의미거리측정방법을 활용한 분산 온톨로지 간 자동 정렬 방법 연구*

A Study on an Automatic Alignment Method of Distributed Ontology by Using Semantic Distance Measure Method

황상규(Sang-Kyu Hwang)**
변영태(Young-Tae Byun)***

초 록

시멘틱 웹은 현재의 월드와이드웹의 진화된 모습으로 컴퓨터와 인간이 서로 협업할 수 있도록 컴퓨터가 이해할 수 있는 지식데이터베이스인 온톨로지 기술을 활용한다. 그러나, 온톨로지를 활용하여 정보의 의미를 이해하고 처리 가능하도록 데이터의 표현형식이 표준화 되더라도, 각기 다른 개발자가 서로 다른 개념하에 구축한 온톨로지를 기반으로 작성된 데이터는 상호 불일치 문제를 유발할 수 있다. 따라서, 서로 다른 개념 하에 구축된 온톨로지 간에는 상호 서로 다른 온톨로지 간 정렬작업이 필요하다. 서로 다른 온톨로지 개념노드 간 자동화 처리된 의미정렬 시 인간전문가가 참으로 판단한 사실을 거짓으로 잘못 판단하는 문제상황(false negative)에 의해 정렬오류문제가 발생하게 되는데, 본 연구에서는 서로 다른 온톨로지 개념노드 간 의미정렬과정에서 발생하는 false negative 오류를 최소화 할 수 있는 알고리즘을 새롭게 개발, 제시하였다.

ABSTRACT

Semantic web technology is the evolution of current World Wide Web including a machine-understandable knowledge database, ontology, it may be enable machine and people to work together. However, problems arise when we try to communicate with different data, which are annotated by different ontologies created by different people with different concepts. Thus, to communicate between ontologies, it needs to align between heterogeneous ontologies. When it is aligned between concept nodes of heterogeneous ontologies, one of main problems is a misalignment situation caused by false negative of automatic ontology mapping. So, in this paper, we present a new method to minimize the false negative error in the process of aligning concept nodes of different ontology.

키워드: 온톨로지 자동정렬, 온톨로지 정렬오류, 의미 유사성
automatic ontology alignment, ontology misalignment, semantic similarity

---

* 본 논문은 2007학년도 홍익대학교 학술연구진흥비에 의하여 지원되었음.
** 홍익대학교 컴퓨터공학과 인공지능연구실(kid4@naver.com) (제1저자)
*** 홍익대학교 컴퓨터공학과 인공지능연구실(ytbyun@hongik.ac.kr) (공동저자)
■ 논문접수일자 : 2009년 11월 18일 ■ 최초심사일자 : 2009년 11월 23일 ■ 게재확정일자 : 2009년 12월 3일
■ 정보관리학회지, 26(4): 319-335, 2009. [DOI:10.3743/KOSIM.2009.26.4.319]

1. 개 요

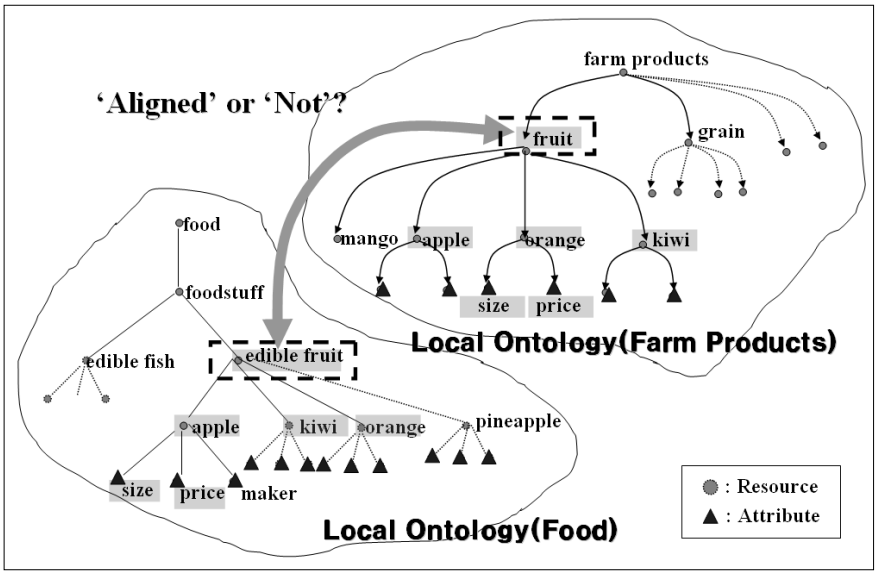
최근 특정 주제영역 도메인에서 사용되는 어휘 개념에 대한 정확한 명시적 정의, 어휘 간 계층적 관계 등을 정형화한 지식 데이터베이스인 온톨로지가 여러 다양한 분야의 정보검색시스템 설계 시 효율적 지식기반 정보처리기법으로 도입, 활용되고 있다. 그러나 대부분의 경우에 있어 온톨로지 구축 및 활용은 개발언어 및 데이터 표현양식의 차이 등의 이유로 제한된 도메인 영역에서만 이루어진다. 온톨로지의 구축 및 활용이 제한되는 가장 큰 이유는 개발자들은 RDF, DAML, OWL 등 다양한 온톨로지 개발언어를 이용하여 각기 서로 다른 개념하에 자신만의 온톨로지를 구축, 운영하고 있기 때문이다. Andrew(2008)는 이렇게 각각의 주제영역별로 개별적으로 구축 운영되는 온톨로지를 지역 온톨로지(local ontology)라고 정의하였으며, 서로 다른 주제영역의 지역 온톨로지 간 매핑, 정렬하기 위한 연계 도구로써 구축, 활용되는 온톨로지를 도메인 온톨로지(domain ontology)라고 정의하였다. 도메인 온톨로지는 상호관련이 있는 서로 다른 주제 영역의 개념노드 간 의미적 연관관계 정보를 제공함으로써, 서로 다른 주제영역의 지역 온톨로지 개념노드 간 상호연계를 가능하게 해준다. 도메인 온톨로지는 서로 다른 주제영역 어휘 간의 매핑, 정렬을 위한 도구로써 활용하기 위한 전제조건으로 설명하고자 하는 대상객체에 대한 명시적 정의(explicit definition)를 제공해야 한다. 예를 들어 '배'라는 어휘는 상황에 따라 '먹을 수 있는 배' 또는 '교통수단으로써의 배'라는 어휘 의미의 중의성 문제를 유발하기 때문에 도메인 온톨로지는 대상객체에 대해 여러 가지 방법으로 명시적 정의를 제공한다. 서로 다른 도메인 온톨로지를 대상으로 상호관련이 있는 수많은 개념노드 간 의미적 연관관계를 식별하기 위해서는 많은 비용과 시간이 요구되며, 연관관계 식별을 보다 효율적으로 수행하기 위한 방안으로 유사도 계산 알고리즘과 같은 자동화된 처리기술을 도입, 적용할 수 있다. 그러나, 자동화된 알고리즘이 실제 참인 정렬관계를 비정상으로 판정하는 false negative발생 시, 비정상으로 판정된 연산결과는 인간전문가의 검토 대상에서 제외됨에 따라 대규모의 온톨로지 개념노드 간 정렬작업 수행함에 있어 정렬작업의 정확도를 낮추는 주요요인으로 작용할 수 있다. 이에 본 연구에서는 각기 서로 다른 두 온톨로지 개념노드 간 자동화된 유사도 계산 시 발생하는 false negative문제 해결을 위하여 새로운 유사도 계산 방법을 개발하였다. 다음 2장에서는 자동화된 유사도 계산 시 발생하는 정렬오류(misalignment) 문제상황을 보다 명료하게 정의하였으며, 3장에서는 false negative 문제 해결을 위해 기존 유사도 계산 알고리즘의 개선 방안을 기술하였다. 4장에서는 현실세계와 보다 유사한 환경하에 실험을 통해 false negative문제의 발생 빈도를 살펴보고 false negative문제 해결에 있어 새롭게 개선된 알고리즘과 기존 방법과의 성능차이를 비교 분석하였으며, 마지막으로 5장에서는 연구수행결과를 종합, 정리하였다.

