의미적 유사성에 기반한 온톨로지 선택 랭킹 모델

Ontology Selection Ranking Model based on Semantic Similarity Approach

오선주(Sunju Oh)*, 안중호(Joongho Ahn)**, 박진수(Jinsoo Park)***

초 록

지식 재사용 측면에서 기존의 온톨로지를 재사용할 수 있다면 많은 자원을 절약할 수 있을 것이다. 그러나 기존의 온톨로지를 활용하기 위해서는 보다 발전된 온톨로지 검색 기능이 요구된다. 현재까지 이루어진 관련 연구들에서는 주로 렉시컬 매칭기법을 사용하여 온톨로지를 검색하였다. 그러나 의미적 측면에서 문제점이 있으므로 본 연구에서는 관계의 의미적 유사성에 기반한 온톨로지 선택 랭킹 모델을 제안한다. 본 연구는 개념간 계층 구조와 관계를 온톨로지 검색에 이용함으로써 온톨로지의 선택 랭킹을 효과적이며 실질적으로 개선하였다. 또한 실험을 통해 연구 모델의 결과와 선행 연구의 결과, 온톨로지 전문가의 랭킹 결과를 비교 분석하고 연구 모델의 타당성을 검증하였다. 본 연구 결과는 온톨로지 검색 연구를 이론적으로 발전시켰을 뿐 아니라 실무적인 측면에서 실무자들이 온톨로지를 쉽게 찾아 재사용할 수 있도록 한다.

ABSTRACT

Ontologies have provided supports in integrating heterogeneous and distributed information. More and more ontologies and tools have been developed in various domains. However, building ontologies requires much time and effort. Therefore, ontologies need to be shared and reused among users. Specifically, finding the desired ontology from an ontology repository will benefit users. In the past, most of the studies on retrieving and ranking ontologies have mainly focused on lexical level supports. In those cases, it is impossible to find an ontology that includes concepts that users want to use at the semantic level. Most ontology libraries and ontology search engines have not provided semantic matching capability. Retrieving an ontology that users want to use requires a new ontology selection and ranking mechanism based on semantic similarity matching.

We propose an ontology selection and ranking model consisting of selection criteria and metrics which are enhanced in semantic matching capabilities. The model we propose presents two novel features different from the previous research models. First, it enhances the ontology selection and ranking method practically and effectively by enabling semantic matching of taxonomy or relational linkage between concepts. Second, it identifies what measures should be used to rank ontologies in the given context and what weight should be assigned to each selection measure.

키워드 : 온톨로지, 온톨로지 선택, 온톨로지 재사용, 의미적 유사성, 온톨로지 평가, 온톨로지 랭킹
Ontology, Ontology Selection, Ontology Reuse, Semantic Similarity, Ontology Evaluation, Ontology Ranking

---

* 서울대학교 경영대학 박사과정
** 서울대학교 경영전문대학원 경영대학 교수
*** 교신저자, 서울대학교 경영전문대학원 경영대학 부교수
2009년 03월 11일 접수, 2009년 05월 07일 심사완료 후 2009년 05월 11일 게재확정.

1. 서 론

인터넷 혁명은 다양한 형태의 정보와 지식을 시간과 장소에 관계없이 공유하게 함으로써 인간에게 생활의 편리함을 제공하여 주었다. 그러나 이와 같은 웹 기술의 발달이 인간이 추구하는 지적 욕구를 완전히 충족시켜주지는 못하는 실정이다. 특히, 인터넷 검색을 통해 유용한 정보를 쉽게 찾을 수 있지만 의미적 일치성 측면에서 정확도는 떨어지고 의미상 관련성이 떨어지는 정보를 본래의 의도와 관계 없이 받아보는 상황에 있다. 웹 검색에서의 이와 같은 문제점은 온톨로지 사용을 통해 해결될 수 있다.

온톨로지는 정보의 의미까지 체계적으로 공유할 수 있도록 하는 정보 공유의 인프라를 제공해 준다[8]. 점점 더 많은 온톨로지들이 현재 개발 중에 있으며 온토링구아(Ontolingua), DAML 온톨로지 라이브러리, 프로티지(Protégé) 라이브러리와 같이 이미 개발된 온톨로지 라이브러리 등도 다수 존재한다. 온톨로지를 새롭게 개발하는 작업은 많은 시간과 노력을 필요로 하는 작업이므로 기존의 온톨로지를 재사용 할 수 있다면 자원의 활용 측면에서 도움이 될 것이다. 더 나아가 이질적 시스템과 자원들이 온톨로지를 공유함으로써 쉽게 통합되고 상호운영성이 높아질 수 있다. 그러나 대부분의 현존하는 온톨로지 라이브러리들은 온톨로지 검색 기능이 미흡하여 사용자들이 필요로 하는 개념 혹은 개념들간의 관계를 포함한 온톨로지를 찾는 것이 어려운 실정이다. 또한 재사용할 수 있는 온톨로지 수가 증가함에 따라 이와 같은 어려움은 더욱 심화될 것으로 예상되고 있다[1].

그와 같은 현상을 개선하기 위해 두 가지 핵심 기술의 발전이 필요하다. 그 중 한 측면은 온톨로지를 사용자의 용도에 맞게 찾아 선택하여 사용할 수 있는 환경을 만드는 것이다. 또 다른 측면은 시맨틱 매칭(semantic matching) 기술을 발전시키고 이 기술을 온톨로지 선택에 적용하는 것이다.

본 논문의 목적은 사용자의 요구사항에 적합한 온톨로지를 선택할 수 있는 프레임워크를 제안하는 것이다. 특히, 개념간 관계의 의미적 유사성에 기반한 온톨로지 선택 모델을 설계하고자 한다. 이러한 목표를 달성하기 위해서 다음과 같은 질문에 대한 답을 얻고자 하였다.

- 사용자 요구 사항에 적합한 온톨로지를 어떻게 선택하고 랭킹(ranking)할 것인가?
- 주어진 요구사항에 맞는 온톨로지를 선택하기 위한 기준은 무엇인가?
- 선택 기준을 어떻게 측정할 것인가?

본 연구에서는 온톨로지 선택 기준을 설계하고 온톨로지 랭킹을 위한 새로운 방법을 개발하였다. 많은 선행연구에서 의미적 유사성에 관하여 연구하였지만[7, 9, 14, 15, 18] 의미적 유사성 이론을 온톨로지 선택에 적용한 예는 찾아 보기 어렵다. 본 연구에서는 의미적 유사성에 기반하여 개념 사이의 관계를 식별함으로써 온톨로지를 선택하는 방법을 새로이 설계하였다. 또한 실험을 통한 실증적 방법을 통해 새로 설계한 방법의 타당성을 검증하였다.

본 연구의 구성은 다음과 같다. 제 2장에서

는 온톨로지 선택과 랭킹에 관련된 선행 연구들을 살펴보며, 제 3장에서는 의미적 유사성에 근거하여 온톨로지를 선택하고 랭킹하는 프레임워크를 제안한다. 제 4장에서는 설계한 방법의 효과를 입증하기 위해 선택 방법의 프로토타입을 구현하여 실험한 내용을 논의한다. 제 5장에서는 본 연구의 선택 알고리즘과 선행 연구의 알고리즘, 그리고 온톨로지 전문가의 평가 결과를 상호 비교한 결과를 보여준다. 제 6장에서는 연구의 시사점을 제시하며 끝으로 제 7장에서는 결론과 향후 과제를 논의한다.

2. 관련 연구들

이 장에서는 본 연구와 관련된 온톨로지 검색, 랭킹에 관한 기존 연구들을 간략히 살펴본다.

2.1 온톨로지 검색 및 랭킹 관련 연구

온톨로지 검색 및 랭킹에 관한 활발한 연구가 진행되어 왔다[1, 2, 4, 10, 11]. 이러한 연구들 중 대표적인 연구에 대해 온톨로지 선택 기준을 중심으로 살펴본다.

- **Swoogle**

Swoogle[4]은 RDF, OWL로 작성된 웹 문서들을 인덱스화하여 트리플(triple) 저장소에 저장하고 있는 시맨틱 웹 검색엔진으로, 현재 11,000여 개의 온톨로지를 포함하고 있다. Swoogle은 구글(Google) 검색엔진의 페이지랭크(PageRank) 알고리즘과 유사한 방식으로 온톨로지간의 링크 수에 기반하여 온톨로지 검색의 우선 순위를 결정한다.

- **OntoSelect**

대부분의 온톨로지 라이브러리들은 정적으로 온톨로지들을 등록하도록 설계되어 있는 반면 OntoSelect[2]는 웹을 모니터링하고 동적으로 온톨로지를 등록하는 라이브러리이다. OntoSelect는 커버리지, 구조, 연결성 등 세 가지 검색 기준에 따라 온톨로지를 검색하는 기능을 제공한다.

