (Se Young Park)
이 영 화 ††† 김 권 양 ††††
(Young-Hwa Lee) (Kweon-Yang Kim)

요 약 온톨로지는 웹과 데이타베이스에서 추출하거나 사람으로부터 직접 얻은 정보들을 기계가 접근할 수 있는 개념과 개념 간의 관계 정보로 표현한다. 온톨로지에서 표현된 개념과 관계 정보를 활용할 경우, 의미적인 관계에 따라 정보를 새롭게 재구성할 수 있다. 본 논문에서는 웹상의 인물검색 사이트에서 추출한 정보를 Protege API를 이용해 OWL기반의 IT-People Event Ontology의 스키마에 맞게 온톨로지화 한다. IT-People Event Ontology는 인물과 관련해 잘 변하지 않는 정보와 시간에 따라 변할 수 있는 사건 정보를 표현하고 있다. 이벤트 온톨로지에 기반한 의미 정보 검색은 입력 질의를 온톨로지에 정의된 의미관계에 따라 처리하고, 질의의 유형에 따라 이벤트 온톨로지에 있는 인물 정보를 검색해서 재구성된 결과를 보여준다. 따라서 기존 시스템들의 인물 검색 결과와 비교했을 때, 사용자의 질의 요구에 보다 적합한 검색 결과를 보여 준다.

키워드 : 이벤트 온톨로지, 의미 정보 검색, 변하는 정보

Abstract An ontology provides an explicit specification of concepts and relations on information extracted from database or on human knowledge. Using an ontology, The information can be reconstructed according to semantic relations. In this paper, IT-People Event Ontology is constructed using people information extracted from web portals. IT-People Event Ontology represents constant information and time-temporal information on people. A system using this ontology outputs the well-organized reconstructed information on a specific individual in interest, and then the reconstructed information is suitable for users' demand.

Key words : IT-People, Event Ontology, semantic, temporal information

1. 서 론

웹은 실시간으로 방대한 정보가 쏟아져 나오고 있고, 컴퓨터만 켜면 누구나 다양한 정보를 찾고 이용할 수 있다. 하지만, 정보량이 많아지면서 사람이 직접 필요한 정보를 찾고, 활용하기까지 많은 시간과 노력이 소요된다. 따라서, 정보 검색 시스템은 관련된 정보를 많이 찾는 것 보다는 사용자의 요구에 맞는 정확한 정보를 검색하는 것이 요구된다. 만약, 컴퓨터가 검색 대상 정보와 질의를 의미적으로 처리할 수 있다면 질의 요구에 맞는 정보를 재구성해서 사용자에게 제공할 수 있을 것이다.

현재, 상용화된 검색 시스템들은 문서와 질의를 키워드 기반으로 처리하고 있다[1]. 질의 키워드가 출현하는 빈도수가 높은 문서를 검색하기 때문에 관련된 문서를 다수 찾아 주고 있지만, 검색된 문서가 실제로 사용자의 요구에 맞는지는 보장할 수 없다. 예를 들어, "이재용의 이력사항은?"이라고 질의하면, '이재용', '이력사항'이 포함된 문서를 검색하지만, 이력사항과 관련된 '졸업', '경력', '인적사항' 등의 내용은 포함하지 않을 수도 있다. '이력사항'을 의미적으로 해석할 수 없기 때문이다.

인물 검색을 목적으로 하는 사이트의 경우, 데이타베이스로 저장된 정보를 HTML 테이블 형태로 제공하고 있다. 예를 들어, "1950년 대구에서 태어난 경기고 출신 행정공무원"이라는 질의를 통해 해당 내용을 만족하는 인물 정보를 검색한다. 이 경우, 인물 정보를 출생일, 출신지역, 출신학교, 직업 정보에 따라 분류하고, 이러한 분류 정보에 해당하는 키워드를 입력받아 인물을 검색한다. 하지만, 인물의 '학력', '이력사항', '출신학교'와 같이 인물과 관련된 검색 대상을 직접적으로 처리하지는 못한다. "이재용의 학력사항"을 검색하기 위해서는 이재용을 찾고, 검색된 인물 관련 내용 중에서 학력과 관련된 내용을 사용자가 직접 찾아야 한다.

