

일본어 DODDLE와 범용 온토로지를 이용한 도메인 온토로지의 구축 및 평가

홍윤기[†], 김태석^{**}, 山口 高平^{***}

요 약

인터넷의 확산으로 방대한 정보가 웹에 넘쳐나고 있지만, 사용자가 필요한 정보를 얻기 위해서는 검색 시스템 이용이 필수적이다. 사용자가 원하는 정보를 검색 시스템 결과로부터 얻는 것이 힘들기 때문에, 검색 시스템의 결과 향상에 관하여 다양한 연구가 진행되고 있다. 온토로지의 구축은 많은 비용이 필요하기 때문에, 일본어 온토로지 구축을 용이하게 해주는 일본어 DODDLE이라는 툴을 사용한다. 본 논문에서는 범용 온토로지를 이용한 온토로지 구축 라이프 사이클, 문서 검색과 관련된 온토로지(Ontology) 구축 방법과 구축된 온토로지를 이용하여 검색 결과를 향상시키는 방법을 제안한다. 그리고 제안된 방법의 효율성을 입증하기 위하여 사례 연구법을 사용하였고, 로켓운용지원 온토로지를 구축하여 실험한 결과를 보여준다.

Constructing Domain Ontologies Using Japanese DODDLE and General Ontologies

Yun-Ki HONG[†], Tai-Suk Kim^{**}, Takahira YAMAGUCHI^{***}

ABSTRACT

With the advancement of the Internet, bulky information overflows in the Web. When the Internet user wants to get the necessary information, it is essential to use the retrieval system. It is not easy for the user to get the information from the result of the retrieval system. Various research activities have been advanced for the result improvement of the retrieval system. Although retrieval results can be improved with ontology, it usually takes lots of costs for users to construct the Japanese domain ontology. This paper discusses how to integrate search result refinement and domain ontology refinement using the domain ontology tool called Japanese DODDLE, and how to improve the research result using the constructed ontology. To prove the effectiveness of the suggested methodology, the case study with rocket operation is performed and it shows that the methodology can be promising.

Key words: Ontology(온토로지), Ontology Lifecycle(온토로지 라이프 사이클), Concept Searching(개념 검색), Document Searching(문서 검색), Semantic Web(시맨틱 웹)

1. 서 론

오늘날에는 인터넷이나 정보 단말기의 보급과 정보 기술의 발달로 정보가 넘치고 있는 환경에 살고

있고, 방대한 정보들로부터 사용자가 필요한 정보를 얻기 위해서 대부분 검색 시스템을 사용한다. 그러나 검색 시스템을 사용하는 경우에도 그 결과가 방대하기 때문에 원하는 정보를 정확하게 얻는 것이 어렵

※ 교신저자(Corresponding Author): 김태석, 주소: 부산광역시 부산진구 엄광로 995(614-714), 전화: 051)890-1707, FAX: 051)890-1724, E-mail: tskim@deu.ac.kr

접수일: 2005년 8월 10일, 완료일: 2005년 10월 17일

[†]일본 KEIO대학 이공학연구과 개방환경과학 오픈시스템

매니지먼트 전수 박사과정

^{**} 종신회원, 동의대학교 공과대학 소프트웨어공학과 교수

^{***} 일본 KEIO대학 이공학부 관리공학과 교수

(E-mail: yamaguti@ae.keio.ac.jp)

다. 그래서 검색 시스템의 정밀도를 향상시키기 위한 연구가 다양하게 진행되고 있다.

검색 시스템의 정밀도를 향상시키기 위해서는 컴퓨터가 지식을 이해할 수 있도록 해주는 온토로지 기술을 주목할 필요가 있다. 온토로지란, 인공지능의 입장에서 『개념화의 명시적인 기술』이라고 정의된다[1]. 여기에서 개념화라고 하는 것은 대상(세계)에 관하여 흥미를 갖는 개념과 그러한 대상들 사이의 관계를 명확히 하는 것을 가리킨다.

비즈니스 프로세스 모델의 구축에 온토로지를 적용하면, 먼저 비즈니스 자체를 구성하는 개념을 조사하고 분석하여 그것들 사이에 관한 관계를 정리하고 개념을 결정하며, 각각의 구성요소들을 명확하게 구별하기 위한 속성을 결정하는 개념을 기술한다. 그러한 개념 중에는 오브젝트(명사에 대응)뿐만 아니라, 활동을 나타내는 동사적인 것도 존재할 수 있다. 그리고 활동을 대상으로 하는 오브젝트와 각각의 구성요소들을 관계 짓는 제약 등 이들 전부를 체계화한 것을 온토로지라고 부른다.