- **OntoKhoj**

OntoKhoj[11]는 온톨로지를 검색하고 랭킹할 뿐만 아니라, 이들을 분류하고 웹 크롤(crawl) 기능을 제공하는 시맨틱 웹 포털이다. 구글의 페이지랭크와 유사한 OntoRank 알고리즘을 온톨로지에 적용하는데, 이는 서로 다른 링크에 가중치를 부여하는 등 좀 더 확장된 기능을 사용한다.

- **OntoSearch2**

OntoSearch2[10]는 웹 상에서 온톨로지를 검색하고 질의어를 처리하는 시맨틱 웹 검색엔진이다. Ontosearch2의 핵심은 DL-Lite언어를 위한 추론엔진으로 SPARQL 질의 언어를 사용하여 온톨로지를 검색하고 질의를 처리하는 프레임워크를 제공한다.

- **AKTiveRank**

AKTiveRank[1]는 각각의 온톨로지에 대해 CMM(Class Match Measure), CEM (CEntrality Measure), SSM(Semantic Similarity Measure) 그리고 DEM(DEnsity Meas-

ure)의 네 가지 측정 방법을 적용하여 값을 계산하고 이들 값에 적절한 가중치를 부여하여 합산한 총 합계 값으로 온톨로지의 랭킹을 매기는 방식을 적용하였다. CMM은 검색어가 온톨로지 내의 클래스들과 얼마나 잘 일치하는가를 나타내며, CEM은 찾고자 하는 개념이 계층구조의 중간에 위치하는가를 측정한다. 보편적으로 계층구조의 중간에 있는 개념은 최상위 혹은 최하위에 있는 개념에 비해 다른 개념과 연관관계가 많이 있을 수 있으므로 풍부하게 표현된다고 할 수 있다. SSM은 찾고자 하는 개념들이 계층구조상 얼마나 인접하게 위치하는가를 나타내준다. DEM은 개념들이 얼마나 풍부하게 잘 정의되어 있는가를 나타내주는 측정방식으로 개념의 부모, 자식, 그리고 연관된 개념의 수를 합산하여 계산한다. 본 연구에서는 AKTiveRank에서 정의한 이들 측정 방식을 도입하여 사용한다.

2.2 의미적 유사성 관련 연구

Rada et al.[14]은 개체(entity)간의 의미적 유사성을 의미적 간격(semantic distance)을 이용하여 측정하였다. Resnik[15]은 계층 구조에서 개체간의 유사성이 개체간 공유되는 개념의 정보 내용에 따라 결정된다고 보았다. Lin[9]은 공유하는 부모 노드와 검색어의 정보 콘텐츠를 모두 고려하였으며 Jang et al.[7]은 계층구조와 함께 의미적 간격을 고려하는 방법을 설계하였다.

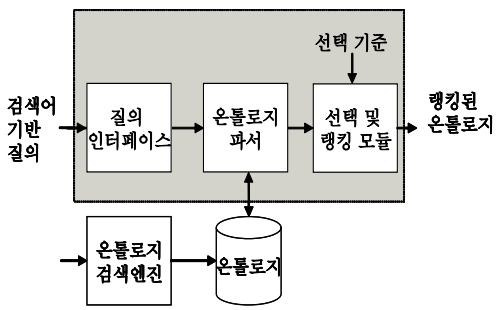
한편 Turney[18]는 개념의 구분보다 개념들간의 관계를 구분하는 것이 유사성 측정을 위해 보다 중요하다고 주장하며 관계적 유사성(relational similarity)을 강조하였다. 그는 개념끼리 유사한 정도를 의미하는 속성적 유사성(attributional similarity)이 의미적 유사성이란 용어로 잘못 사용되어 왔다고 주장하였다.

3. 연구 모델

본 장에서는 사용자의 의도에 보다 적합하게 온톨로지를 선택할 수 있는 온톨로지 선택 및 랭킹 모델을 제시하고자 한다. 즉, 사용자가 필요로 하는 개념과 개념간의 관계를 보유하였는지를 척도로 온톨로지를 선택할 수 있도록 함으로써 온톨로지 검색에 있어 의미적 측면을 강화하고 정확도를 높일 수 있다. 우선, 제안하는 연구 모델의 전체적 아키텍처를 제시하고, 모델의 주요 사상인 시맨틱 매칭(semantic matching)이론, 온톨로지 선택 기준과 측정 방식에 대해 기술한다.

3.1 연구 모델의 아키텍처

본 연구의 온톨로지 선택 및 랭킹 모델은 질의 인터페이스, 온톨로지 파서(parser), 선



〈그림 1〉 모델 아키텍처

택 및 랭킹 모듈로 구성된다(<그림 1> 모델 아키텍처 참조).

모델의 주요 구성요소의 하나인 질의 인터페이스는 온톨로지 사용자나 에이전트로부터 요구사항을 받아들여 키워드 기반 질의를 생성한다. 키워드 기반 질의는 검색하고자 하는 개념어를 찾아주는 검색 방식이다. 연구 모델에서는 개념어뿐 아니라 개념간 관계어도 검색할 수 있다. 사용자가 입력한 개념어 혹은 관계어로 온톨로지를 검색하여 검색어와 완전 일치 또는 일부분 일치하는 개념을 가진 온톨로지를 찾아 준다. 본 연구 모델은 온톨로지 검색엔진을 이용하여 웹에 산재되어 있는 온톨로지들을 검색하여 미리 온톨로지 저장소(repository)를 구축해 놓는다. 이때, 온톨로지 검색엔진은 이미 공개적으로 사용되고 있는 Swoogle[4]을 이용한다.

온톨로지 파서는 온톨로지 구문을 읽어 토큰화하고 내부적인 데이터 구조로 전환하여 준다. 선택 및 랭킹 모듈은 선택 기준에 따라 온톨로지들의 순위를 계산한다. 본 연구에서 사용하는 온톨로지 선택 기준은 검색하고자 하는 개념과 관계를 가장 많이, 가장 근접하게 포함하고 있는 온톨로지를 찾는 것이다. 그리고 찾고자 하는 개념이 관련 개념을 통해 가장 자세히, 풍부하게 표현되어 있는 것을 찾는 것이다. 각 선택기준에 대해서는 제 3.3.1절에 자세히 설명한다.

3.2 의미적 유사성(Semantic Similarity)을 기반으로 한 온톨로지 선택

온톨로지 선택 및 랭킹이란 로컬 서버에 있는 온톨로지 저장소를 대상으로 검색어와 의미적으로 유사한 개념을 가지는 온톨로지를 찾아내어 순위를 매기는 것이라고 정의한다. 본 절에서는 기존 온톨로지 선택 모델의 문제점을 살펴보고 이를 해결하기 위해 시맨틱 매칭기법을 온톨로지 선택에 적용한다.

3.2.1 기존 온톨로지 선택의 문제점

기존 온톨로지 선택 모델의 큰 문제 중 하나는 온톨로지 검색의 매칭 수준(level)이다. 기존의 온톨로지 선택에 대한 연구에서는 렉시컬(lexical) 수준의 키워드 매칭에 주로 의존하였다. 그러나 이 방법을 사용했을 경우 몇 가지 문제점이 발생할 수 있다. 첫째, 의미적으로 유사한 개념을 찾는 것이 어렵다. 즉, '친구'와 '동무'는 같은 개념임에도 불구하고 '친구'로 '동무'를 검색할 수 없다. 둘째, 다의어의 경우, 서로 다른 의미로 사용되었을 경우에도 구분이 안 된다. 예를 들면 '사과'는 먹는 사과가 있고 용서를 구하는 사과가 있지만 문맥 없이는 이를 구분하기 힘들다. 셋째, 개념간의 관계를 찾을 수 없다. 예를 들면 '과일'과 '색상'과의 관계를 이용하여 '빨간 사과'를 찾을 수 없었다.

그러므로 이와 같은 문제점을 해결하기 위해 첫째, 검색어의 동의어와 다의어를 처리할 수 있는 기능과 둘째, 단순히 렉시컬 수준의 매칭뿐만 아니라 검색어의 의미적인 차이도 구분할 수 있는 개념간 관계를 이용한 매칭기법이 효과적이라 할 수 있다.

3.2.2 의미적 유사성을 기반으로 한 시맨틱 매칭

기존 온톨로지 선택기법의 단점을 개선하기 위해 본 연구에서는 의미적 유사성에 기

반한 시맨틱 매칭기법을 사용한다. 의미적 유사성은 지식검색과 시스템 통합과 관련된 연구에서 많이 논의되어 왔으나 그 의미와 측정 방식이 상이한 측면이 있다.