2. 관련연구

2.1 온톨로지 간 정렬문제

일반적으로 대부분의 온톨로지는 대상객체에 대한 명시적 정의를 위해 다양한 방안을 모색하게 되는데, 명시적 정의를 위한 대표적 방안으로 대상객체를 한 단어로 표현하기 보다는 유사어휘의 집합인 개념노드로 표현함으로써 어휘의미중의성 문제를 예방할 수 있다. 온톨로지의 일종으로 간주되는 어휘사전인 워드넷(WordNet)에서 개념노드는 문맥에서 단어들이 서로 대체되었을 때 의미적으로 큰 차이가 없는 단어들의 집합으로 표현된다. 예를 들어 동일한 개념을 설명하기 위한 유사어휘의 집합인 {부모, 부모님, 어버이}는 대상객체에 대한 명시적 정의로써 하나의 개념노드가 될 수 있다. 이렇게 대상 온톨로지의 수나 종류도 다양해지고 연관되어 처리해야 할 정보의 양도 급격히 늘어나면서, 의미 유사성(semantic similarity)을 활용한 효율적 온톨로지 관리 및 활용의 필요성이 부각되고 있다. 의미 유사성이란 대상객체의 어휘의미적 정보를 사용해 관련 객체와의 유사성을 측정하는 것이다. 서로 다른 온톨로지 개념노드 간 의미정렬도 이러한 의미 유사성에 근거해 정렬여부를 판단할 수 있다. 예를 들어 우리가 과일(fruit)이라는 단어를 들었을 때 먹을 수 있는 과일(edible fruit)을 연상하는 것이 의미 유사성이다.

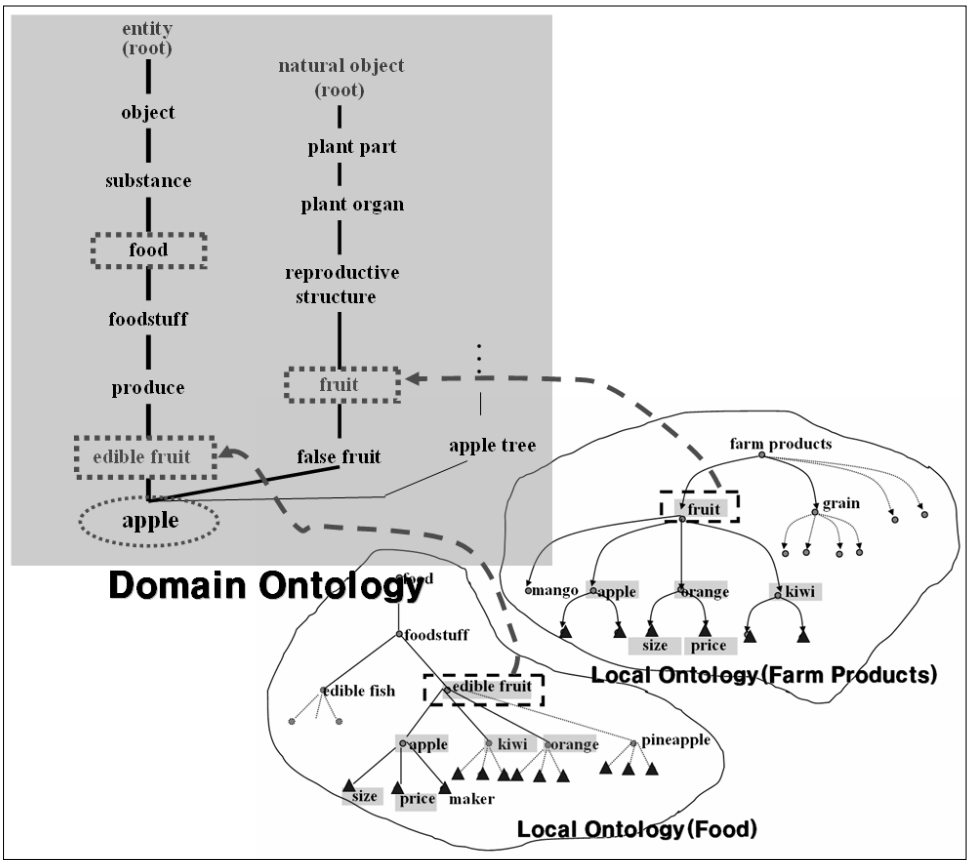
〈그림 1〉에서처럼 농업을 주제로 한 지역 온톨로지의 개념노드인 'fruit'과 음식물을 주제로 한 지역 온톨로지 개념노드인 'edible fruit'간에는 'apple', 'orange', 'kiwi'와 같은 공통의 자손노드가 여럿 존재하며, 이를 근거로 두 개념노드 간 의미정렬 여부를 판단할 수 있다. 이러한 온톨로지 개념노드 간 의미정렬작업은 많은



〈그림 1〉 온톨로지 개념노드 간 의미정렬

비용과 시간이 요구되며, 정렬작업을 보다 효율적으로 수행하기 위한 방안으로 유사도 계산 알고리즘과 같은 자동화된 처리기술을 도입, 적용할 수 있다. 근래는 여러 가지 유사도 계산 방법 가운데 벡터기반 유사도 계산 방법을 이용하여 온톨로지 유사도를 계산하는 연구가 많이 진행되고 있으며, 예를 들어 Ce Zhang(2008), Rajesh(2008)는 온톨로지 정렬과 온톨로지 정합을 위해 벡터기반 유사도 계산알고리즘을 적용하였다. 벡터기반 유사도 계산방법이란 간단히 말해 두 개념노드 간 공통어휘 빈도수를 통해 유사도를 계산하는 방법이다. 벡터기반 유사도 계산방법에서는 〈그림 2〉와 같이 서로 다른 지역 온톨로지의 두 개념노드를 도메인 온톨로지를 통해 상호 매핑 시 'fruit'과 'edible fruit'간에는 'apple'이라는 공통의 자손노드가 1개 존재하며, 이를 근거로 두 개념노드 간 의미정렬 여부를 판단할 수 있다. 공통어휘 빈도수 기반 유사도 계산방법의 가장 큰 단점은 개념노드와 자손노드 간 연관밀접도가 유사도 계산과정에 반영되지 못한다는 점이다. 예를 들어 〈그림 2〉에서 'edible fruit' 대신으로 'food'와 'fruit'간의 공통어휘 빈도수 기반 유사도를 계산해도 공통의 자손노드는 동일하게 'apple' 1개만 존재함으로써 유사도 계산결과값은 기존과 동일하다. 그러나, 'edible fruit'이 'food'보다는 'fruit'이라는 개념노드와 보다 밀접한 관계를 가짐에 따라 보다 높은 유사도 계산값을 가지는 것이 타당하다. 이러한 문제점을 해결하기 위해 상황에 따라 공통 자손노드에 각기 서로 다른 가중치값을 부여할 수도 있는데, 온톨로지를 대상으로 한 유사도 계산알고리즘에 있어 가중치값 부여를 위해 가장 직관적으로 많이 쓰이는 방법으로 의미거리기반 유사도 계산방법이 있다. 거리기반 유사도 계산방법은 보통 서로 다른 온톨로지의 두 개념노드를 대상으로 조상-자손 관계에 있는 공통된 자손(혹은 부모)노드와의 링크거리를 계산함으로써 정렬여부를 판단한다. 예를 들어 〈그림 2〉에서 공통 자손노드인 'apple'에서부터 'edible fruit'과 'fruit'까지 링크거리 합은 3이며, 'apple'에서부터 'food'와 'fruit'까지 링크거리 합은 6으로 상대적으로 낮은 가중치값을 가지도록 가중치를 부여할 수 있다.