다양한 분야에서 온토로지의 개발이 이루어지고 있는데, 기능모델링 및 모델데이터 표준화 등의 공학 분야[8], 계통연구 등의 생물학[9], 법학[10] 등, 특정 연구 분야에 있어서는 실용단계까지 개발이 이루어지고 있다. 그렇지만, 일반기업에 있어서는 정보공유, 지식공유 등의 목적으로 온토로지 구축이 이루어지지 못하고 있는데, 그 이유는, 첫째, 일반기업에 있어서 온토로지의 유효성이 충분히 인지되어 못했고, 둘째, 온토로지 구축 및 유지·보수에 엄청난 코스트(시간과 비용)가 들기 때문이다.

본 논문의 목적은 앞에서 언급한 검색 시스템의 정밀도 향상시키기 위하여 범용 온토로지를 이용하여 영역 온토로지 성능을 향상시키기 위한 라이프사이클과 문서 검색을 관련지어 검색 정밀도 문제를 해결하는 것이다.

2. 관련연구

본 절에서는 사용자에게 필요한 정보를 제공하기 위해 구축한 영역 온토로지에 대하여 지금까지 진행해온 연구 시나리오에 대한 설명과 특정 분야 영역 검색에 필요한 온토로지의 구축과 성능을 향상시키기 위한 라이프사이클에 대해 설명한다.

2.1 시나리오

그림 1은 본 논문에서 제안하는 시스템의 전체적인 시나리오를 나타내고 있다.

전문 용어를 키워드로 전문 분야의 문서를 검색할 경우에 키워드에 엄밀하게 일치하는 문서가 발견되면 좋지만, 키워드에 엄밀하게 일치하지 않는 경우에는 검색 결과가 없는 것 보다는 입력한 키워드에 유사한 문서를 제공하는 쪽이 사용자에게는 유익한 정보 제공하는 것이라고 할 수 있다. 입력된 키워드와 유사한 문서 검색은 키워드 검색 시스템에 동의어 검색이나 개념 계층을 이용한 검색 기능의 추가로 구현 할 수 있다.

동의어 및 개념 계층을 이용한 검색기능은 온토로지를 이용하여 구현한다. 검색기능에 이용할 온토로지로서 범용 온토로지를 사용할 경우에는 온토로지가 이미 구축되어 있기 때문에 비용이 적다는 장점이 있으나, 문서 검색에서 요구되는 검색 정밀도가 낮다는 단점을 가지고 있다. 다른 방법으로 영역 온토로지를 이용할 경우에는 온토로지 구축비용은 많이 들지만 검색 시스템에서 요구하는 검색 정밀도는 높다는 장점을 가진다.

본 논문에서는 정밀도가 높은 검색 시스템이 필요하기 때문에, 영역 온토로지를 이용하기로 결정했다. 영역 온토로지의 높은 구축비용을 경감시키기 위해서, 일본어 DODDLE(Domain Ontology rapid Development Environment)를 이용하여 반자동으로 범용 온토로지로부터 영역 온토로지를 구축하였다. 틀

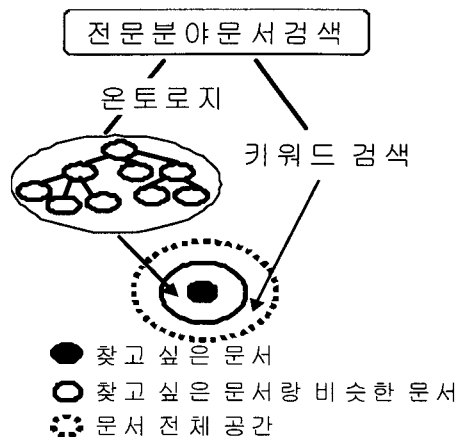


그림 1. 시나리오

로부터 구축된 영역 온토로지(초기 영역 온토로지)는 특정 전문 영역의 검색에 특화되어 있지 않기 때문에 초기 영역 온토로지를 전문 분야 문서 검색 시스템에 탑재한 후, 사용자의 검색 작업 반복을 통하여 온토로지를 이용한 검색 결과에 관한 평가를 온토로지에 다시 반영시켰다. 이 과정을 통하여 최종적으로는 전문 분야의 검색에 특화된 영역 온토로지를 구축할 수 있었으며, 동시에 검색 결과의 정밀도도 향상시킬 수 있다.