이러한 기존 검색의 문제점을 해결하기 위한 노력으

· 본 논문은 정통부 및 정보통신연구진흥원의 정보통신선도기반기술개발사업의 연구결과로 수행되었습니다.
· 이 논문은 2007 한국컴퓨터종합학술대회에서 'Event 온톨로지 기반의 의미 정보 검색'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 경북대학교 컴퓨터공학과
yjhan@sejong.knu.ac.kr
†† 종신회원 : 경북대학교 컴퓨터공학과 교수
sypark@sejong.knu.ac.kr
††† 정 회 원 : 경북대학교 컴퓨터공학과 교수
yhlee@sejong.knu.ac.kr
†††† 정 회 원 : 경일대학교 컴퓨터공학과 교수
kykim@kiu.ac.kr

논문접수 : 2007년 10월 2일
심사완료 : 2007년 12월 28일

로 온톨로지를 이용한 검색 방법이 연구되고 있다. Wallace [2]는 온톨로지에 정의된 개념을 이용해 질의를 확장하는 방법을 제안했다. 질의어의 각 단어를 개념적으로 해석하지만, 각 개념들 사이에 관계를 독립적인 것으로 보기 때문에 단어 사이의 의미 관계를 반영하지는 못한다. 질의 단어들 간의 관계를 통해 질의어에 대한 가중치를 반영한 연구도 있다[3]. 하지만, 이러한 연구는 검색 대상을 문서로 하기 때문에 결국, 사용자는 필요한 정보를 직접 재구성해야한다. SEAL[4]은 온톨로지 자체를 검색 대상으로 하는 검색 서비스를 제공한다. 질의어는 온톨로지에 있는 개념 관계에 따라 해석되고, 사용자 요구에 맞는 정보를 제공한다. 하지만, 온톨로지를 이용해 인물에 대해 변하는 정보를 표현하고 재구성된 인물 정보를 출력하는 검색 모델을 제시한 예는 없다.

본 논문에서는 IT-People Event Ontology(ITPEO)[1]의 스키마를 이용해서 서로 다른 소스로부터 추출된 인물정보를 검색하고, 사용자의 질의 요구에 맞게 재구성된 결과를 보여 줄 것이다. 검색 시스템이 질의어를 온톨로지의 개념과 개념 간의 관계 정보로 이해함으로써 사용자의 요구에 맞는 정보를 검색할 수 있다.

먼저, 2장에서 ITPEO를 소개한다. 다음으로 3장에서 온톨로지 기반 사건 정보 검색 시스템의 전체 구성을 설명하고, HTML 테이블 정보와 온톨로지간의 매핑관계를 설명한다. 이어서 4장에서 ITPEO에서 정의하고 있는 이벤트 개념에 기반한 정보검색 방법을 제시한다. 마지막으로 5장에서는 기존의 검색 시스템과 온톨로지 기반 사건 정보 검색 시스템을 비교 분석한다.

2. IT-People Event Ontology

ITPEO는 IT관련 인물 및 조직의 활동을 시간과 장소 정보에 따라 구분되는 사건으로 표현하고 이들 간의 의미관계를 정의한 지식기반이다. 크게 인물(People), 조직(Organization), 시간(Time), 장소(Region), 상품(Product)에 대한 개념과 이들 간의 관계로 표현될 수 있는 사건(Event)을 정의한다[5]. 인물, 조직, 장소, 상품은 잘 변하지 않는 정보를 속성(property)으로 가지게 된다. 예를 들어 인물의 경우, 이름, 생년월일, 출생지가 속성이다.