본 연구에서는 제 2.2절에서 살펴본 바와 같이 기존의 연구에서 정의하고 있는 의미적 유사성의 뜻을 검토하고 이를 기반으로 의미적 유사성을 새로이 정의하고자 한다. 앞서 Turney 연구에서 의미적 유사성의 불완전성이 지적되었으며, 이전의 연구들에서 속성적 유사성이 의미적 유사성으로 인식되어 왔고 관계적 유사성이 무시되어 왔다[18]. 그러므로 본 연구의 의미적 유사성 정의에는 속성적 유사성뿐 아니라 관계를 고려하는 유사성이 추가되어야 한다. 즉, 온톨로지 내의 특정 개념이 의미적으로 유사하려면 개념의 속성이 유사해야 할 뿐만 아니라 개념이 다른 개념과 형성하는 관계도 유사하여야 한다. 그러므로 의미적 유사성은 다음과 같이 개념(속성) 유사성, 관계 유사성과 계층관계 유사성의 통합으로 정의할 수 있다.

$$Sim(o,\ T,\ r) = \lambda_1 Sim_c(o,\ T) + \lambda_2 Sim_r(o,\ T,\ r) + \lambda_3 Sim_{taxo}(o,\ T),$$

이때 $\lambda_1 + \lambda_2 + \lambda_3 = 1$이며, $T = \{t_1, \cdots t_n\}$이고 $R = \{r, t_i, t_j\}$이다. λ_i는 각 유사성에 대한 가중치이며 변동 가능한 값이다. T는 검색어 집합이며 관계 매칭을 위해서는 적어도 두 개 이상의 검색어를 포함하여야 한다. t_i와 t_j는 검색어이며 r은 관계 R의 명칭(label)이다.

즉, 본 연구에서는 관계의 의미적 유사성에 비중을 두고 깊이 있는 연구를 하기 위해 관계적 유사성을 세분화하여 일반적 관계를 다루는 관계 유사성 $Sim_r(o,\ T,\ r)$과 IS-A 관계를 다루는 계층관계 유사성 $Sim_{taxo}(o,\ T)$으로 구분하였다. 본 절에서는 의미적 유사성을 구성하는 개념 유사성, 관계 유사성, 그리고 계층관계 유사성을 측정하기 위해 각각의 매칭 방법에 대해 설명한다.

또한 본 연구에서 사용하는 온톨로지 o, 검색어 t_i의 집합 T, 관계 R과 관계의 명칭 r 등의 기호들은 이하 동일한 의미로 사용한다.

3.2.2.1 개념 매칭(Concept Matching)

속성적 유사성, 즉 개념 유사성은 완전 일치에 의한 매칭 혹은 부분 일치에 의한 매칭을 모두 반영하여 유사성을 측정하는 방식이다. 전체 매칭 수에 대한 완전 일치되는 매칭 수의 비율로 나타내진다.

$$Sim_c(o,\ T) = \frac{|T \cap X|}{|T \cap X| + \alpha |X - T|}$$

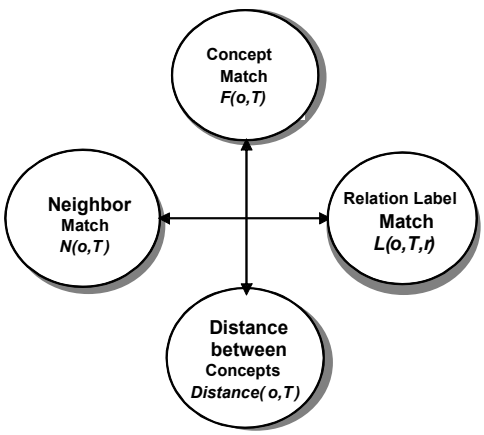
$0 \le \alpha \le 1$이고 T는 검색어 집합이며, X는 온톨로지 o에서 검색어에 대응되는 개념들의 집합이다. α는 부분 매칭의 가중치 함수로서 $\alpha(X) = 1/(1 + depth(X))$로 구해진다. -는 차집합이며 | |는 관계차수(cardinality)를 의미한다.

3.2.2.2 관계 매칭(Relation Matching)

개념 사이의 관계를 측정하는 연구들은 활발히 이루어졌으나[6, 12, 13, 18] 온톨로지 선택에 있어 개념 사이의 관계를 적용한 연구 시도는 없었다. 본 연구는 관계의 의미적 유사성을 적용한 온톨로지 검색을 통해 온톨

로지 선택 및 랭킹을 보다 효과적으로 할 수 있는 방안을 제시한다. 즉, 의미적 유사성의 측정지표인 RMM(Relation Match Measure)을 설계하고 적용하는 실험을 수행한다. RMM은 본 절의 뒷 부분에 있는 관계일치도에서 정의가 자세히 되어 있다. 실험 결과를 통해 어떤 온톨로지가 사용자가 제시한 개념과 관계를 잘 표현하고 있는 지 보여준다.

관계의 의미적 유사성은 <그림 2>의 네 가지 요인에 의해 영향을 받는다. 네 가지 영향 요인의 구체화를 위해 $F(o, T)$, $L(o, T, r)$, $N(o, T)$, $Distance(o, T)$ 등 네 개의 매치 함수를 설계하였다.

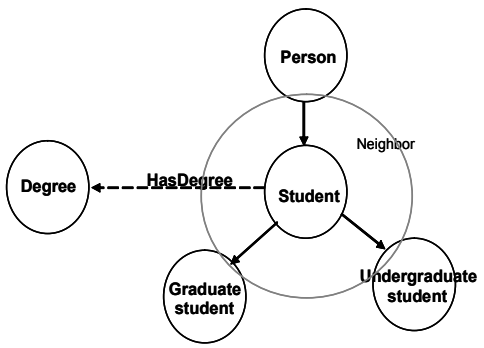


〈그림 2〉 관계의 의미적 유사성의 영향요인

개념매치(Concept Match)인 $F(o, T)$는 개념 사이의 대응도를 나타내며 $Sim_c(o, T)$로 측정된다. 즉, 관계를 구성하는 도메인(domain)과 레인지(range) 개념이 정확히 혹은 부분적으로 일치하는지에 따라 값이 결정되는 함수이다.

관계명칭매치(Relation Label Match)인 $L(o, T, r)$은 관계의 의미를 나타내는 명칭 r의 일치 정도를 나타내며 $Sim_e(o, c_i, c_j, r)$로 측정된다. 이는 검색어간의 관계가 개념 c_i와 c_j간의 관계와 의미적으로 어느 정도 일치하는지를 나타내준다. 온톨로지내 개념을 나타내는 기호 c_i, c_j는 이하 동일한 의미로 사용한다.

이웃매치(Neighbor Match)인 $N(o, T)$은 관계의 도메인과 레인지가 정확하게 일치하지 않더라도 도메인과 레인지 개념의 인근 노드와 관계가 일치하면 매치를 인정할 수 있도록 반영하기 위한 함수이다. 다시 말해, 관계의 의미적 매치를 좀 더 확장 해석하여 인근 노드까지 확장하는 것이다. 또한 $N(o, T)$는 $Sim_n(o, c_i)$로 측정된다. 이웃매치의 예는 <그림 3>에 나타나 있다. <그림 3>에서 Person과 Degree간의 관계를 검색한다고 가정하면, 두 노드간 아무 관계가 없지만 Person의 이웃인 Student가 Degree와 관계를 가진다. 이때 원래 검색하려는 관계는 Student와 Degree간의 관계와 이웃매치 된다.



〈그림 3〉 이웃매치(Neighbor Match)

개념간 거리 $Distance(o, T)$는 관계를 구성하는 노드들 사이의 링크 수를 기준으로 간격을 구하는 함수이다. 두 개념간의 간격

이 관계의 의미적 유사성과 반비례하는 특성이 있다.

사용자가 찾고 싶어하는 개념과 개념들간의 관계를 온톨로지가 보유하였는지는 관계의 의미적 유사성에 따라 결정되는데 다음에 정의된 RMM, Taxo에 의해 측정될 수 있다.

- **관계일치도(Relation Match Measure : RMM)**

RMM은 검색하려고 하는 검색어들 사이의 관계와 온톨로지 내의 개념과 개념 사이의 관계와의 의미적 유사도를 계산하는 측정 방식이다. RMM은 찾고자 하는 개념을 단순히 문자 그대로만 찾는 한계에서 벗어나 다른 개념과 형성하고 있는 관계를 검사함으로써 검색하고자 하는 개념에 대한 의미적인 측면을 보완할 수 있다. RMM은 검색하려고 하는 관계를 구성하는 도메인 개념과 레인지 개념뿐 아니라 관계의 명칭까지 일치하는지 검사한다. 또한 찾고자 하는 관계를 구성하는 개념 사이의 간격에 반비례하는 특성을 가진다.