서로 다른 온톨로지 개념노드 간 유사도의 자동화된 계산에 있어 유사도 계산 알고리즘 자체의 성능적 고려 외에도 참인 정렬관계를 비정상으로 판정하는 false negative문제에 대한 대응방안 마련이 요구된다. 그 이유로 자동화된 알고리즘이 잘못된 정렬관계를 정상으로 판정하는 false positive문제는 인간전문가가 알고리즘 수행결과인 적합데이터를 검토하는 과정에서 문제상황을 식별, 정정할 수 있으나, 참인 정렬관계를 비정상으로 판정하는 false negative인 경우에는 대규모의 정렬작업 수행 시 알고리즘 수행결과 부적합데이터는 인간전문가의 검토 대상에서 처음부터 제외시킴에 따라 문제상황인식 자체가 불가능하기 때문이다. 예를 들어 〈그림 2〉에서 자동화된 유사도 계산방법이 'fruit'과 'edible fruit'간에 공통의 자손노드가 아니라 공통의 조상노드 유무기준으로 의미정렬 여부를 판단할 경우, 'fruit'과 'edible fruit'간에는 공통의 조상노드가 없으므로 두 개념노드 간 의미연관성은 없다고 잘못 판단할 수 있다. 이 경우 인간전문가는 문제상황을 식별, 정정할 수 있으나, 자동화된 시스템은 참인 정렬



〈그림 2〉 도메인 온톨로지 기반 서로 다른 개념노드 간 의미정렬여부 판단

관계를 비정렬로 판단함으로써 false negative 문제가 발생하게 된다. 이에 본 연구에서는 현실세계와 보다 유사한 환경하에 실험을 통해 온톨로지 개념노드 간 자동화된 의미정렬 시 false negative문제가 정말 중요한 고려요소인지 여부를 비교실험분석을 통해 확인하였으며, false negative를 예방하기 위한 방안으로 기존 유사도 계산 방법의 개선 방안을 모색해보았다.

2.2 온톨로지 정렬 개념정의

서로 다른 온톨로지 개념노드 간 의미정렬 시 발생하는 정렬오류 문제상황을 보다 명확히 정의하기 위하여, 온톨로지 간 정렬, 온톨로지 개념노드 간 의미정렬 및 정렬오류를 다음과 같이 논리식 형태로 정의하였다.

- **Definition 1** (온톨로지 간 정렬) 두 개의 지역 온톨로지 O_1과 O_2가 존재하고 두 온톨로지 간 매핑을 위한 도메인 온톨로지 O'가 존재할 때 O_1에서 O'으로의 매핑 $MO_1 \rightarrow O'$와 O'에서 O_2로의 매핑 $MO' \rightarrow O_2$가 존재한다면, 두 지역 온톨로지 O_1과 O_2간 정렬관계 집합은 $RO_1 \rightarrow O_2$라고 정의한다.

- **Definition 2** (개념노드 간 정렬) 온톨로지 O_1에 속한 개념노드 t($t \in O_1$)와 온톨로지 O_2에 속한 개념노드 t'($t' \in O_2$)존재하고 두 개념노드 간에서 정렬관계 〈t, r, t'〉이 존재한다면, 개념노드 간 정렬관계 r은 두 지역 온톨로지 O_1과 O_2간 성립되는 정렬관계집합 중 하나의 엘리먼트($\exists r \in R_{O_1 \rightarrow O_2}$)이다.
- **Definition 3** (개념노드 간 의미정렬) 의미 유사성이란 대상 객체의 어휘의미적 정보를 사용해 관련 객체와의 유사성을 측정하는 것이다. '개념노드 간 의미정렬'이란 서로 다른 두 지역 온톨로지 개념노드 상호 간 의미 유사성에 근거해 정렬여부를 판단한다.

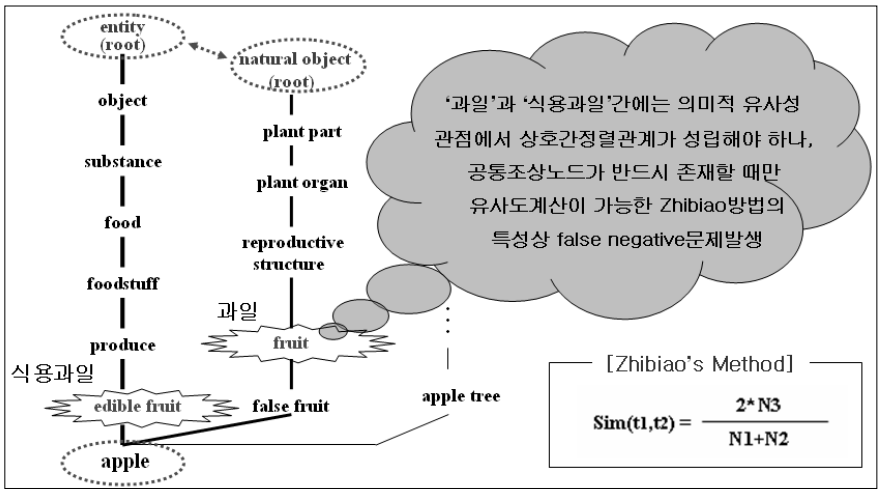
두 개의 지역 온톨로지 O_1과 O_2간 존재하는 개념노드 간 모든 정렬관계집합을 계산하기 위해서는 많은 연산처리가 수행되어야 한다. 이에 현실적으로 두 지역 온톨로지 개념노드 간에 존재하는 모든 연관 관계를 비교, 식별하는 대신 보통 가장 유사도가 높은 연관 관계를 기준으로 두 개념노드 간 의미정렬여부를 판단한다. 예를 들어 거리기반 유사도 계산방식에서는 두 개념노드를 대상으로 조상-자손 관계에 있는 공통된 자손노드를 모두 식별, 판단하기보다는 두 개념 노드에서 가장 가까운 자손노드까지의 링크거리 합을 기준으로 두 개념노드 간 의미정렬여부를 판단할 수 있다.