사건(Event)은 특정한 사건 내용(Predicate)을 중심으로 시간, 장소 정보에 따라 변할 수 있는 정보를 표현한다. 그림 1은 이벤트 온톨로지를 중심으로 한 스키마 구성과 시간이 따르는 사건 정보 인스턴스 예를 보여주고

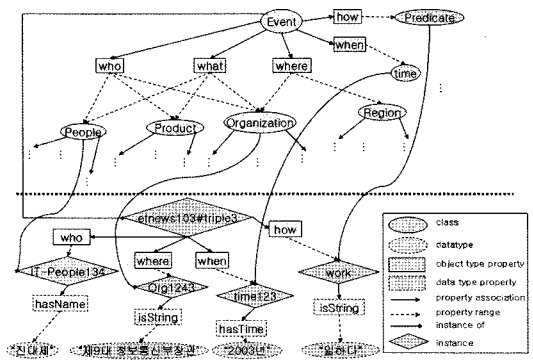


그림 1 ITPEO의 인스턴스 예

있다. 사건의 주체로서 "진대제"는 "People"이라는 개념을 가지고, "제 9대 정보통신부장관"은 조직의 직위 정보로서 "Organization"이라는 개념에서 정의된다. "2003년"이라는 시간 정보는 "Time"으로 표현되고 "일하다"는 사건의 내용으로서 인스턴스 "work"에 대한 문자값을 의미한다. 전체 사건은 출처와 날짜, 추출 번호로 표현된 유일한 ID값(etnews103#tripe3)을 가진다.

ITPEO는 전자신문을 대상으로 사람이 직접 추출한 정보와 일반 텍스트에서 패턴에 기반해 추출한 정보로 이루어져 있다. 신문 기사의 경우, 어느 정도 정형화된 문체를 사용하기 때문에 표현 유형을 정의함으로써 정보를 추출할 수 있다. 자연어 처리 기술과 문장의 특정 패턴에 기반하여 일반 텍스트로부터 정보를 추출하고, 온톨로지화 하는 연구는 진행 중에 있다[5].

일반 텍스트 문서에서 인스턴스 및 속성(property) 생성을 완전 자동화하는 방법은 아직 연구되고 있는 분야이다. Celjuska와 Vargas-Vera[6]는 비구조 문서로부터 인스턴스 후보를 생성하고 최종적으로 사람의 판단에 따라 인스턴스를 생성하는 반자동 방법을 제안했다. Bernardo 와 동료 연구원들[7]은 텍스트 문서로부터 제한된 형태의 인스턴스를 생성한다.

3. 온톨로지 기반 사건 정보 검색 시스템

ITPEO는 서로 다른 영역의 소스로부터 온토롤지의 개념과 개념관계에 따라 정보를 인스턴화해서 저장한다. 데이타 베이스나 웹, 혹은 비구조화된 문서로부터 자동으로 추출된 정보와 사람의 지식을 온토롤지로 표현한다(그림 2).

본 논문에서는 주로 HTML 테이블 정보를 추출하여 온토롤지화 한다. 인물의 이름, 출생지, 출생(생년월일) 등을 기본정보로 하고, 직업, 소속, 학력, 경력, 수상내역과 같이 시간에 따라 변할 수 있는 정보는 이벤트의 인스턴스로 입력한다. 정보 추출을 위해 표의 속성과 온톨

1) ITPEO는 정통부 및 정보통신연구진흥원의 정보통신선도기반기술개발사업의 연구결과로 구축되었습니다.

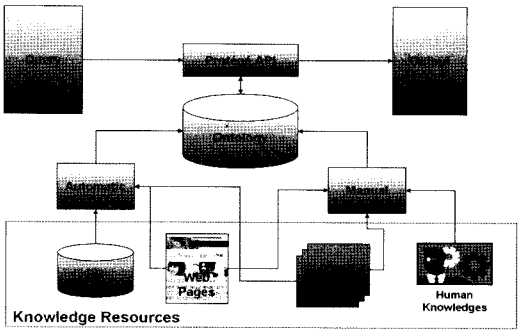


그림 2 온톨로지 기반 사건 정보 검색 시스템

로지 개념 매핑 테이블을 구성하고 속성값을 대응하는 개념의 인스턴스 및 데이터값으로 한다.