- **정의 1-RMM**

검색어 t_1, t_2 사이의 관계 R_s는 R_s = <r, t_1, t_2>로 나타내지며 온톨로지내 도메인 개념 c_d와 레인지 개념 c_r 사이의 관계 R_c는 R_c = <r, c_d, c_r>로 나타내진다. 이때, 주어진 온톨로지 o, 검색어 T = {t_1, t_2}, 관계 R_s, R_c의 관계 명칭 r에 대해 RMM은 $Sim_r(o, T, r)$로 다음과 같이 정의할 수 있다.

$$RMM[o, t_1, t_2, r] = Sim_r(o, T, r) = \sum_d \sum_r Sim_r(o, c_d, c_r, r)$$

$$Sim_r(o, c_d, c_r, r) = \frac{(F(o, c_d) - N(o, c_d))(F(o, c_r) - N(o, c_r))(1 + L(o, c_d, c_r, r))}{Distance(o, c_d, c_r)}$$

$$= \left[\frac{Sim_c(o, c_d) \times Sim_c(o, c_r)}{Distance(o, c_d, c_r)} - \frac{Sim_c(o, c_d) \times Sim_n(o, c_r)}{Distance(o, c_d, c_r)} - \frac{Sim_n(o, c_d) \times Sim_c(o, c_r)}{Distance(o, c_d, c_r)} + \frac{Sim_n(o, c_d) \times Sim_n(o, c_r)}{Distance(o, c_d, c_r)} \right] \times (1 + Sim_e(o, c_d, c_r, r))$$

$$Sim_n(o, c_i) = \frac{1}{(1 + number\ of\ concepts\ in\ neighborhood\ of\ c_i)}$$

$$Sim_e(o, c_d, c_r, r) = \begin{cases} 1 & c_d, c_r\text{사이에 관계가 있고 관계 명칭이 } r\text{과 완전 매칭되는 경우} \\ 0 & \text{그외의 경우} \end{cases}$$

$Distance(o, c_d, c_r)$ = c_d와 c_r 사이의 최소 링크수 ■

3.2.2.3 계층관계 매칭(Taxonomy Matching)

찾고자 하는 개념과 개념들 사이의 종적관계(IS-A)를 검색하기 위해서 검색 대상 온톨로지들이 검색 개념과 개념들간 관계를 보유하였는지 측정할 수 있는 방식이 필요하다. 본 장에서는 개념과 개념 사이의 종적 관계의 의미적 유사성의 정도에 따라 온톨로지를 선택할 수 있도록 종적 관계의 의미적 유사성을 측정하는 방법을 아래와 같이 정의한다.

- **계층관계일치도(Taxonomy Match Measure : Taxo)**

Taxo는 트리 구조상에서 개념과 개념 사이의 종적관계의 일치 정도를 나타낸다. 즉, 검색하려고 하는 검색어들 사이의 종적관계와 온톨로지 내의 개념과 개념 사이의 종적 관계와의 의미적 유사도를 계산한다. 즉, 온톨로지 내에서 검색어와 일치하는 개념을 찾고 개념간의 종적관계 유무를 검사한다. 종

적관계의 이행적(transitive) 특성을 반영하여 조상과 후손의 관계를 종적관계의 확장으로 간주한다. 트리 구조상 노드 간의 링크 수를 의미적 간격으로 정의하면 Taxo값은 이 간격에 반비례하는 특성이 있다.

• **정의 2-Taxo**

주어진 온톨로지 o내 부모 개념 c_p와 자식 개념 c_c, 검색어 $T = \{t_1, t_2\}$에 대해 $Taxo$는 다음과 같이 정의할 수 있다.

$$Taxo[o, t_1, t_2]$$
$$= Sim_{taxo}(o, T) = \frac{F(o, T)^2}{Distance(o, T)}$$
$$= \sum_p \sum_c Sim_{taxo}(o, c_p, c_c)$$
$$= \sum_p \sum_c \frac{Sim_c(o, c_p) \times Sim_c(o, c_c)}{Distance(o, c_p, c_c)}$$

이때, $F(o, T)$와 $Distance(o, T)$는 앞의 정의와 동일하다. ■

3.3 온톨로지 랭킹

온톨로지 랭킹이란 온톨로지 선택기준에 따라 온톨로지의 순위를 매기는 작업을 의미한다. 본 절에서는 온톨로지 선택에 관한 선행 연구를 바탕으로 온톨로지 랭킹을 위한 선택 기준을 설계하고 선택 기준에 따라 온톨로지를 측정하는 방법을 제시한다.

3.3.1 온톨로지 선택 기준(Ontology Selection Criteria)

선택 기준은 다음의 질문과 관련된다.

(1) 검색어가 온톨로지에 정의된 개념과 의미적으로 어느 정도 일치하는가?
(2) 검색어가 온톨로지의 주요 개념인가?
(3) 검색어가 온톨로지 내에서 얼마나 풍부하게 표현되어 있나?

위의 사항을 만족하는 선택 기준으로 개념간 관계의 의미적 유사성(semantic similarity), 토픽 커버리지(topic coverage), 표현밀도(density)를 선정하였다(<표 1> 참조). 그 중에서 개념간 관계의 의미적 유사성을 가장 중요한 기준으로 정하였다. 선행 연구에서는 온톨로지 선택에 있어서 개념간 관계의 유사성은 무시되어 왔으나 개념간 관계의 유사성을 이용한 검색은 검색 결과의 정확도를 높여줄 수 있고[16], 제 3.2.1절의 기존 선택 방식의 단점을 해결할 수 있으므로 중요성이 크다고 할 수 있다.

〈표 1〉 온톨로지 선택 기준

선택 기준	설 명
의미적 유사성	찾고자 하는 검색어가 온톨로지 내의 개념 그리고 개념간의 관계와 의미적으로 유사한가?
토픽 커버리지	검색어가 온톨로지 내의 개념들과 얼마나 잘 일치하는가?
표현밀도	검색어와 일치하는 개념들이 다른 개념들과 연관관계를 가지며 잘 정의되어 있나?

토픽 커버리지는 온톨로지가 포함할 수 있는 지식 도메인의 범위를 나타낸다. 현재까지의 연구에서 토픽 커버리지는 소규모의 온톨로지들을 대상으로 한 온톨로지 선택에 있어 가장 신뢰성 있는 방법으로 알려져 있다[5].

검색어와 일치하는 온톨로지 내부의 개념 어들이 다른 개념들과 관계를 보다 많이 맺고 있을 경우, 개념을 통해 나타내는 정보의 양이 많고 풍부하며 표현밀도가 높다고 할 수 있다.

3.3.2 온톨로지 선택을 위한 측정 방식

온톨로지 선택 기준을 적용하기 위해서 구체적인 측정 방식이 필요하다. 특히, 관계의 의미적 유사성 기준을 위해 본 연구에서 새로이 제안한 측정 방법인 RMM, Taxo를 이용하였다. 그러나 다른 선택 기준인 토픽 커버리지와 표현밀도 기준에 대해서는 이미 선행 연구에서 측정 방식이 개발되어 있기 때문에 새로이 개발하지 않고 선행 연구의 방법을 그대로 사용하였다. Alani et al.[1]의 연구에서 토픽 커버리지 기준을 위해 CMM을, 그리고 표현밀도 기준을 위해 SSM과 DEM을 도입하여 사용하였다(<표 2> 참조).

Alani et al.[1]의 연구에서는 CEM, SSM, DEM 등 세 가지 방식으로 표현밀도를 측정하였다. 그러나 본 연구에서는 CEM은 사용하지 않았다. 왜냐하면 선행 연구에서 다른 선택 기준에 비해 표현밀도가 CEM, SSM, DEM 등 상대적으로 여러 가지 방식으로 측정 되었으므로, 본 연구에서는 DEM과 유사한 CEM을 사용하지 않고 DEM, SSM만을 사용 하였다. 즉, 개념과 관련을 맺는 주변 개념의 수를 세는 DEM은 개념이 온톨로지 계층구조상에 중간 위치에 오는지를 측정하는 CEM과 유사한 결과를 가져오기 때문이다.