2.2 온토로지 라이프 사이클

전문 분야의 문서 검색 시스템에 필요한 온토로지를 처음부터 구축하는 것은 시간과 비용이라는 측면에서 효과적이지 못하다. 그래서 일본어 DODDLE이라는 툴을 사용하여 초기 영역 온토로지를 작성하였고 검색 시스템에 탑재시켰다. 시스템의 사용자가 탑재된 온토로지를 이용하여 검색을 수행할 때마다 검색 결과에 관한 평가를 온토로지에 반영시켰다. 이러한 반복 과정을 통해서 최종적으로 특정분야 검색에 보다 적합한 영역 온토로지 로 진화되는 동시에 검색 시스템의 검색 정밀도도 향상되었다. 이러한 작업 과정을 “온토로지 라이프 사이클”이라고 한다.

3. 온토로지 구축 도구

본 논문에서 제안된 시스템을 구축하기 위해 사용

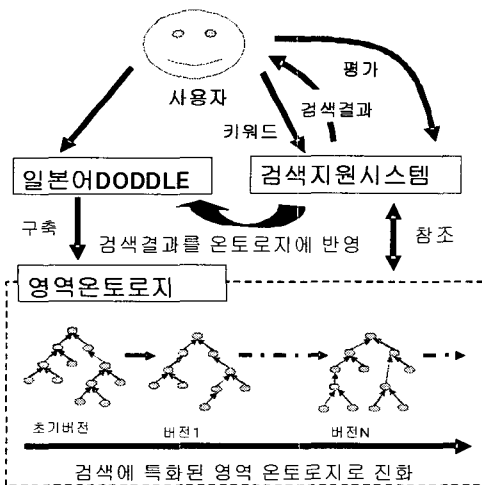


그림 2. 온토로지의 진화

한 일본어 온토로지 구축지원 툴과 개념 계층을 구축하기 위한 전자 사전의 사용에 대해서 설명하고 문서 검색을 위해 사용된 시스템에 대하여 설명한다.

3.1 일본어 DODDLE 시스템

일본어 DODDLE은 일본어 영역 온토로지 구축 지원 툴이다. 툴 내부의 처리과정은 그림 3과 같다. 시스템의 입력은 온토로지를 구성하기 위한 어휘(입력 어휘)다. 입력어휘는 일본어 형태소해석시스템인 茶釜(ChaSen)[7]으로 명사/동사만을 추출하여 어휘의 기본형 입력어휘로 선택하는 전처리를 통하여 구축하였다. 입력 모듈에서는 입력 어휘와 EDR[4]의 개념을 대응시킨다. 개념 계층 구축 모듈에서는 입력 모듈에서 대응된 개념(입력 개념)집합과 말단 노드라고 하는 부분 목을 EDR으로부터 추출한다.

이때에 입력 개념간의 위상 관계(선조·친자·형제 관계)를 유지하는 것에 공헌하는 중간 개념과 공헌하지 않는 중간 개념을 판별한다. 판별 결과 불필요한 중간 개념은 삭제된다(전정 작업). 최종적으로 입력 개념 집합과 필요한 중간 개념 집합을 이용하여 영역 온토로지가 구축된다. 그리고 구축된 온토로지는 OWL (Web Ontology Language)[2]형식의 파일로 출력될 수 있다.