4. 이벤트를 기반으로 한 정보검색

이벤트가 가질 수 있는 사건내용(predicate), 주체(who), 대상(what), 장소(where), 시간(time) 속성은 질의 내용에 있는 키워드의 의미와 관계를 가진다. 예를 들어 "삼성전자 이재용의 이력사항은?"이라고 질의할 경우, "삼성전자"라는 문자정보를 가지는 인스턴스는 Organization이라는 개념을 갖기 때문에 주체(who)나 대상(what), 장소(where) 속성의 값이 될 수 있다. "이재용"이라는 문자정보를 가지는 인스턴스는 People이라는 개념을 갖기 때문에 주체(who)나 대상(what) 속성의 값이 된다. "이력사항"은 "졸업하다", "일하다", "수상하다" 등의 문자정보를 가지는 사건내용(Predicate)의 인스턴스를 질의 대상으로 한다.

이러한 질의문과 이벤트 개념과의 관계를 이용해서 질의 유형에 따라 질의 내용을 반영하는 결과를 출력한다.

4.1 질의 유형

자연어 질의는 형태소 분석과 구문 분석, 그리고 개체명 인식 정보를 바탕으로 질의 유형이 결정된다[8]. 본 검색 시스템은 유형이 분석된 자연어 질의를 입력으로 받는다. 예를 들어 다음과 같은 유형을 처리하게 된다.

그림 3의 유형1은 기본적으로 사람 이름을 대상으로 인물 검색을 한다. 조직이름으로 대상 인물을 제약할 수도 있다. 유형2는 유형1의 결과에서 질의영역을 만족하는 정보를 검색한다. 예를 들어, 질의영역으로 이력사항, 경력사항, 학력사항, 수상내역 등이 있다. 경력사항은 어디에서 일했고, 어디에 소속이라는 정보를 포함한다. 학력사항은 어떤 학교를 졸업했는지에 관한 정보이다. 이력사항은 경력, 학력, 수상 내역을 모두 포함하는 정보를 검색대상으로 한다. 유형3은 유형2의 결과에서 조건을 만족하는 검색을 수행한다. 현재는 사건 정보를 시간

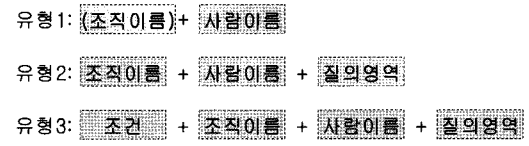


그림 3 질의 유형

조건에 따라 검색을 수행한다.

이러한 유형을 이용해 처리할 수 있는 자연어 질의 예는 다음과 같다.

1. 이재용의 {이력, 경력, 학력, 수상내역}은?
2. 삼성전자 이재용의 이력사항은?
3. 2002~2006년 사이에 삼성전자 진대제의 이력은?

질의1은 이름이 "이재용"인 모든 인물을 검색대상으로 한다. 질의2는 유형2에 해당한다. 먼저 검색된 "이재용"이라는 인물 중 삼성전자에 소속된 인물을 찾고 이력사항에 해당하는 "졸업하다", "수상하다", "일하다"와 관련된 사건 정보를 검색한다. 질의3은 유형3에 해당한다. 질의2에서 검색된 결과 중 시간이 2002년에서 2006년 사이인 정보를 검색한다. 검색된 사건 정보는 시간 순서에 따라 정렬된 형태로 출력되고, 인물은 기본 정보를 함께 출력함으로써 동명이인을 판단할 수 있는 근거를 사용자에게 제공한다.

4.2 이벤트에 기반한 정보 검색

인물에 대한 검색은 기본적인 정보와 시간에 따라 변할 수 있는 사건에 대한 검색을 대상으로 한다. 기본 정보는 이름, 생년월일, 출생지에 해당하며, 동명인에 대해 구분할 수 있는 근거가 된다.

사건 정보 검색은 Protege API를 통해 ITPEO의 이벤트를 기반으로 구현하고 있다.