〈표 2〉 온톨로지 선택 기준과 측정 방식

선택 기준	측정방식	연구자
의미적 유사성	RMM, Taxo	본 연구에서 제안
토픽 커버리지	CMM	Alani et al.[1]
표현밀도	DEM, SSM	Alani et al.[1]

3.3.3 랭킹(Ranking)

온톨로지 랭킹이란 각 온톨로지에 대해 측정 방식들을 적용하여 값을 계산하고 이들을 합산한 값에 따라 온톨로지의 순위를 매기는 것이다. 합산 값이 클수록 높은 우선 순위를 가지게 된다. 이때, 온톨로지 선택의 목적에 맞게 각 척도의 가중치가 조정될 수 있다. 각 척도에 대한 가중치는 각 척도의 상대적 중요도를 나타낸다. 본 연구에서는 온톨로지 랭킹을 위해 다음과 같은 합산 식을 적용하였다.

- **관계 검색**

Total[o] = α CMM[o, T] + β RMM[o, t_1, t_2, r] + γ SSM[o, T] + δ DEM[o, T].
이때 $\alpha + \beta + \gamma + \delta = 1$이다.

- **계층관계 검색**

Total[o] = α CMM[o, T] + β Taxo[o, t_1, t_2] + γ SSM[o, T] + δ DEM[o, T].
이때 $\alpha + \beta + \gamma + \delta = 1$이다.

4. 실 험

본 장에서는 연구 모델의 타당성과 우수성을 검증하기 위해 실험을 통한 실증적 방법을 이용하였다. 이를 위해 온톨로지 선택을 위한 측정 알고리즘의 프로토타입을 구현하

고 실행하였다. 프로토타입을 실행한 결과는 다음 장에서 온톨로지 전문가의 랭킹 결과와 이전 선행 연구와의 비교를 통해 연구 모델의 타당성을 증명하는데 사용된다.

실험을 위한 방법론으로 실험실 실험 방법을 적용하였는데 실험실 실험 방법은 실험의 주변 환경을 제어할 수 있는 경우 효과적인 방법이다[3].

4.1 실험 설계

본 연구의 실험은 다음과 같이 이루어졌다. 의미적 유사성에 기반한 온톨로지 선택 방법이 타당한지를 검증하기 위해 첫째, 온톨로지내 관계의 의미적 유사성을 측정하는 알고리즘을 구현하고, 각각의 온톨로지들에 알고리즘을 적용하여 값을 계산하였다. 둘째, 각 측정 방식의 가중치를 융통성 있게 조절할 수 있다고 가정하고 각 실험에서 측정 방식의 가중치를 변화시키며 세 가지 유형의 실험을 하였다. 본 연구의 실험은 가상의 온톨로지가 아닌 실제 웹 상에 개방된 온톨로지들을 대상으로 실험을 하였으므로 객관성이 있다고 할 수 있다.

4.1.1 데이터

실험의 대상 온톨로지들을 수집하기 위해 온톨로지 검색엔진인 Swoogle을 이용하여 인터넷상의 온톨로지들을 검색하였다. 대학교육 관련 온톨로지를 구하기 위해 “Student”와 “University” 등의 두 키워드를 사용하여 온톨로지들을 검색하여 그 결과를 저장소에 저장하였다. 검색 결과로 <표 3>에 나타난 바와 같이 aargh.owl, agent.owl, 그리고 akt_ontology_LITE.owl 등 10여 개의 OWL파일을 얻을 수 있었다.

〈표 3〉 실험 대상 온톨로지 목록

온톨로지	온톨로지 이름
A	aargh.owl
B	agent.owl
C	akt_ontology_LITE.owl
D	iswc.owl
E	ka.owl
F	koala.owl
G	semiport.owl
H	swrc.owl
I	univ-bench.owl
J	univ.owl

4.1.2 실험 환경 설정 및 시나리오

실험 서버는 윈도우 XP를 사용하였다. 또한 비주얼 C++ 6.0 개발 환경에서 프로토타입을 구현하고 다음과 같은 사례를 설정하고 실험을 수행하였다.

- **사례**

MIS(Management Information Systems) 전공 대학원생인 길동은 대학 교육에 대한 연구를 하고 있다. 그는 Person과 Degree의 관계에 관심이 많았고 관련 정보를 체계적으로 정의하고 활용하기 위해 온톨로지를 사용하려고 한다. 그의 최종적인 목적은 연구에 사용할 가장 적합한 온톨로지를 찾아 내어 적용하는 것이다. 이를 위해 University와 Student란 개념을 기본적으로 가지는 온톨로지들을 실험 대상으로 수집하고 이들 온톨로지 그룹을 대상으로 그의 연구에 필요한

개념과 관계들을 검색하려고 한다. 그는 수집한 온톨로지들을 가지고 다음과 같은 작업을 하였다.

(1) Person과 Degree 개념이 정의가 잘 되어 있고 그들 사이에 관계가 정의되어 있는지 찾아본다.
(2) Person과 Student개념이 잘 정의되어 있고 그들 사이가 종적관계로 정의되어 있는지 찾아본다.

위의 사례에서 (1)과 (2)의 두 가지 실험을 하였다. 각각의 실험에서 각 온톨로지별로 측정 방식을 적용하여 계산하고 가중치를 두고 합산하였다. 최종적으로 합산된 값을 기준으로 온톨로지들의 순위를 매겼다. 그런데 (1)과 (2) 각각의 실험에서 가중치를 변화시키며 각기 세 번의 실험을 하였다. 즉, 모든 측정 방식에 같은 비중의 가중치를 두는 방법, Alani et al.[1]의 연구인 AKTiveRank에서 좋은 결과를 내었던 가중치 조합을 사용하는 방법, 그리고 본 연구에서 새로이 설계한 관계의 의미적 유사성을 강조한 가중치 조합을 사용하는 방법을 써서 실험을 반복하였다. 본 실험은 정확한 가중치 조합을 구하기 보다는 의미적 유사성 기준을 추가한 본 연구 모델의 타당성을 검증하는 것이 목표이므로 타당성이 검증될 때까지 의미적 유사성과 관련된 RMM, Taxo에 두 배, 세 배, 혹은 여러 배수를 곱하는 방식을 택하였다. 한편, 도메인에 적합한 가중치 조합의 설정으로 유사성 측정에 보다 정확한 결과를 얻을 수 있는데, 이를 위해 회귀 모델의 적합화(model fitting) 기법을 사용하여 적합한 가중치를 구할 수 있다.

- **실험 1**

Person과 Degree 개념과 두 개념 사이의 관계를 가지는 온톨로지를 찾기 위해 각 온톨로지에 대해 다음과 같이 CMM, RMM, SSM, DEM등의 측정 방식을 적용한 값을 합산하여 전체 값을 계산한다.

$\mathrm{Total}[o] = \alpha\ \mathrm{CMM}[o, T] + \beta\ \mathrm{RMM}[o, t_1, t_2, r] + \gamma\ \mathrm{SSM}[o, T] + \delta\ \mathrm{DEM}[o, T]$.

이때 $\alpha + \beta + \gamma + \delta = 1$ 이며 각 측정 방식에 대하여 가중치를 변화시키며 다음과 같이 세 번의 실험을 실행하였다.

실험 1a(동일한 가중치를 적용한 경우). 즉, $\alpha = \beta = \gamma = \delta = 0.25$.

실험 1b(선행 연구에서의 가중치 조합을 적용한 경우). 즉, $\alpha = 0.1$, $\beta = 0.3$, $\gamma = 0.2$, $\delta = 0.4$.

실험 1c(관계의 의미적 유사성을 강조한 경우). 관계의 의미적 유사성을 강조하기 위해 의미적 유사성과 관련된 RMM에 두 배의 가중치를 부여하였다. 즉, $\alpha = 0.2$, $\beta = 0.4$, $\gamma = 0.2$, $\delta = 0.2$.

- **실험 2**

Person과 Student로 구성된 종적관계를 가지는 온톨로지를 찾기 위해 각 온톨로지에 대해 다음과 같이 측정치를 계산하고 합산하여 전체 값을 계산한다.

$\mathrm{Total}[o] = \alpha\ \mathrm{CMM}[o,T] + \beta\ \mathrm{Taxo}[o,$

T] + ɣ SSM[o,T] + δ DEM[o, T]

이때 α + β + ɣ + δ = 1이며 가중치를 변화시키면서 다음과 같이 세 번의 실험을 실행하였다.

실험 2a(동일한 가중치를 적용한 경우). 즉 α = β = ɣ = δ = 0.25.

실험 2b(선행 연구에서의 가중치 조합을 적용한 경우). 즉, α = 0.1, β = 0.3, ɣ = 0.2, δ = 0.4.