일본어 DODDLE과 같은 온토로지 구축 지원 툴

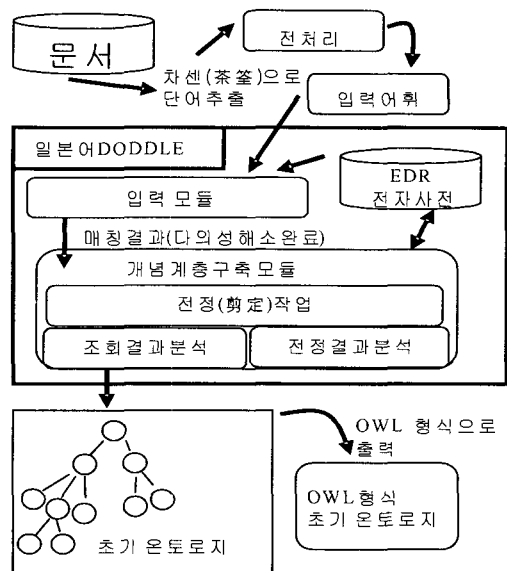


그림 3. 일본어 DODDLE 시스템의 처리과정

의 선행 연구로서, 영역 온토로지 구축 지원 환경 DODDLE-OWL(Domain Ontology rapid Development Environment - OWL Extension)[5]이라고 하는 시스템이 있다. DODDLE-OWL은 시스템의 입력으로 온토로지를 구성하는 어휘, 입력된 어휘와 대응관계를 참조하기 위한 전자 사전인 WordNet[3]과 텍스트 코퍼서를 이용한다. 그리고 영역 온토로지를 OWL 형식의 파일로 출력할 수 있다.

3.2 EDR 전자사전

일본어 DODDLE에서는 개념 계층을 구축하기 위하여 일본 독립행정법인 정보통신연구기구의 EDR 전자사전을 사용한다. EDR 전자 사전은 일본어 단어 사전(27만 어휘), 영어 단어 사전(19만 어휘), 개념 사전(41만 어휘), 일영 대화 사전(23만 어휘), 영일 대화 사전(16만 어휘), 일본어 공기 사전(90만 어휘), 영어 공기 사전(46만 어휘), 일본어 코퍼스(20만 문), 영어 전 자료(12만 문), 전문 용어 사전(정보 처리)으로 구성되어 있다. 일본어 DODDLE에서는 상기의 사전 중에서 일본어 단어 사전, 영어 단어 사전, 개념 사전을 사용한다.

3.3 일본어 DODDLE

그림 4와 그림 5는 일본어 DODDLE을 보여준다. 그림4는 온토로지를 작성하기 위한 어휘를 입력한 후의 화면을 나타내고, 그림5는 그림 4에서 매칭된 어휘 리스트를 EDR의 개념 사전을 참조하면서 개념 계층을 구축한 후의 화면을 나타내고 있다.

그림 4의 틀에 어휘를 입력하면 EDR 전자사전을 비교한 후, 매칭 한 어휘 리스트를 ①에 출력하고,

매칭 하지 않은 어휘를 ⑦에 출력한다. 사용자가 ①의 어휘 리스트에서 1개의 단어를 선택하면, ②는 ①에서 선택된 단어를 표제어에서 포함하고 있는 개념 ID를 출력한다. ②에서 1개의 개념 ID를 선택하면 ③에 일본어 표제어, ④에 영어 표제어, ⑤에 일본어 설명, ⑥에 영어 설명이 표시된다. 어휘의 다의성을 해소하기 위해 ③, ④, ⑤, ⑥의 정보를 사용한다.

그림 5는 구축된 개념 계층을 참고하여 개념 계층을 다시 편집하거나 추가하는 기능(③, ④, ⑤, ⑥)과 개념 변동 관리를 위한 2개의 기능(⑦, ⑧)이 준비되어 있다. 각 구성 요소에는 다음과 같은 기능을 한다.

- ① : 입력 어휘 중, EDR과 매칭 하지 않은 어휘 리스트(개념에 추가해야 할 어휘가 있기 때문에)
 - ② : 개념 계층 구축 결과를 트리 구조(tree structure)로 표현
 - ③ : ②에서 개념을 선택한 어휘의 일본어 표제어
 - ④ : ②에서 개념을 선택한 어휘의 영어 표제어
 - ⑤ : ②에서 개념을 선택한 어휘의 일본어 설명
 - ⑥ : ②에서 개념을 선택한 어휘의 영어 설명
 - ⑦ : 조합 결과 분석의 결과
- 조합 결과로부터 재이용 가능한 영역과 불가능한(개념 변동이 발생하고 있다고 추정되는)영역으로 분할하고, 재이용 불가능한 영역을 이동하는 것에 의하여 개념 변동을 해소한다.
- ⑧ : 전정 결과 분석

초기 모델에 있어서 동일한 부모 노드(상위 개념)를 갖는 형제 노드(node)사이의 전정 작업에 있어서, 제거된 중간 개념수의 차이가 큰 경우에 그 계층 관계를 재구성하도록 제시한다.