- **Definition 4** (개념노드 간 정렬오류) 온톨로지 개념노드 간 의미정렬여부를 위해 사용하는 자동화된 유사도 계산 알고리즘이 실제 참인 정렬관계를 비정상으로 판정하는 false negative상황 발생 시, 자동화된 유사도 계산 알고리즘에 의해 발생한 문제상황을 '개념노드 간 정렬오류'라고 정의한다.

2.3 온톨로지 개념노드 간 정렬오류 문제

도메인 온톨로지 내에서 각기 서로 다른 주제영역에 속한 두 개념노드 간 자동화된 의미정렬 판단방법으로 보통 어휘의미 간 유사도 계산 알고리즘을 사용한다. 본 논문에서는 온톨로지를 대상으로 한 유사도 계산방법에 있어 가장 직관적으로 많이 쓰이는 방법으로 거리기반 유사도 계산방식 중 하나인 Zhibiao(1994) 방법을 적용하였다. 거리기반 유사도 계산방식에 하나인 Zhibiao방법은 두 개념노드에서 가장 가까운 자손노드까지의 링크거리 합을 기준으로 두 개념노드 간 의미정렬여부를 판단한다. Zhibiao방법의 유사도 계산식은 다음 〈그림 3〉과 같다.

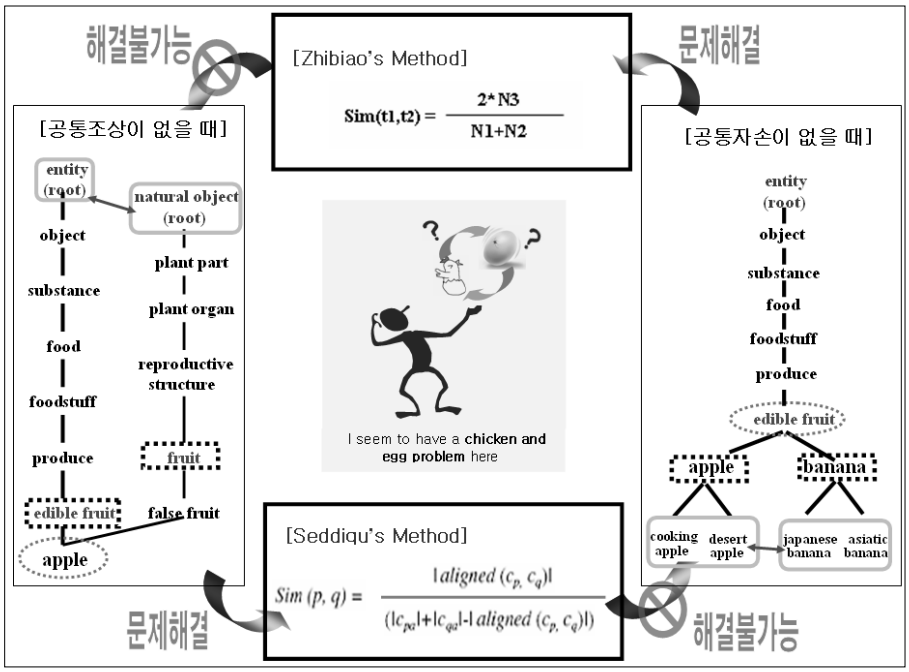
〈그림 3〉에서 N1, N2는 두 개념노드 t1, t2로부터 가장 최단거리에 있는 공통의 조상노드까지의 링크거리합(노드 간 연결 링크 수)이며, N3은 MRCA에서부터 최상위 노드(Root)까지의 최단링크 거리이다. 기존 Zhibiao방법은 개념노드 '과일(t1=fruit)'과 '식용과일(t2=edible fruit)'과 같이 공통조상노드가 존재하지 않는 경우에는 유사도 계산이 불가능하다. 이를 공통조상노드에 기반하여 유사도 계산을 수행하는 Zhibiao방법 만의 제약사항이라고 간주할 경우, 공통자손노드인 'apple'을 가지고 유사도



〈그림 3〉 공통조상노드 부재 시, 발생되는 false negative문제

계산을 수행하도록 Zhibiao방법을 변형할 수 있다. 그러나, 〈그림 3〉의 예제에서처럼 농업을 주제로 한 지역 온톨로지의 개념노드인 t1과 음식물을 주제로 한 지역 온톨로지 개념노드인 t2간 공통의 최상위 노드가 존재하지 않을 경우에는 여전히 N3의 값은 0이 되며, Sim(t1, t2) = 0이 됨에 따라 false negative문제가 계속 발생한다. 공통조상노드가 반드시 존재할 때만 유사도 계산이 가능한 Zhibiao방법의 false negative문제해결을 위하여 공통의 최상위 노드가 없어도 공통자손노드가 존재하면 유사도 계산이 가능한 방법론을 새롭게 조사하였으며, 그 과정에서 순수하게 공통자손노드 기반으로 유사도 계산이 가능한 Seddiqui(2006)방법을 확인하였다. Seddiqui방법에서는 "자식노드들 간의 높은 유사성이 존재할 때 부모노드 간에도 유사성이 높다."라는 가정 하에 〈그림 4〉와 같이 유사도 계산 알고리즘을 제시하고 있다.

〈그림 4〉의 Seddiqui방법에서 p, q는 유사도를 계산하는 대상인 두 개념노드이며, Cpa, Cqa는 두 개념노드 p, q각각에서 정렬관계가 성립하는 자손노드의 개수를 의미한다. aligned(Cp, Cq)는 두 개념노드 p,q간에 정렬관계가 성립하는 공통 자손노드 수이다. Seddiqui방법은 공통조상노드가 없어도 공통자손노드가 존재하면 유사도 계산이 가능한 방법으로 〈그림 4〉에서 'fruit'과 'edible fruit'간에 'apple'이라는 공통의 자손노드가 존재함에 따라 두 개념노드 간 의미적 연관성이 존재함을 확인할 수 있다. 그러나, Seddiqui방법은 공통자손노드가 반드시 존재할 때만 유사도 계산이 가능한 방법으로 〈그림 4〉의 우측 그림과 같이 두 개념노드 '사과(p=apple)'와 '바나나(q=banana)'간에 정렬관계가 성립하는 공통 자손노드가 존재하지 않을 경우에는 aligned(Cp, Cq) 값은 0이 된다. 결과적으로 Sim(p,q) = 0이 됨에 따라 공통자손노드가 존재하지 않는 경우 새로운 false negative문제가 발생함을 확인할 수 있다. 이에 본 연구에서는 공통조상 또는 공통자손노드가 존재하지 않는 경우 발생하는 false negative문제 해결을 위하