그림 4는 질의문 "삼성전자 이재용의 이력사항은?"에 대한 검색 예이다. 삼성전자는 Organization에 대응되고, 이재용은 People에 대응된다. 먼저, 이벤트의 인스턴스 중에서 Organization이 "삼성전자"이고 People이 "이재용"인 것을 찾는다.

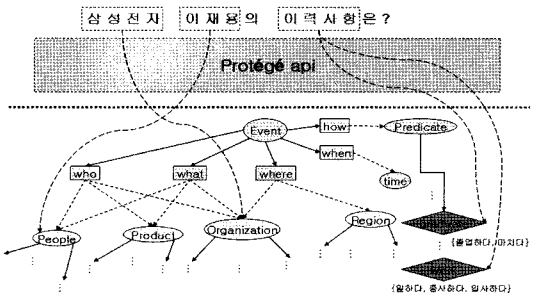


그림 4 이벤트에 기반한 정보 검색

질의 영역에 해당하는 검색대상은 Predicate의 인스턴스 집합으로 표현된다. 따라서, 이력사항은 "graduate", "work"를 검색 대상으로 한다. "graduate"은 소스 정보에 따라 "졸업하다"로 표현될 수도 있고, "마치다"로 표현될 수도 있다. "work"의 경우도 "일하다", "종사하다", "입사하다"등 다양한 형태로 표현될 수 있다.

또한, Predicate의 인스턴스들은 서로 같은 문자 정보를 값으로 가질 수도 있다. 예를 들어, 일을 마쳤다는 의미로 "end"라는 인스턴스와 졸업했다는 의미로 "graduate"라는 인스턴스는 모두 "마치다"라는 문자 표현이 가능하다. 시스템은 이러한 문자 정보에 관계없이 "graduate", "work"라는 개념을 대상으로 검색한다. 최종 검색된 결과는 시간 순서에 따라 정렬된 형태로 출력한다.

5. 추출 정보 및 검색 결과 분석

웹사이트로부터 9,257명의 인물에 대한 기본 정보를 추출했다. 이벤트의 인스턴스에 해당하는 각 인물의 경력, 학력, 수상내역은 총 53,469개를 추출했다(표 1).

표 1 추출된 정보

인물	수상내역	경력	학력	합계
9,257	16,889	24,532	12,048	62,726

기본 정보의 값에 해당하는 지역이름과 사건 정보의 소속에 해당하는 회사, 학교 등의 이름은 각각 유일한 이름을 갖는 객체로 보고 모두 인스턴스화 되었다. 같은 이름의 다른 인물이 있을 수 있으므로 생년월일, 출생지에 따라 구분될 수 있는 인스턴스로 입력되었다.

그림 5는 "삼성전자 이재용의 이력사항은?"에 대한 검색결과이다. graduate 개념에 해당하는 "졸업하다", "마치다"에 해당하는 사건과, work 개념에 해당하는 "일하다", "소속이다", "취임하다", "운영하다"에 해당하는 사건 정보를 출력해서 보여주고 있다. 각 정보는 시간 순서에 따라 정렬된 형태로 출력되었다. 그림 6은 키워드 기반 검색 시스템에서 질의한 결과이다. "삼성전자"와 "이재용"이 가장 많이 포함되었고, 최근에 사람들이 가장 많이 찾은 문서를 검색해서 보여 주고 있다. 검색된 결과는 질의의 키워드가 포함된 관련된 문서일 뿐 정확하게 사용자의 요구를 반영했다는 보장이 없다. 따라서, 사용자는 각 문서를 검색하면서 직접 이력사항에 해당하는 정보를 찾아야 한다.