실험 2c(계층관계의 의미적 유사성을 강조한 경우). 계층관계의 의미적 유사성을 강조하기 위해 의미적 유사성과 관련된 Taxo에 두 배의 가중치를 부여하였다. 즉, α = 0.2, β = 0.4, ɣ = 0.2, δ = 0.2.

4.2 실험 결과

4.2.1 실험 1의 결과

<표 4>의 계산 값을 분석해 보면 온톨로지 B와 온톨로지 F만이 RMM값이 0이 아닌 것을 알 수 있다. 이는 온톨로지 B와 온톨로지 F만이 사용자가 찾고자 하는 Person과 Degree의 관계를 포함하고 있다는 것을 나타내 준다. CMM값은 모든 온톨로지들에 대해 비교적 일관된 값을 가지는 것을 볼 수 있다. 이는 대부분의 온톨로지들이 Person 혹은 Degree 검색어를 포함하고 있다는 것을 의미한다. 특히 두 검색어 모두를 포함하고 있는 온톨로지 B, 온톨로지 F, 온톨로지 J 등은 다른 온톨로지에 비해 상대적으로 높은 CMM값을 가지는 것을 확인할 수 있었다. DEM은 주어진 검색 개념들이 주변의 다른 개념들과 관계를 형성하는 정도를 나타내는 값으로 온톨로지 E와 I가 상대적으로 높은 DEM값을 가진다. 이는 다른 온톨로지에 비해 온톨로지 E와 I에서 검색 개념이 다른 개념들과 많은 연관관계를 가지며 보다 자세하게 표현되었다고 할 수 있다. SSM은 모듈화와 관련되는 값으로 온톨로지 계층구조내 검색 개념이 인접하게 위치하여 일부를 손쉽게 추출하여 다른 온톨로지 생성에 사용할 수 있음을 나타내주는 값이다. 온톨로지 B와 C가 높은 SSM값을 보여주고 있다. 이것은 이들 온톨로지에서 Person과 Degree가 인접하게 위치한다는 것을 의미한다.

〈표 4〉 실험 1의 측정치

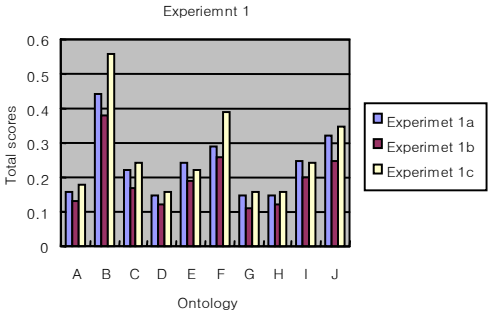
온톨로지	CMM	Taxo	RMM	DEM	SSM	CEM
A	0.36	0.0	0	0.13	0.06	0.64
B	1.0	0.0	0.3	0.31	1.0	0.50
C	0.55	0.0	0	0.22	0.93	0.71
D	0.18	0.0	0	0.27	0.0	1.0
E	0.18	0.0	0	0.6	0.0	1.0
F	0.45	0.0	0.3	0.17	0.0	0.75
G	0.18	0.0	0	0.20	0.0	0.60
H	0.18	0.0	0	0.27	0.0	0.60
I	0.18	0.0	0	0.67	0.0	0.60
J	0.91	0.0	0	0.29	0.33	0.27

서로 다른 가중치를 각각의 측정 방식에 부여하여 세 번의 실험을 하였을 때 RMM에 높은 가중치를 부여한 실험 1c가 다른 실험에 비하여 비교적 온톨로지들 간에 큰 차이를 보여주었다(<표 5> 참조). 모두 동일한

가중치를 각각의 측정 방식에 부여한 실험1a에서 세 번의 실험 중 평균적인 분포를 보여주었다. 그러나 <그림 4>에 나타난 바와 같이 대략 세 번의 실험 모두에서 유사한 형태의 분포를 보여주고 있다.

〈표 5〉 실험 1 결과

온톨로지	실험 1a	실험 1b	실험 1c
A	0.16	0.13	0.18
B	0.44	0.38	0.56
C	0.22	0.17	0.24
D	0.15	0.12	0.16
E	0.24	0.19	0.22
F	0.29	0.26	0.39
G	0.15	0.11	0.16
H	0.15	0.12	0.16
I	0.25	0.20	0.24
J	0.32	0.25	0.35



〈그림 4〉 실험 1 결과

4.2.2 실험 2의 결과

<표 6>의 계산 값을 분석해 보면 대부분의 온톨로지들이 높은 Taxo값을 가지는 것을 알 수 있다. 이는 온톨로지 A와 온톨로지 C만을 제외하고 모든 온톨로지들이 사용자가 찾고자 하는 Person과 Student를 가지며 두 개념 사이에 직접적인 종적관계를 포함하고 있다는 것을 나타내 준다. 온톨로지 C는 이행적 종적관계를 가지므로 상대적으로 낮은 Taxo값을 가진다. 또한 RMM값이 모두 0으로 나타난 것은 모든 온톨로지에서 Person과 Student간에 종속관계가 아닌 다른 관계를 가지지 않고 있다는 것을 의미한다. CMM값은 모든 온톨로지들에 대해 비교적 일관된 값을 가지는 것을 볼 수 있다. 이는 대부분의 온톨로지들이 Person 혹은 Student 개념을 포함하고 있다는 것을 의미한다.

〈표 6〉 실험 2의 측정치

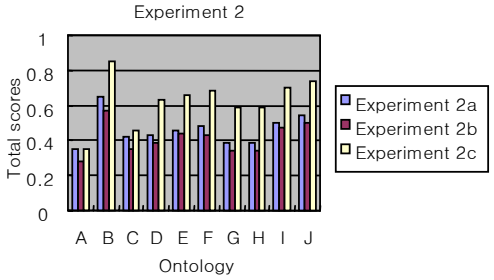
온톨로지	CMM	Taxo	RMM	DEM	SSM	CEM
A	0.87	0	0	0.10	0.79	0.43
B	1.0	1.0	0	0.24	1.0	0.50
C	0.87	0.20	0	0.15	0.87	0.43
D	0.50	1.0	0	0.13	0.49	0.67
E	0.50	1.0	0	0.30	0.49	0.58
F	0.63	1.0	0	0.10	0.67	0.56
G	0.50	1.0	0	0.07	0.36	0.70
H	0.50	1.0	0	0.07	0.36	0.70
I	0.63	1.0	0	0.25	0.64	0.40
J	0.75	1.0	0	0.40	0.53	0.64

실험 2의 결과가 <표 7>과 <그림 5>에 나타나 있다. Taxo에 높은 가중치를 부여한 실험 2c와 동일 가중치를 부여한 실험 2a에서 검색하는 계층관계가 없거나 직접적이지 않은 온톨로지 A, C를 제외하고 대부분의 온톨로지들이 실험 2c에서 높은 결과 값을 나타내었다. 일반적으로 세 번의 실험 모두에서 유사한 형태의 분포를 보여주고 있다.

즉, 가중치 차이에 관계없이 거의 비슷한 형태의 랭킹을 보여주는 안정적인 분포를 나타내고 있다.

〈표 7〉 실험 2 결과

온톨로지	실험 2a	실험 2b	실험 2c
A	0.35	0.28	0.35
B	0.65	0.57	0.85
C	0.42	0.35	0.46
D	0.43	0.39	0.63
E	0.46	0.44	0.66
F	0.48	0.43	0.68
G	0.39	0.34	0.59
H	0.39	0.34	0.59
I	0.50	0.47	0.70
J	0.54	0.50	0.74



〈그림 5〉 실험 2 결과

연관관계를 구성하는 Person과 Degree로 검색하였던 실험 1의 결과와 비교하여 계층관계를 구성하는 Person 그리고 Student로 검색하는 실험 2에서는 모든 온톨로지들에 대해 계산 결과 값이 상대적으로 높은 합계 값을 가지는 것을 볼 수 있는데(<표 5>, <표 7>참조) 이는 대부분의 온톨로지들이 Person 그리고 Student와 두 개념 간의 종적 관계를 가졌다는 것을 의미한다.

5. 평 가

본 장에서는 연구모델이 선행 연구보다 의미적 일치성 측면에서 우수하다는 것을 보이고자 한다. 그래서 연구모델의 실행 결과를 온톨로지 모델링 전문가가 온톨로지 순위를 매긴 결과, 그리고 선행 연구인 AKTive-Rank의 랭킹 결과와 상관 분석을 행하였다.

5.1 온톨로지 전문가 평가

연구모델에 의해 계산된 랭킹 결과와 비교하기 위해 온톨로지 모델링 전문가들을 대상으로 질문지를 통한 온톨로지 랭킹 실험을 실시하였다. 실험에는 여섯 명의 온톨로지 전문가가 선정되어 참여하였다. 실험에 참여한 전문가들은 MIS 전공분야의 온톨로지를 전문으로 연구하는 박사과정 연구원들로 정하였다.