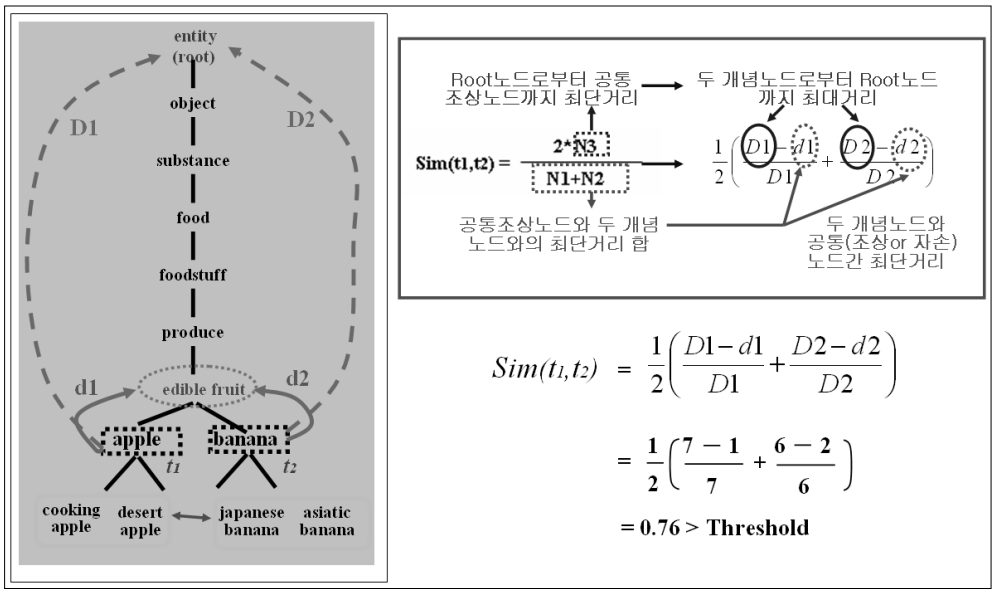
〈그림 4〉 개념노드 간 유사도 계산 시, 순환 반복되는 false negative문제

여, 기존 Zhibiao방법을 개선한 새로운 유사도 계산 방법을 개발하였다.

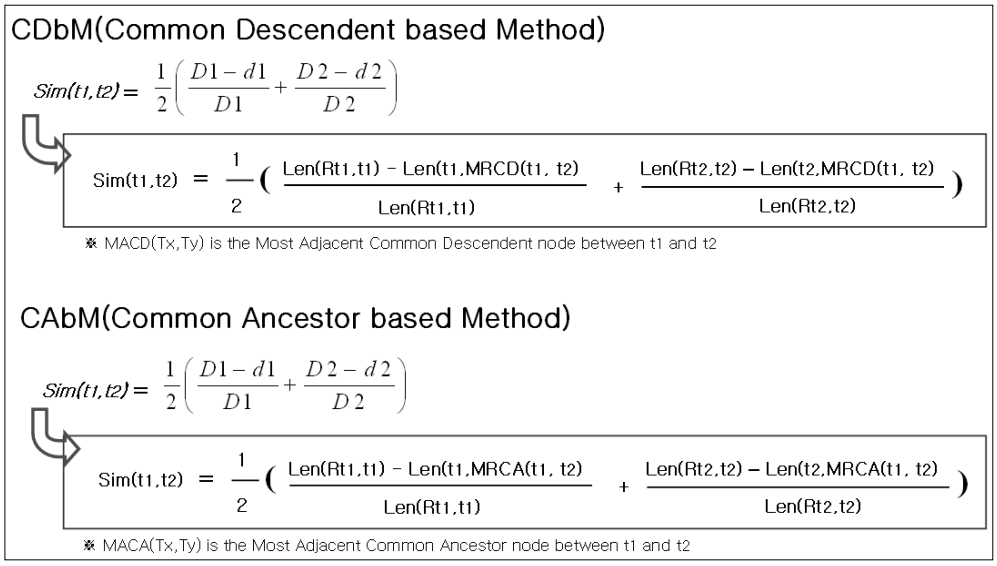
3. 의미거리측정 방법개선을 통한 정렬오류해결

단일 주제영역으로 구성되어 최상위 노드가 하나인 Tree형태(〈그림 4〉의 우측 그림 참조)가 될 수 있는 지역 온톨로지와는 달리 다양한 주제 영역의 집합인 도메인 온톨로지는 최상위 노드가 다수 존재하는 Forest구조(〈그림 4〉의 좌측 그림 참조)를 가지게 된다. Forest 구조를 가지는 도메인 온톨로지의 아키텍처 특성상 유사도 계산 시 공통의 최상위 노드가 존재하지 않음에 따라 결과값이 0이 되는 상황은 false negative문제가 발생하는 주요원인이 되며, false negative문제해결을 위해서는 기존 유사도 계산 알고리즘의 개선이 요구된다. 이에 공통의 자손노드가 존재하나 최상위 노드가 존재하지 않을 경우에도 유사도 계산이 가능하도록 〈그림 5〉와 같이 기존 Zhibiao방법을 개선하였다.

Forest 구조를 가지는 일반적인 도메인 온톨로지의 아키텍처 특성상 공통조상노드가 존재하지 않을 경우 공통자손노드 기반으로 유사도 계산이 가능토록 〈그림 5〉와 같이 기존 Zhibiao 알고리즘을 1차로 개선하였으며, 공통조상 또는 공통자손노드 둘 중 하나만 존재해도 유사도 계산이 가능토록 각각의 유사도 계산 수식을 표준화하기 위하여 다음 〈그림 6〉과 같이 개선된 Zhibiao알고리즘을 2차로 변환하였다.



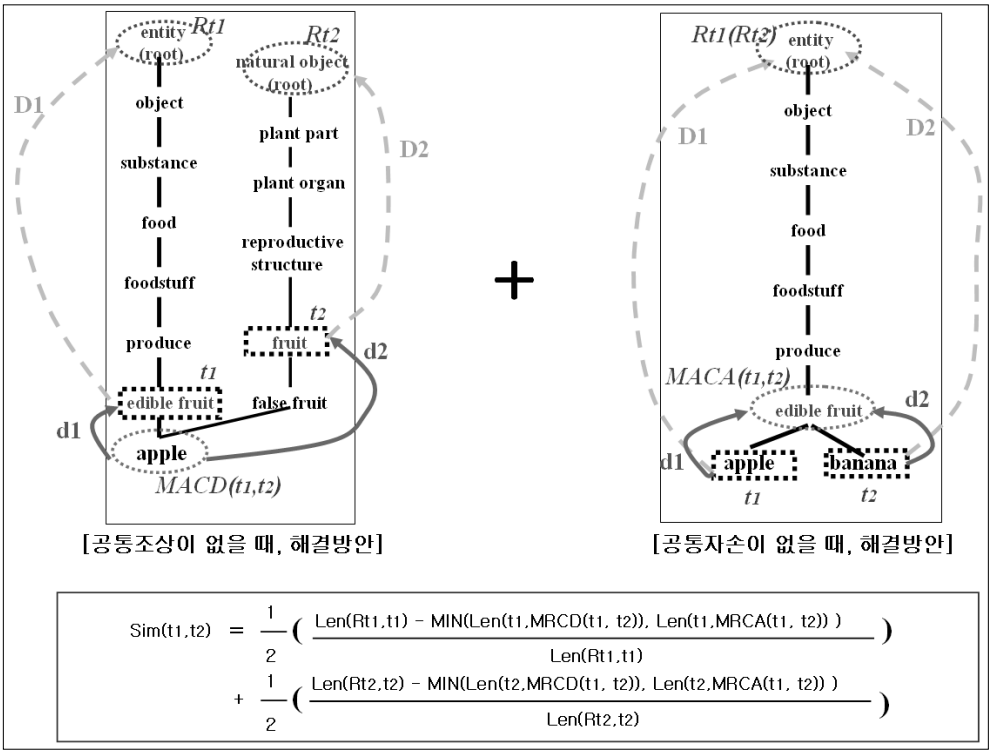
〈그림 5〉 Zhibiao알고리즘 개선 방안



〈그림 6〉 공통자손/공통부모 기반 유사도 계산 알고리즘 간 표준화

공통자손 기반 유사도 계산 알고리즘인 CDbM (Common Descendent based Method)에서 Rt1, Rt2는 두 개념노드 t1, t2 각각의 최상위 노드이며, Len(Rt1, t1)은 최상위 노드 Rt1으로부터 개념노드 t1까지의 최단링크거리이다. MRCD(t1, t2)는 두 개념노드 t1, t2에서 가장