그림 7은 인물 검색 사이트에서 "이재용"이라는 인물 중에서 환경부 장관 출신의 이재용을 검색한 결과의 일

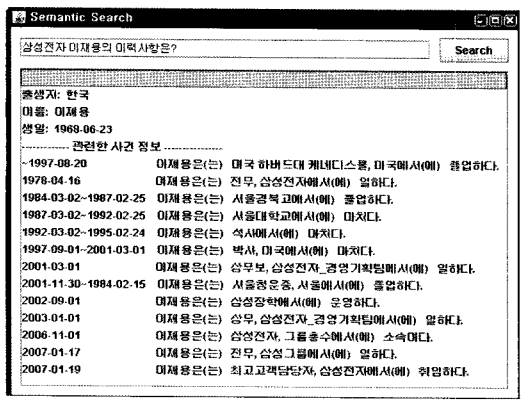
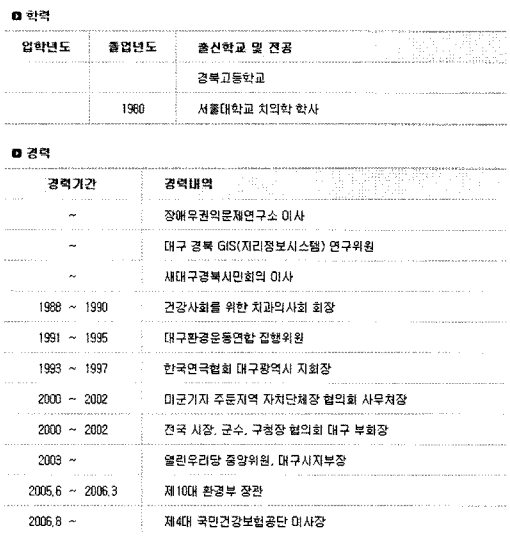


그림 5 "삼성전자 이재용의 이력사항은?

그림 6 키워드 기반 검색 결과



학력

입학년도	졸업년도	출신학교 및 전공
		경북고등학교
	1980	서울대학교 치의학 학사

경력

경력기간	경력내역
~	장애우권익문제연구소 이사
~	대구 경북 GIS(지리정보시스템) 연구위원
~	새대구경북시민회의 이사
1988 ~ 1990	건강사회를 위한 치과의사회 회장
1991 ~ 1995	대구환경운동연합 집행위원
1993 ~ 1997	한국연극협회 대구광역시 지회장
2000 ~ 2002	미군기지 주둔지역 자치단체장 협의회 사무처장
2000 ~ 2002	전국 시장, 군수, 구청장 협의회 대구 부회장
2003 ~	열린우리당 중앙위원, 대구시지부장
2005.6 ~ 2006.3	제10대 환경부 장관
2006.8 ~	제4대 국민건강보험공단 이사장

그림 7 인물 검색 사이트

부이다. 데이타 베이스의 구조의 맞게 정보를 보여주기 때문에 시간 정보에 따라 정보를 재구성하지는 못한다. 또한 질의를 통해 "학력", "경력"에 해당하는 내용을 직접적으로 추출해서 보여주지 않는다.

그림 8은 "1993~2006년 사이에 이재용의 경력사항은?"이라고 질의한 결과이다. "이재용"이라는 이름을 가진 모든 인물들을 검색하고 그중에서 시간 조건에 맞는 경력사항에 관한 정보를 검색한 결과이다. reply1에 해당하는 "이재용"은 그림 7의 인물에 해당한다.

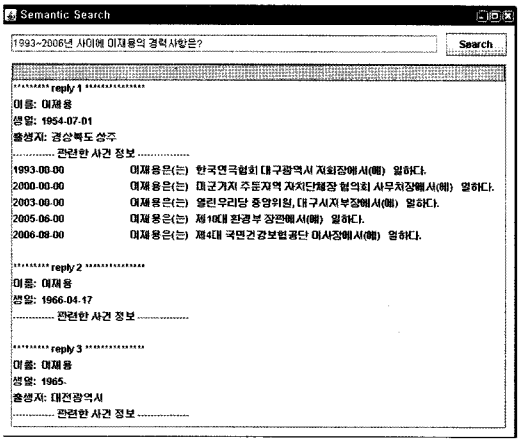


그림 8 "1993~2006년 사이에 이재용의 경력사항은?"