우선 전문가들에게 실험 과정에 따라야 할 지침과 질문지, 대상 온톨로지의 내부 구조를 알 수 있는 그래픽 형태의 자료와 온톨로지 원문 OWL 파일, 그리고 각각의 온톨로지가 포함하는 개념간 관계 리스트를 제공하였다. 이렇게 제공된 질문지와 자료들은 전문가들이 정확한 분석을 할 수 있도록 충분한 시간을 준 뒤 회수 하였다. 질문지는 가능한 단순하고 명료하게 그리고 실험에서 온톨로지를 선택하는 기준 점수로 사용할 수 있도록 작성되었다(부록 : 질문지 참조). 각 질문에 대해 전문가들이 각각의 온톨로지에 점수를 매기게 하였다. 또한 전문가들은 질문지 및 제공 데이터에 대해 의견을 제시할 수 있도록 하여 질문지에 반영하였다. 회수된 질

문지로부터 점수를 합산하여 온톨로지의 순위를 매겼다.

전문가들의 랭킹 결과는 각각의 온톨로지에 대해 각기 다른 순위를 나타냈으며 여러 전문가들의 랭킹 결과의 평균을 구하여 상관분석에 사용하였다. 전문가들의 랭킹 결과의 표준편차는 1.2536으로 나타났다.

5.2 선행 연구의 랭킹 결과

본 연구에서 참조한 선행 연구 AKTive-Rank는 온톨로지 측정 방식으로 CMM, CEM, DEM 그리고 SSM을 사용하였다. 이들 측정 방식을 본 연구의 실험 환경에 동일하게 적용하여 CMM, CEM, DEM, SSM등 네 개의 측정치를 다시 계산하였다. 그리고 선행 연구에서 사용한 최선의 가중치 조합(CEM 0.3, CMM 0.1, DEM 0.4, SSM 02)을 그대로 적용하였다.

5.3 비교 평가 분석

실험 1과 실험 2의 결과들을 전문가들의 온톨로지 랭킹 결과와 비교하기 위해 스피어만 랭크 상관 분석을 사용하였다. 스피어만 랭크 상관 분석은 서열 관계인 랭킹 데이터 간 상관 관계를 분석하는데 사용된다[17]. 평가 변수끼리 완전한 연관관계에 있으면 +1 값을 가진다.

실험 1과 실험 2의 결과와 전문가 평가, AKTiveRank의 실행 결과가 <표 8>에 정리되어 있다.

<표 8>의 실험 1과 실험 2의 결과와 전문가의 평가 결과를 상관 분석하여 <표 9>과

〈표 8〉 온톨로지의 랭킹 결과

온톨로지	실험 1a	실험 1b	실험 1c	실험 2a	실험 2b	실험 2c	전문가 평가	AKTive Rank
A	9	7	7	10	10	10	9	7
B	1	1	1	1	1	1	1	1
C	6	6	4	7	7	9	2	5
D	10	8	8	6	6	6	10	8
E	5	5	6	5	4	5	7	4
F	3	2	2	4	5	4	3	6
G	7	10	8	8	8	7	5	10
H	7	8	8	8	8	7	6	8
I	4	4	4	3	3	3	8	3
J	2	3	3	2	2	2	4	2

같은 스피어만 상관계수를 구하였다. 또한 비교를 위하여 AKTiveRank의 결과와 전문가 평가 결과를 상관 분석하여 <표 9>에 정리하였다. 비교 결과 본 연구 모델인 실험 1과 실험 2를 전문가 평가 결과와 상관 분석한 결과는 유의한 상관 관계를 가진 반면 선행 연구인 AKTiveRank의 상관계수는 유의

〈표 9〉 전문가 평가 결과와의 스피어만 상관분석

실 험	스피어만 상관계수
실험 1a	.640*
실험 1b	.470
실험 1c	.665*
실험 2a	.653*
실험 2b	.622*
실험 2c	.586*
AKTiveRank	.404

주) *유의수준 0.05에서 유의.

하지 않은 것으로 판명되었다.

또한, 실험 1이 실험 2보다 상대적으로 높은 상관 관계를 가지며 RMM에 높은 가중치를 적용한 실험 1c가 가장 높은 상관 관계를 나타내었다. 이로써 본 실험에서 사용된 RMM은 사용자의 평가와 유사한 랭킹 결과를 나타내도록 강한 영향을 주는 것을 알 수 있었다.

6. 연구의 시사점

본 연구는 관계의 의미적 유사성에 근거하여 온톨로지를 검색하는 효과적인 방법을 제안하였다. 현재까지의 연구에서 온톨로지 선택 기준으로 토픽 커버리지, 인지도, 표현밀도, 모듈성, 성능 등이 고려되어 왔다[16]. 각각의 선택 기준이 중요하나, 근본적으로 검색 대상 개념이 온톨로지의 주요 개념이면서 의미상 정확하게 표현되었는지 구별할 수 있어야 한다. 토픽 커버리지는 가장 널리 사용되어 왔지만[16] 의미적 매칭의 정확성이 뒷받침되지 못하면 결과적으로 의미상 차이가 있는 개념을 포함한 다수의 온톨로지들이 검색 결과로 얻어진다. 따라서 검색어에 대한 매칭의 의미상 정확도가 반드시 고려되어야 한다. 특히, 관계의 의미적 유사성을 기준으로 온톨로지들을 검색하면 상대적으로 관련성이 적은 관계들을 여과하여 검색 결과를 크게 줄일 수 있다. 또한 본 연구의 실험에서 관계의 의미적 유사성에 높은 가중치를 주었을 때 온톨로지 랭킹 결과가 온톨로지 전문가들의 랭킹 결과와 의미 있는 상관관계를 가짐을 실험을 통하여 검증하였다.

그러나 당장 실무에 관계의 의미적 유사성 기준을 적용하여 온톨로지를 선택하는 것은 한계가 예상된다. 왜냐하면 현재 개발된 많은 온톨로지들은 개념 간의 관계가 명확하게 정의되지 못한 경향이 있기 때문이다. 예를 들면 “개념 A와 개념 B가 C의 관계가 있다”라고 명확히 정의하지 않고 단지 “개념 A는 개념 B와 관련되어 있다”라고 정의하거나 관계를 정의하는데 적극적이지 않았던 측면이 있다는 점이다. 즉, 이것은 현재까지 온톨로지 구축이 개념의 정의와 계층구조 정의를 중심으로 진행되어 왔다는 것을 의미한다. 또한 관계를 형식화하는 것이 개념을 형식화하는 것보다 더욱 어렵기 때문인 것으로도 추정된다[18]. 향후 개념 간의 관계를 명확히 정의하기 위한 이론적 발전과 실무적 노력이 이루어지면 본 연구 모델이 보다 활발하게 적용될 수 있으리라 예측된다. 특히 현재 온톨로지 검색 기능이 미약한 온톨로지 라이브러리나 온톨로지 검색 엔진에 본 연구 모델이 적용된다면 개념간 관계를 이용한 의미적 유사성이 높은 온톨로지를 효과적으로 검색할 수 있을 것이다

7. 결론 및 향후 과제

본 연구에서는 온톨로지내 개념간 관계의 의미적 유사성을 측정하기 위해 제안한 새로운 온톨로지 측정 방식인 RMM과 Taxo을 적용한 선택 및 랭킹 모델을 설계하였다. 본 연구 모델은 온톨로지가 가지는 개념의 일치성뿐만 아니라 표현의 풍부성, 개념간 관계와 같이 구조적 요인을 고려한 선택 모델이다. 다시 말해 개념의 문자 수준의 일치성뿐

만 아니라 더 나아가 의미적 측면의 일치성을 고려하기 위해 개념이 다른 개념과 맺고 있는 관계를 식별함으로써 의미적 일치성을 향상시킨 모델이다. 또한 모델을 실증적으로 검증하기 위해 연구모델의 프로토타입을 개발하고 전문가들의 랭킹 결과와 비교함으로써 우수성을 증명하였다.

본 연구는 학문적 측면에서 첫째, 온톨로지 선택과 랭킹을 위해 관계의 의미적 유사성에 근거한 새로운 방법을 시도하였고, 둘째, 제안한 모델의 타당성을 실증적으로 증명함으로써 온톨로지 선택을 위한 이론을 발전시켰다. 또한 실무적 측면에서 온톨로지에 대한 전문 지식이 없는 실무자들도 주요 개념 또는 개념간 관계를 이용하여 필요한 온톨로지를 용이하게 찾을 수 있는 실용적 방법을 제안하였다.