가까운 거리에 있는 공통자손노드(the Most Adjacent Common Descendent node)이며, MRCA(t1, t2)는 두 개념노드 t1, t2에서 가장 가까운 거리에 있는 공통조상노드(the Most Adjacent Common Ancestor node)이다. Len(t1, MRCD(t1, t2))는 개념노드 t1로부터 공통자손노드 MRCD(t1, t2)까지 최단링크거리이다. 새로 개발한 CDbM은 공통자손노드가 존재하지 않을 경우 유사도 계산이 불가능하며, 이 경우 기존 공통조상노드 Zhibiao방법을 CDbM과 융합이 가능하도록 계산식을 변형한 알고리즘인 CAbM(Common Ancestor based Method)을 적용할 수 있다. 기본적으로 CAbM은 기존 Zhibiao알고리즘과 동일하게 공통조상노드 기반 유사도 계산방법이며, CAbM을 만든 목적은 공통조상 또는 공통자손노드 어느 둘 중 하나만 존재해도 유사도 계산이 가능토록 기존 Zhibiao알고리즘을 CDbM과 동일한 형태로 변환한 알고리즘이다. CAbM에서 MRCA (t1, t2)는 두 개념노드 t1, t2의 공통조상노드이며, Len(t1, MRCA(t1, t2))는 개념노드 t1로부터 공통조상노드 MRCA(t1, t2)까지 최단링크거리이다. 마지막으로 공통조상 또는 공통자손노드 둘 중 하나만 존재해도 유사도 계산이 가능토록 CDbM과 CAbM을 하나로 합하여 새로 만든 Hybrid방법은 다음 〈그림 7〉과 같다. CDbM과 CAbM, Hybrid방법 3가지 모두 유사도 계산의 결과값은 반드시 0과 1사이



〈그림 7〉 false negative 문제해결을 위한 새로운 Hybrid방법

의 값을 가져야 하며, 0 또는 1의 값을 가지게 되는 경우에는 두 개념노드 간 의미정렬관계가 성립하지 않음으로써 해당 결과값은 정렬대상에서 제외(discard)한다. 새로 만든 Hybrid방법은 공통의 조상/자손노드가 어느 하나만 존재하고 공통의 최상위 노드가 존재하지 않더라도 유사도 계산이 가능함으로써, 공통조상 또는 공통자손노드가 존재하지 않음으로써 발생하는 false negative문제를 모두 해결하였다.

4. 실험 및 평가

본 논문에서는 현실세계와 보다 유사한 환경하에 실험을 수행하기 위하여 온톨로지로 활용가능한 어휘의미사전인 워드넷을 도메인 온톨로지로 선정하였다. 온톨로지의 일종으로 간주되는 워드넷은 여러 분야에 걸쳐 어휘 간 의미유사성을 식별하기 위한 도구로 널리 활용되어 왔다. 워드넷은 약 11만 5천 건의 개념노드로 구성되어 있는데, 워드넷에서의 개념노드는 문맥에서 단어들이 서로 대체되었을 때 의미적으로 큰 차이가 없는 단어들의 집합으로 표현된다. 예를 들어 동일한 개념을 설명하기 위한 유사어휘의 집합인 {부모, 부모님, 어버이}는 하나의 개념노드가 될 수 있다. 워드넷은 이러한 개념노드 간 동의어 관계(Synonyms), 반의어관계(Antonyms), 상/하위관계(Hyper/Hypo-nyms), 부분/개체관계(Melonym) 등 수백만 건의 관계(relationship) 정보를 가지고 있으며, 이를 기반으로 현실세계와 보다 유사한 실험수행이 가능하다. 본 논문의 실험을 위하여 Bernard (2009)의 'WordNet SQL Builder'를 사용, 다음 〈표 1〉과 같이 총 82,115개의 명사형 개념노드를 추출하였다.

〈표 1〉 도메인별 명사형 노드수

도메인번호	도메인명	노드수
C03	noun.tops	51
C04	noun.act	6650
C05	noun.animal	7509
C06	noun.artifact	11587
C07	noun.attribute	3039
C08	noun.body	2016
C09	noun.cognition	2964
C10	noun.communication	5607
C11	noun.event	1074
C12	noun.feeling	428
C13	noun.food	2573
C14	noun.group	2624
C15	noun.location	3209
C16	noun.motive	42
C17	noun.object	1545
C18	noun.person	11087
C19	noun.phenomenon	641
C20	noun.plant	8030
C21	noun.possession	1061
C22	noun.process	770
C23	noun.quantity	1275
C24	noun.linkdef	437
C25	noun.shape	341
C26	noun.state	3544
C27	noun.substance	2983
C28	noun.time	1028

추출된 워드넷 명사형 노드는 〈표 1〉에서처럼 총 26개의 다양한 도메인 카테고리로 구성되어 있으며, 이는 워드넷이 각기 서로 다른 주제영역의 지역 온톨로지 간 매핑을 위한 도구로써 도메인 온톨로지의 역할을 충분히 수행할 수 있음을 보여준다. 각기 서로 다른 주제영역의 지역 온톨로지를 워드넷과 매핑하였을 때,

관련성이 높은 도메인과 관련성이 낮은 도메인 그룹을 묶어 다음 〈표 2〉와 같이 유사실험군과 대조실험군을 편성하였다. 유사/대조 실험군은 워드넷의 명사형 26개 도메인 중 3개 그룹(noun.animal, noun.plant, noun.process)을 기준으로 각기 서로 다른 도메인그룹에 속한 개념노드 간 정렬관계 빈도가 높은 관련 도메인 묶어 유사실험군으로, 정렬관계 빈도가 낮은 도메인과 묶어 대조 실험군을 편성하였다.

예를 들어 〈그림 2〉와 같이 농업을 주제로 한 지역 온톨로지의 개념노드인 'fruit'은 도메인 온톨로지인 워드넷의 도메인 'C20(noun.plant)'과 매핑이 가능하며, 이와 동시에 워드넷의 도메인 'C13(noun.food)'과도 매핑 가능하다. 이러한 특성을 고려하여 주제영역이 상호 유사한 워드넷의 도메인 C20과 C13을 하나로 묶어 유사실험군으로 지정하였으며, 주제영역이 서로 상이한 워드넷의 도메인을 묶어 대조실험군을 편성하였다. 만약 기존 Zhibiao방법의 false negative문제가 특수한 주제영역에만 발생한다면, 해당 주제영역 에서만 새로운 Hybrid방법을 적용하고 나머지 경우에는 연산횟수가 적은 기존 CAbM 혹은 Zhibiao방법이나 CDbM 방법을 적용하는 것이 보다 효율적이다. 이에 본 연구에서는 false negative문제의 특수성 여부를 판단하기 위해 유사실험군과 대조실험군을 편성, 주제영역별로 매핑 성공율을 상호 비교/분석함으로써, false negative문제가 특수한 경우에만 발생하는지 여부를 확인하였다.