만약, "환경부장관 이재용"으로 인물에 대한 추가 정보를 줄 경우, reply2와 reply3의 정보는 제외된다. 그림 7과 달리 시간 조건을 반영한 검색 결과를 보여주고 있다. 또한 학력, 경력, 수상내역 등 다양한 정보 중에서 경력에 해당하는 정보만 검색해서 보여준다.

6. 결론 및 향후 과제

키워드 기반 검색은 질의어를 포함하는 관련 문서를 검색해서 보여주지만 자연어 질의의 요구를 정확하게 반영하지는 못한다. 데이터베이스를 기반으로 한 검색은 신뢰할 수 있는 정보를 제공해 주지만, 변하는 사건정보나 부분적인 개념관계에 대해 유연하게 대처하지 못한다.

온톨로지를 이용할 경우, 질의어를 의미적으로 처리함으로써 질의 요구를 반영한 검색을 할 수 있다. 또한, 웹 혹은 데이타베이스의 다양한 정보를 자동으로 온톨로지화 함으로써 인물과 관련된 변화하는 사건 정보도 개념 단위로 처리할 수 있다.

또한, 개념 단위의 정보들은 각각을 하나의 객체로 다룰 수 있기 때문에 다양한 형태의 표현이 가능하다. 온톨로지에서 얻은 사건 정보를 자연어 형태로 생성하거나 인물과 연관된 정보를 그래프 형태로 표현함으로써 사용자가 이해하기 좋은 형태로 결과를 보일 수 있다.

웹상에서는 인물에 관해 여러 사이트에서 정보를 제공한다. 향후 과제로 여러 사이트에서 다르게 표현된 인물 정보를 하나의 온톨로지로 통합하는 연구를 할 계획이다. 동일 인물에 대해 어떤 사이트에서는 이름, 출생, 출생지 정보를 제공하고, 다른 사이트에서는 성별 정보를 제공할 수 있다. 또는, 출생이라는 표현을 생년월일로 표현하는 경우도 있을 것이다. 출생과 생년월일을 개념 수준에서 이해한다면 여러 사이트의 정보를 하나로 통합할 수 있다.

또한, ITPEO의 술어 정보를 확충함으로써 처리할 수 있는 질의 형태를 확장할 수 있다. 그리고 이미 구축된 조직 혹은 인물 온톨로지와 호환해서 응용함으로써 다양한 응용에 유용하게 쓰일 수 있다.

참 고 문 헌

[1] Baeza-Yates R., and Ribeiro-Neto B., Modern Information Retrieval, *Addison Wesley*, 1999.

[2] Wallace A. P., and Ana M. C., "An Ontology Based-Approach for Semantic Search in Portals," *In Proceedings of the 15th International Workshop on Database and Expert Systems Applications*, pp. 127-131, 2004.

[3] Jose A. R., Eduardo M., Jorge B., and Arantza I., "Seraching the Web: From Keywords to Semantic Queries," *In Proceedings of the Tird International Conference on Information Technology and Applications*, pp. 244-249, 2005.

[4] Maedche A., Staab S., Stojanovic N., Studer R., Sure Y., "SEmantic portAL: The SEAL Approach," *In Fensel D., Hendler J. A., Lieberman, H., Wahlster W. (eds.), Spinning the Semantic Web. MIT Press*, pp. 317-359, 2003.

[5] 경북대학교, 온톨로지 검증 및 온톨로지 기반 인스턴스 생성에 관한 연구, 최종 보고서, pp. 52-65, 2006.

[6] Celjuska D., and Vargas-Vera M. "Ontosophie, A Semi-Automatic System for Ontology Population from Text," *In proceeding of the International Conference on Natural Language*. Hyderabad, India (2004).

[7] Bernardo M., Emanuele P., Octavian P. and Manuela S., "Ontology Population from Textual Mentions: Task Definition and Benchmark," *In Proceeding 2nd Workshop on Ontology Learning and Population at COLING/ACL2006*, Sydney Australia, pp. 26-32, 2006.

[8] Stralkowski T., Natural language information retrieval, *Information Processing & Management*, pp. 397-417, 1995.