향후의 연구 과제는 다음의 두 가지로 요약할 수 있다. 첫째, 질의 검색어의 동의어와 상의어로 검색어를 확장시키는 질의확장(query expansion) 기능을 이용하여 온톨로지를 검색함으로써 보다 정확성을 높이는 것이다. 둘째, 사용자의 요구사항과 특성에 맞게 온톨로지를 선택할 수 있는 컨텍스트 기반 온톨로지 선택 모델을 개발하는 것이다.

참 고 문 헌

[1] Alani, H., and Brewster, C., Ontology Ranking based on the Analysis of Concept Structures, In Proceedings of the Third International Conference on Knowledge Capture(K-CAP 05), Banff, Canada, 2005.

[2] Buitelaar, P., Eigner, T., and Declerck, T., OntoSelect : A Dynamic Ontology Library with Support for Ontology Selection, In Proceedings of the Demo Session at the International Semantic Web Conference, Hiroshima, Japan, 2004.

[3] Cook, T. D., and Campbell, D. T., Experimental and quasi-experimental designs for research, Chicago : Rand McNally, 1979.

[4] Ding, L., Finin, T., and Joshi, A., Swoogle : A Search and Metadata Engine for Semantic Web, In Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management, 2004.

[5] Gangemi, A., Catenacci, C., Ciaramita, M., and Lehmann, J., Modeling Ontology Evaluation and Validation, ESWC, LNCS 4011, 2006, pp. 140-154.

[6] Green, R., Beans, C., and Myaeng, S., Semantics of relationships : An Interdisciplinary perspective, Information Science and Knowledge Management, 2002, pp. 91-110, Kluwer.

[7] Jang, J. J., and Conrath, D. W., Semantic similarity based on corpus statistics and lexical taxonomy, In Proceedings of International Conference on research in Computational Linguistics,

Taiwan, 1998.

[8] Park, J., and Ram, S., Information Systems Interoperability : What Lies Beneath?, ACM Transactions on Information Systems, Vol. 22, No. 4, October 2004, pp. 595-632.

[9] Lin, D., An Information-Theoretic Definition of Similarity, In Proceedings of the International Conference on Machine Learning, Morgan Kaufmann, Madison, Wisconsin, USA, 1998.

[10] Pan, Jeff Z., Tomas, E., and Sleeman, D., Ontosearch2 : Searching and Querying Web Ontologies, In Proceedings of the IADIS International Conference WWW/Internet 2006.

[11] Patel, C., Supekar, K., Lee, Y., and Park, E. K., OntoKhoj : A Semantic Web Portal for Ontology Searching, Ranking and Classification, In Proceeding of the Workshop on Web Information and Data Management, ACM, 2003.

[12] Pennacchiotti, M., and Pantel, P., Ontologizing Semantic Relations, In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL, July 17-18, 2006, pp. 793-800.

[13] Pennacchiotti, M., and Pantel, P., A Bootstrapping Algorithm for Automatically Harvesting Se mantic Relations, In Proceedings of Inference in Computational Semantics(ICoS-2006), Buxton, England, 2006.

[14] Rada, R., Mili, H., Bicknell, E. and Blettner, M., Development and application of a metric on semantic nets, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 19, 1989, pp. 17-30.

[15] Resnik P., Semantic Similarity in a Taxonomy : An Information-Based Measures and its Application to Problems of Ambiguity in Natural Language, Journal of Artificial Intelligence research, Vol. 11, 1999, pp. 95-130.

[16] Sabou, P., Lopez, V., Motta, E., and Uren, V., Ontology Selection : Ontology Evaluation on the Real Semantic Web, In Proceedings of the Evaluation of Ontologies on the Semantic Web Workshop, held in conjunction with WWW 2006.

[17] Spearman, C., The Proof and Measurement of Association between Two Things, Amer. J. Psychol., Vol. 15, 1904.

[18] Turney, P. D., Measuring Semantic Similarity by Latent Relational Analysis, In Proceedings of the Nineteenth International Joint on Artificial Intelligence(IJCAI-05), 2005, pp. 1136-1141, Edinburgh, Scotland.

〈부 록 : 질문지〉

다음 사례는 온톨로지를 이용한 연구의 한 예입니다. 다음 사례를 읽고 주어진 질문에 답변해 주시기 바랍니다.

- **사례**

MIS전공 대학원생인 길동은 대학 교육에 대한 연구를 하고 있다. 그는 Person과 Degree의 관계에 관심이 많았고 관련 정보를 체계적으로 정의하고 활용하기 위해 온톨로지를 사용하려고 한다. 그의 최종적인 목적은 연구에 사용할 가장 적합할 온톨로지를 찾아 내어 적용하는 것이다. 이를 위해 University와 Student 개념을 기본적으로 가지는 온톨로지들을 수집하고 이들 온톨로지 그룹을 대상으로 그의 연구에 필요한 개념들을 검색하려고 한다. 그는 수집한 온톨로지를 가지고 다음과 같은 작업을 하려고 한다.

- Person과 Degree 개념이 정의가 잘 되어 있고 그들 사이에 관계가 정의되어 있는지 찾아 본다.
- Person과 Student 개념이 잘 정의되어 있고 그들 사이가 종적(IS-A)관계로 정의되어 있는 지 찾아본다.

이 때, 찾고자 하는 개념의 동의어는 고려하지 않으며 검색어를 포함하는 개념은 고려하기로 합니다. 예를 들면 Student 검색 시 Graduate Student를 고려하기로 합니다. 또한 이들 개념이 잘 정의되어 있는지를 판단하는 기준으로 3가지 기준을 사용하려고 하며 기준은 다음과 같습니다.

(1) Topic Coverage : 찾고자 하는 개념이 모두 정확하게 정의되어 있는가?
(2) Semantic Similarity in Relation : 개념이 의미적으로 잘 정의되어 있는지 보기 위해 다른 개념과의 관계가 잘 정의되어 있는가?
(3) Richness : 찾고자 하는 개념이 다른 개념과의 연관관계(부모, 자식, 조상, 형제 관계)를 많이 가지는가?

- **질문**

1. 각각의 온톨로지는 사례와 관련된 개념들을 모두 포함합니까?
2. 각각의 온톨로지는 사례와 관련하여 개념간 관계를 모두 포함합니까? 만약 관련 개념이 포함되어 있지 않아서 관계를 포함하지 않아도 동일한 기준을 적용하여 판단합니다. 즉, 이러한 경우, 개념과 관계 모두 포함하지 않으므로 더 낮은 점수를 받게 됩니다(2번~6번까지 동일하게 적용됨).

3. 각각의 온톨로지는 사례와 관련하여 개념간 종적(IS-A)관계를 모두 포함합니까?
4. 각각의 온톨로지는 사례에 필요한 조상, 자식, 형제 개념을 모두 포함합니까?
5. 사례에서 필요로 하는 모든 개념들이 온톨로지 트리 상의 중간에 위치합니까?
6. 사례에서 필요로 하는 모든 개념들이 온톨로지 트리 상에서 인접하게 위치합니까?
7. 사례의 작업을 하기 위해 가장 적합한 온톨로지는 어느 것입니까?

저 자 소 개

오선주 (E-mail : ohsunju7@snu.ac.kr)
1986년 서울대학교 계산통계학과 (학사)
1993년 서울대학교 계산통계학과 전산과학전공 (이학석사)
2004년~현재 서울대학교 경영대학 MIS전공 (박사과정)
관심분야 온톨로지, e-비즈니스, Information Technology Architecture 등

안중호 (E-mail : jahn@snu.ac.kr)
1975년 서울대학교 문리과대학 외교학과 (정치학사)
1980년 서울대학교 행정대학원 (행정학석사)
1987년 New York University (Stern School, 경영학 석·박사)
1987년~1988년 미국 Fordham 대학, Baltimore 대학, 동국대학교 조교수
1994년 서울대 연구부처장
1999년 한국경영정보학 회장
2000년 한국퍼실리티메니지먼트학 회장
1989년~현재 서울대학교 경영대학 및 경영전문대학원 교수
관심 분야 IT 거버넌스, BPM, e-비즈니스 정보기술전략, PR, ERP 등

박진수 (E-mail : jinsoo@snu.ac.kr)
1999년 The University of Arizona 경영학 (박사)
1999년~2002년 University of Minnesota 조교수
2002년~2005년 고려대학교 경영대학 조교수
2005년~현재 서울대학교 경영대학 및 경영전문대학원 부교수
관심분야 온톨로지, 시맨틱 웹, 정보시스템 통합, 지식공유