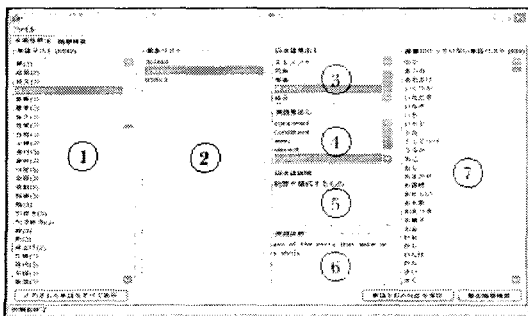


그림 4. 일본어 DODDLE 어휘입력

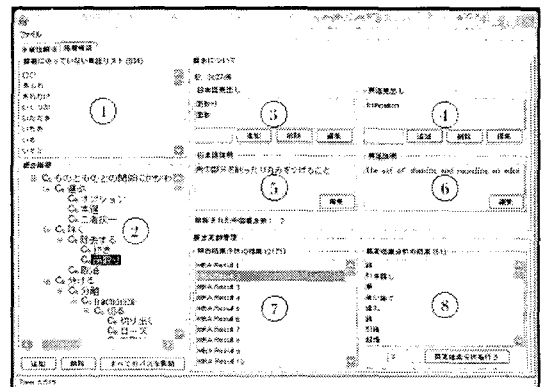


그림 5. 일본어 DODDLE 개념계층구축

3.4 문서검색시스템 : gxfinder

(㉞)Galaxy Express(GE)의 문서 검색 시스템인 gxfinder의 인터페이스 화면은 그림 6과 같다. gxfinder를 이용하여 검색을 수행할 때 사용 가능한 옵션으로는 동의어 및 유의어가 있다.

동의어 옵션은 EDR전자사전의 특징인 1개의 개념에 여러 개의 표제어가 기술되어 있는 것을 이용하는 옵션이다. 이 표제어를 검색 키워드로 전개하고 싶은 경우에 동의어 옵션을 사용하면 된다. 유의어 옵션은 입력된 키워드의 상위 개념 표제어와 입력 키워드의 형제 개념의 표제어를 검색 키워드로 전개하는 옵션이다.

4. 사례 연구 및 평가

본 논문에서 제안된 시스템의 평가를 위해서 사례 연구법을 적용하였다. 일본어 DODDLE에 의하여 구축된 초기 영역 온토로지(OWL 형식)를 (㉞)GE의 문서 검색 시스템에 탑재하여 gxfinder로 검색실험을 하였다.

4.1 단어추출과 온토로지 구축

초기 영역 온토로지를 구축하기 위한 입력으로서, (㉞)GE가 소유한 10,500개의 문서를 茶釜(Chasen)이라는 형태소 해석기를 이용하여 키워드를 추출하였다. 추출된 키워드 결과는 표 1과 같다.

회사가 소유하고 있는 문서에서 추출된 키워드에는 불필요한 단어도 많이 포함되어 있다. 그래서 온토로지 구축을 위한 전처리 과정으로서 반각 문자는 전각 문자로, 동사는 기본형으로 변환, 기호문자는 삭제, 1문자 단어도 삭제하여 입력하였다.

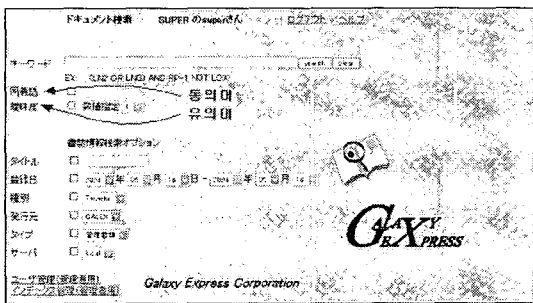


그림 6. gxfinder 인터페이스

표 1. 키워드 추출결과

일반명사	6,926語
그 외 명사	6,191語
동사	3,617語
미지어	6,211語
합계	22,945語

그림 7은 초기 영역 온토로지의 구축 과정에서 나타난 단어수의 변화를 보여준다.

그림 7에서 “16,734” 이라는 숫자는 대상 문서에서 추출한 “일반 명사+그 외 명사+동사”의 합계다. 이번 실험에서는 초기 영역 온토로지를 작성하는데 있어서 필요한 다의성 해소 작업을 수동으로 하였다. 6,226단어의 다의성을 해소하는데 약 15시간이 소요되었다. 이 시간이 초기 온토로지 구축과 관련된 비용이라고 할 수 있다.