〈표 3〉과 〈표 4〉는 1차 유사/대조실험군에서 임의의 두 개념노드 간 유사도 계산결과를 보여주고 있다. 〈표 3〉에서 [가]항은 기존 공통

〈표 2〉 워드넷 기반 유사/대조실험군

실험군	매핑관계 비교 도메인그룹	매핑관계 도메인명(비교 노드수)
유사실험군(1차)	C5_C6	noun.animal / noun.artifact(7509 / 11587)
	C5_C8	noun.animal / noun.body(7509 / 2016)
	C5_C13	noun.animal / noun.food(7509 / 2573)
유사실험군(2차)	C20_C5	noun.plant / noun.animal(8030 / 7509)
	C20_C6	noun.plant / noun.artifact(8030 / 11587)
	C20_C13	noun.plant / noun.food(8030 / 2573)
유사실험군(3차)	C22_C11	noun.process / noun.event(770 / 1074)
	C22_C15	noun.process / noun.location(770 / 3209)
	C22_C28	noun.process / noun.time(770 / 1028)
대조실험군(1차)	C5_C10	noun.animal / noun.communication(7509 / 5607)
	C5_C11	noun.animal / noun.event(7509 / 1074)
	C5_C15	noun.animal / noun.location(7509 / 3209)
대조실험군(2차)	C20_C10	noun.plant / noun.communication(8030 / 5607)
	C20_C11	noun.plant / noun.event(8030 / 1074)
	C20_C15	noun.plant / noun.location(8030 / 3209)
대조실험군(3차)	C22_C5	noun.process / noun.animal(770 / 7509)
	C22_C8	noun.process / noun.body(770 / 2016)
	C22_C13	noun.process / noun.food(770 / 2573)

〈표 3〉 유사실험군(1차) 평가결과

	C05_C06	C05_C08	C05_C13
[가] CAbM	5566	14357	2477
[나] CDbM	1118	176	585
[다] Hybrid	6678	14532	3051
[라] Ratio(1-[가]/[다])	16.65%	1.20%	18.81%
[마] Ratio(1-[나]/[다])	83.26%	98.79%	80.83%

〈표 4〉 대조실험군(1차) 평가결과

	C05_C10	C05_C11	C05_C15
[가] CAbM	627	37	140
[나] CDbM	423	59	79
[다] Hybrid	1050	96	219
[라] Ratio(1-[가]/[다])	40.29%	61.46%	36.07%
[마] Ratio(1-[나]/[다])	59.71%	38.54%	63.93%

조상노드 기반의 유사도 계산방법인 Zhibiao방법을 대신하기 위해 새로 개발한 CAbM을 적용 시 서로 다른 두 온톨로지 노드 간 매핑이 성공한 총 횟수이며, [나]항은 Zhibiao방법의 false negative문제를 해결하기 위해 새로 개발한 공통자손노드 기반의 유사도 계산방법인 CDbM의 성공횟수이다. 예를 들어 워드넷의 도메인 'C20(noun.plant)'과 'C13(noun.food)'을 대상으로 〈그림 3〉과 같이 개념노드 '과일(t1=fruit)'과 '식용과일(t2=edible fruit)'간 유사도 계산이 발생할 경우, CAbM은 매핑이 실패하고 CDbM은 매핑에 성공함으로써 CDbM의 총 성공횟수는 1만큼 증가하게 된다. [다]항은 공통조상노드 기반의 CAbM과 공통자손노드 기반의 CDbM을 합쳐 새로 만든 Hybrid방법의 성공횟수이다. [가]~[다]에서 최초 유사도 계산을 위한 각 방법의 임계치(threshold)는 하한값 0.5에서 시작하여 그 값을 높여가는 방법으로 실험을 수행하였으나, 각 실험군의 도메인 그룹 간 특성 차가 너무 크기 때문에 공통으로 적용 가능한 상한선 임계치값을 도출해는 것이 현실적으로 불가능하였다. 이에 휴리스틱한 방법으로 공통조상노드 또는 공통자손노드까지 거리가 1인 경우를 임시 상한선 임계치로 설정 후 도출된 계산 결과를 다시 한번 인간도메인 전문가가 적합성 판정을 수행하여 최종 온톨로지 간 매핑의 성공여부를 판단토록 하였다. [라]항은 Hybrid방법 대비 공통조상노드가 존재할 때만 유사도 계산이 가능한 CAbM방법을 적용 시, false negative문제 발생빈도를 보여준다. 마찬가지로 [마]항은 Hybrid방법 대비 공통자손노드가 존재할 때만 유사도 계산이 가능한 CDbM방법을 적용 시, false negative문제 발생빈도를 보여준다.

[라] 또는 [마]항에서 false negative문제 발생빈도가 그다지 높지 않다면, Hybrid방법 대신 CAbM 혹은 CDbM을 적용하는 것이 연산속도 측면에서 바람직하다. 〈표 3〉과 〈표 4〉

의 1차 실험결과 중 [라]항과 [마]항을 비교해보면, 대체적으로 유사실험군 평가결과인 [라]항의 값이 상대적으로 비율이 낮음을 확인할 수 있다. 이를 근거로 "유사실험군에서는 공통조상노드가 존재할 때만 유사도 계산이 가능한 CAbM방법을 적용해도 false negative 문제 발생빈도가 낮음에 따라, 유사실험군에서는 Hybrid방법 대신 CAbM방법을 적용해도 된다."는 가설을 수립할 수 있다. 이는 공통조상노드가 존재하지 않음으로써 발생하는 false negative문제가 대조실험군에서만 주로 발생하는 특수한 경우임의 가능성을 보여주며, 유사실험군에는 기존 Zhibiao방법이나 CAbM방법을 적용하는 것이 연산속도 측면에서 바람직하기 때문이다. 그러나, 1차 유사/대조실험군만의 평가결과를 통해 가설을 일반화하기에는 충분치 않으며, 유사/대조실험군을 2, 3차로 추가, 편성하여 동일한 형태의 실험을 반복 수행하였다. 추가 실험은 기존과 1차와 동일한 조건하에 수행하였으며, 2, 3차 추가실험 결과는 다음 〈표 5〉~〈표 8〉과 같다.