그림 8은 10,301단어의 다의성 분포 그래프다. 그래프의 X축은 입력 단어와 동일한 표제어가 포함되어 있는 개념 수로 “다의성(多義性)”이라고 한다. Y축은 다의성의 수가 같은 입력 단어들의 수다. 다의성의 최저 값은 “1”이고 최고 값은 “46”으로 나타났다.

그림 9는 온토로지를 구축할 때 나타난 개념수의 변화를 보여준다. EDR전자사전의 개념 419,570개와 입력 단어 10,301개를 시스템에 입력하여 비교한 결과, 8,324개의 개념이 출력되었다.

8,324개의 개념으로부터 개념 계층을 구축하면 11,371개의 개념이 되었고, 불필요 중간 개념 전정한 후에는 10,177개의 개념이 되었다. 8,324개의 개념으로부터 10,177개의 개념으로 증가한 이유는 개념 계

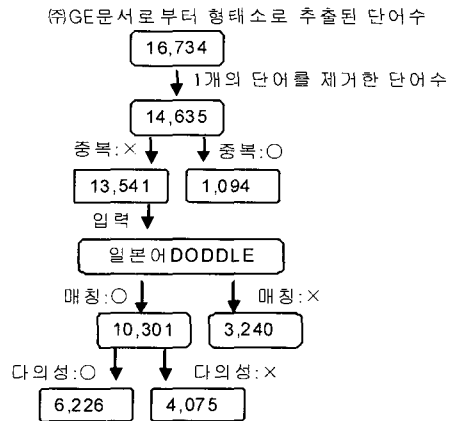


그림 7. 단어수의 변화

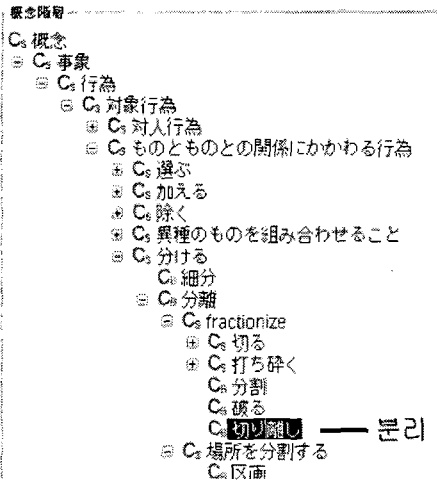


그림 11. "분리" 주변의 개념 계층

③ 키워드 : MST AND 피난

옵션이 없는 경우에는 1개의 문서가 발견되었다. 동의어·매매도 옵션을 이용한 경우에는 "피난"이라는 키워드가 "대피"라고 전개되어 검색 결과가 156건으로 증가한 것은 좋았다고 판단되었다. 그림 12는 "피난"이라는 키워드 주변의 개념 계층을 보여준다.

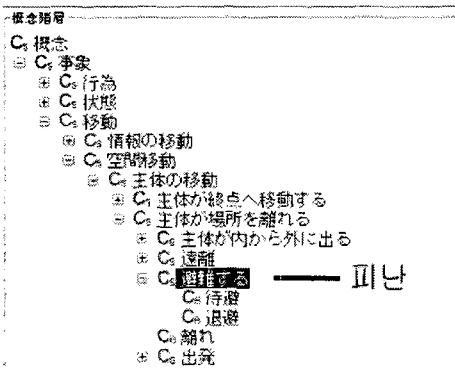


그림 12. "피난" 주변의 개념 계층

5. 결 론

영역 온토로지 구축과 검색 시스템 결과의 향상을 위하여 일본어 DODDLE이라는 툴을 사용한 초기 온토로지를 구축 후, 문서 검색 시스템과의 연동으로 범용 온토로지로 진화시켜 전문 분야의 검색에 특화된 영역 온토로지로 바뀌어가는 본 논문의 제안 방법은 아직 초기 단계이지만 제안된 시스템을 통하여 유효

성을 보였다.

이번의 실험에서는 키워드를 직접 입력하였지만 차후에는 자연언어문장의 입력을 통해서 문장을 입력하고 자동으로 키워드를 추출할 예정(현재로서는 명사와 동사만 추출할 예정)이다. 그리고 추출된 키워드가 문서 검색 시스템인 gxfinder에 쉽게 입력될 수 있도록 시스템을 확장할 예정이다.

그리고 4절에서 언급한 다의성을 수동으로 해소하는 것이 아니라 자동화할 수 있는 전략 및 알고리즘에 관한 새로운 아이디어를 통해서 초기 영역 온토로지 구축 비용을 줄일 예정이다. 그리고 검색 실험을 반복하고, 영역 온토로지에서의 검색 결과를 반영하는 기능 추가 과정을 통해서 본 논문의 최종목적인 영역 온토로지 구축 방법과 검색 시스템의 결과를 향상시키는 방법의 유효성을 보다 상세히 평가할 예정이다.

6. 감 사

사례 연구법의 실험과 평가에 협력하여 주신(주) Galaxy Express의 川村正則씨에게 깊게 감사드립니다.

참 고 문 헌

- [1] 미조구찌 리이치로, 온토로지공학, 인공 지능학회(편), (주)오음사, 도쿄, 2005.
- [2] Michael K. Smith, Chris Welty, and Deborah L. McGuinness, OWL Web Ontology Language Guide, <http://www.w3c.org/TR/owl-guide/>
- [3] Miller, G.A, "WordNet: A Lexical Database for English," *Communications of the ACM*, Vol. 38, No. 11, pp. 39-41, 1995, <http://www.cogsci.princeton.edu/~wn/>
- [4] 일본 전자화 사전 연구소, EDR 전자화 사전(제2판) 사양 설명서, TR2-006(개), 도쿄, 2001.
- [5] T.Morita, Y.Shigeta, N.Sugiura, N.Fukuta, N.Izumi, and T.Yamaguchi, "DODDLE-OWL: OWL-based Semi-Automatic Ontology Development Environment, Evaluation of Ontology-based Tools," *EON2004*, 도쿄, 2004.
- [6] 미조구찌 리이치로, 세만틱웹과 온토로지공학, 인공지능학회, 제20회 AI 신포지엄, pp. 1-28.

2005.

- [7] 마츠모토 유지, 키타우치 아키라, 야마시타 타츠오, 히라노 요시타카, 마즈다 히로시, 타카오 카 카즈마, 아사하라 마사유키, 일본어형태소해석시스템[茶釜] version2.2.1 사용설명서, 나라선단과학기술대학원대학 마츠모토연구실, 나라, 2000.
- [8] 키타무라 요시노부, 공학 도메인 온토로지, 일본인공지능학회, Vol. 19, No. 2, pp. 179-186, 2004.
- [9] 타카이 타카코, 생물학의 기능개념 온토로지, 일본인공지능학회, Vol. 19, No. 2, pp. 137-143, 2004.
- [10] 쿠레마츠 마사키, 야마구치 타가히라, 법률지식의 체계적 정의로써의 법률온토로지, 일본인공지능학회, Vol. 19, No. 2, pp. 144-150, 2004.



홍 윤 기

2000년 동의대학교 컴퓨터공학과 졸업
 2003년 일본게이오대학 이공학 연구과 개방환경과학 사회정보시스템공학 전수석사과정 수료 (2001-2003)
 2003년 동대학 전수박사과정 재학 중

일본인공지능학회 회원
 관심 분야: 자연어처리, 온토로지 구축



김 태 석

1992년 일본KEIO대학 이공학부 계산기과학전공공학박사
 1992년 3월 일본 KEIO대학 이공학부 객원연구원
 1993년 3월~현재 동의대학교 컴퓨터소프트웨어공학과 교수

2000년 3월~2003년 7월 동의대학교 전산정보원장
 2000년 3월~2003년 9월 (재)부산테크노파크 운영위원
 2003년 8월~2005년 12 동의대학교 교무처장
 관심 분야: 인터넷응용, 원격강의, 자연어처리



山口 高平

1979년 오오사카대학 공학부 통신공학과 졸업
 1984년 오오사카대학 대학원 공학연구과 공학박사
 1984년 오오사카대학 산업과학연구소 조교
 1989년 시즈오카대학 공학부 조교수

현재: 게이오대학 이공학부 관리공학과 교수
 1991, 1998년도 인공지능학회전국대회우수논문상.
 일본전자정보통신학회, 일본정보처리학회, 일본인지과학회, AAAI, IEEE-CS, ACM 각회원
 관심 분야: 자연어 처리, 온토로지 구축