2/3차 실험결과 중 2차 유사실험군 〈표 5〉의 C20_C5실험결과를 살펴보면, 1차 실험결과와 달리 [라]항의 매우 높은 비율로 나타나는 기

〈표 5〉 유사실험군(2차) 평가결과

	C20_C05	C20_C06	C20_C013
[가] CAbM	189	4754	40002
[나] CDbM	200	761	1662
[다] Hybrid	389	5515	41638
[라] Ratio(1-[가]/[다])	51.41%	13.80%	3.93%
[마] Ratio(1-[나]/[다])	48.59%	86.20%	96.01%

〈표 6〉 대조실험군(2차) 평가결과

	C20_C10	C20_C11	C20_C15
[가] CAbM	1079	80	1443
[나] CDbM	149	34	23
[다] Hybrid	1228	114	1466
[라] Ratio(1-[가]/[다])	12.13%	29.82%	1.57%
[마] Ratio(1-[나]/[다])	87.87%	70.18%	98.43%

〈표 7〉 유사실험군(3차) 평가결과

	C22_C11	C22_C15	C22_C28
[가] CAbM	1204	129	828
[나] CDbM	462	28	122
[다] Hybrid	1665	157	950
[라] Ratio(1-[가]/[다])	27.69%	17.83%	12.84%
[마] Ratio(1-[나]/[다])	72.25%	82.17%	87.16%

〈표 8〉 대조실험군(3차) 평가결과

	C22_C05	C22_C08	C22_C13
[가] CAbM	0	1904	3
[나] CDbM	0	89	21
[다] Hybrid	0	1992	24
[라] Ratio(1-[가]/[다])	0%	4.42%	87.50%
[마] Ratio(1-[나]/[다])	0%	95.53%	12.50%

존 가설에 위배되는 예외상황이 발생하였음을 확인할 수 있다. 이는 곧 1차 실험군의 평가 결과와 달리 기존 Zhibiao 및 CAbM방법의 false negative문제가 대조실험군에서만 주로 발생하는 특수한 경우가 아니라, 유사/대조실험군 모든 경우에 있어 공통조상노드가 존재하지 않음으로써 발생하는 false negative문제는 서로 다른 두 온톨로지 간 노드 간 매핑 성공율을 낮추는 주요원인으로 작용할 수 있음을 입증한다. 1/2/3차 평가결과를 종합해 판단한 결과, 기존 Zhibiao 및 CAbM방법의 공통조상노드가 존재하지 않음으로써 발생하는 false negative문제는 도메인에 종류에 상관없이 모든 경우에 있어 서로 다른 두 온톨로지 간 노드 간 매핑 성공율을 낮추는 주요원인임을 확인하였다. 마찬가지로 1/2/3차 유사/대조실험군의 [마]항을 비교해보면, 유사/대조실험군 모든 경우에 있어 공통자손노드가 존재하지 않음으로써 발생하는 false negative문제는 서로 다른 두 온톨로지 간 노드 간 매핑 성공율을 낮추는 주요원인으로 작용할 수 있음을 확인할 수 있다. [마]항의 유사/대조실험군 18개 그룹 중 14개 그룹에서 공통자손노드가 존재하지 않음으로써 발생하는 false negative문제의 발생빈도가 50%를 초과하였다. 이는 개선된 Hybrid방법이 대부분의 경우에 있어 기존 유사도 계산방식의 false negative 오류를 예방함으로써, 서로 다른 개념노드 간 매핑 성공율을 높이는데 기여함을 보여준다.

5. 결론 및 향후 연구

대규모의 온톨로지 개념노드 간 의미정렬작업 수행함에 있어 자동화된 알고리즘이 실제 참인 정렬관계를 비정상으로 판정하는 false negative발생 시, 비정상으로 판정된 연산결과는 인간전문가의 검토 대상에서 제외됨에 따라 정렬작업의 정확도를 낮추는 주요요인으로 작용할 수 있다. 본 연구에서는 각기 서로 다른 두 온톨로지 개념노드 간 자동화된 유사도 계산 시 발생하는 false negative문제 해결을 위하여, 기존 의미거리기반 유사도 계산방법인 Zhibiao방법을 개선한 새로운 유사도 계산 방식인 Hybrid 방법을 개발하였다. 유사/대조 실험군을 대상으로 1/2/3차 실험을 통해 false negative문제가 대상도메인 종류에 상관없이 서로 다른 두 온톨로지 간 노드 간 의미정렬 시 매핑 성공율을 낮추는 주요원인임을 확인하였으며, 서로 다른 온톨로지 개념노드 간 의미정렬을 위한 유사도 계산을 위해서는 기존 Zhibiao 방법보다는 새로 개발한 Hybrid방법을 적용하는 것이 보다 효과적임을 확인하였다. 본 논문

에서는 현실세계와 보다 유사한 환경하에 실험 수행을 위하여 다양한 주제들로 구성된 워드넷으로부터 총 82,115개의 명사형 개념노드를 추출, 실험을 실시하였으며, 향후 워드넷 이외 다양한 온톨로지들을 대상으로 실험을 확대해 나갈 예정이다.

참 고 문 헌

김기성. 2007. 시그니처 트리를 사용한 의미적 유사성 검색 기법.『한국정보과학회논문지D: 데이터베이스』, 34(6): 546-553.

김세종. 2009. 용어를 공유하는 패턴 쌍을 이용한 의미 관계 추출.『한국정보과학회논문지C: 컴퓨팅의 실제』, 15(3): 221-225.

김수경. 2008. 새로운 N-ary 관계 디자인 기반의 온톨로지 모델을 이용한 문장의미결정.『정보관리학회지』, 25(4): 43-66.

서은경. 2009. 온톨로지 기반의 연구자정보 검색 인터페이스 설계.『정보관리학회지』, 26(2): 173-194.

이유진. 2009. 시맨틱 디지털도서관 서비스를 위한 서지 온톨로지 구축.『정보관리학회지』, 26(1): 215-230.

변영태, 황상규, 오경묵. 1999. 어휘사전 워드넷을 활용한 의미기반 웹 정보필터링.『정보처리논문지』, 6(11): 3399-3409.

Bou, Bernard. 2009. "Wordnet Sql Builder." 〈http://wnsqlbuilder.sourceforge.net〉.

Carbonetto, Andrew August. 2008. "Ontology Alignment in the Presence of a Domain Ontology." The Degree of Master of Science, University Of British Columbia.

Dimitre, A. 2006. "Information Integration via an End-to-End Distributed Semantic Web System." International Semantic Web Conference.

Hanif, Seddiqui Md., Yohei Seki, and Masaki Aono. 2006. "Automatic Alignment of Ontology Eliminating the Probable Misalignments." ASWC 2006, LNCS 4185: 212-218.

Kavalec, M. 2005. "A Study on Automated Relation Lebelling in Ontology Learning." *Ontology Learning and Population form Text: Methods, Evaluation and Applications*, IOS Press, 44-58.

Maedche, A. 2002. "Measuring similarity between ontologies." *Proceedings of the European Conference on Knowledge Acquisition and Management* (LNAI 2473), 251-263.

Thiagarajan, R., G. Manjunath, and M. Stumptner. 2008. *Computing Semantic Similarity Using Ontologies*. HPLabs Technical Report HPL-2008-87. Available from 〈http://www.hpl.hp.com/techreports/2008/HPL-2008-87.pdf〉

Wu, Zhibiao. 1994. "Verb semantics and lexical selection." *32nd Annual Meeting of the*

Association for Computational Linguistics. 133-138.

Zhang, C., Yu-Jing Wang, Bin Cui, and Gao Cong. 2008. "Semantic similarity based on compact concept ontology." *Proceeding of the 17th international conference on World Wide Web*, 1125-1126.

Ziegler, Patrick. 2006. "Detecting Similarities in Ontologies with the SOQA-SimPack Toolkit." EDBT 2006, LNCS 3896.