한국지능시스템학회 논문지 2012, Vol. 22, No. 1, pp. 42-48
http://dx.doi.org/10.5391/JKIIS.2012.22.1.42

어휘와 구조 정보에 기반한 온톨로지의 다단계 매핑

Multi-level Mapping of Ontologies Based on Lexical and Structural Information

황세찬* · 강신재**†
Se-Chan Hwang*, Sin-Jae Kang**†

* 대구대학교 대학원 컴퓨터정보공학과
** 대구대학교 정보통신대학 컴퓨터 · IT공학부

요 약

시맨틱 웹이 대두되면서 온톨로지의 사용이 점차 늘어나고 있다. 동일한 분야에 관한 온톨로지일지라도 구축 방법과 활용 형태에 따라 같은 개념이 다른 형태로 표현되거나, 다른 개념이 같은 형태로 표현될 수 있다. 이러한 온톨로지들을 공유하고 재사용하기 위해서는 온톨로지의 매핑이 필요하다. 본 논문에서는 온톨로지의 어휘 정보를 이용하여 다단계로 매핑하고, 이 결과를 기반으로 구조 정보의 유사성을 검사하는 방법을 제안한다. 온톨로지에서 어휘 정보가 부여되지 않는 블랭크 노드를 추가로 확장하여 매핑 성능을 향상시켰다. 실험을 통하여 86.38%의 F1-measure값을 얻을 수 있었다.

핵심주제어 : 온톨로지, 온톨로지 매핑, 시맨틱 웹

Abstract

Since the Semantic Web emerged, ontology has been widely used in web environment. Even ontologies belong to the same domain, they may contain same meaning different words, or different meaning same words according to their development background and the type of utilization. In order to share and reuse the ontologies, ontology mapping is required. This paper presents a ontology mapping method that consists of the initial process of multi-level mapping based on lexical information, and the second mapping process using the lexical results and structural similarity. Mapping performance was improved by additionally expanding structural information of blank nodes, which have no lexical information. Through experiments, our method achieved 86.38% in F1-measure.

Key Words : Ontology, Ontology Mapping, Semantic Web

1. 서 론

시맨틱 웹(Semantic Web)[1,2]은 현재의 인터넷과 같은 분산 환경에서 리소스(웹 문서, 각종 화일, 서비스 등)에 대한 정보와 리소스 사이의 관계-의미 정보를 컴퓨터가 처리할 수 있는 온톨로지(Ontology)로 표현하고 처리하는 기술이다. 웹의 창시자인 팀 버너스 리가 1998년에 처음 제안했고, 현재 W3C에 의해 표준화 작업이 진행 중이다.

온톨로지는 시맨틱 웹의 중요 구성요소로, 개념과 속성, 개념간의 관계를 표현하는 프로퍼티 등으로 구성된다. 컴퓨터는 온톨로지로 표현된 개념을 이해하고 지식 처리를 할 수 있는데, 추론, 증명 등의 처리에 온톨로지의 공리(axiom)와 규칙(rule)이 사용되며, 규칙 표현을 위해서 별도의 규칙 언어가 사용된다. 온톨로지의 활용이 늘어나면서 웹 서비스 개발자들이 그들의 필요에 의해 온톨로지를 만들다 보니, 구축 배경이나 활용 목적에 따라 다양한 온톨로지들이 구축되었다. 이렇게 기 개발된 여러 웹 서비스를 결합하여 원하는 검색 결과를 얻거나, 새로운 웹 서비스를 만들고자 할 때, 기존 온톨로지 정보를 서로 연동하여 재사용하기 위한 방법이 필요하게 되었는데, 이러한 기술이 온톨로지 통합(ontology integration)이다. 온톨로지 통합 기술에는 온톨로지 매핑, 온톨로지 병합 등이 있다. 온톨로지 매핑 또는 정렬(ontology mapping 또는 alignment)은 두 온톨로지의 형태를 독립적으로 유지하면서 서로 연결하는 것이며, 온톨로지 병합(ontology merging)은 여러 개의 온톨로지를 구조적으로 하나의 온톨로지로 합하여 만드는 것을 의미한다[3]. 실제적으로 온톨로지 병합은 쉽지 않기 때문에 온톨로지 매핑의 접근법을 많이 취하고 있다.

매핑 시 어려운 문제로는 특정 개념이 서로 다른 온톨로지에서 각각 다른 형태의 어휘로 표현되거나, 의미적으로 다른 개념이 같은 형태로 표현되는 경우가 있다. 또한 온톨로지에서 사용되는 어휘가 영어, 독일어, 프랑

접수일자: 2011년 8월 18일
심사(수정)일자: 2011년 12월 14일
게재확정일자 : 2012년 2월 10일
† 교신저자
이 논문은 2011학년도 대구대학교 학술연구비 지원에 의하여 연구되었음

스어 등 서로 다른 언어로 표현되는 경우도 매핑을 어렵게 하는 경우이다. 이러한 것들은 개념들 간의 단순 문자열 비교만으로는 완전한 매핑을 할 수 없으며, 개념 주변의 구조 정보, 또는 의미 정보 등을 함께 고려하여 매핑을 하여야 한다.

본 논문에서는 온톨로지의 어휘 정보를 이용하여 다단계로 매핑하고, 이 결과를 기반으로 구조 정보의 유사성을 검사하는 방법으로, 기존 유사 연구보다 온톨로지 매핑의 성능을 향상시키는 방법을 제안한다.

2. 관련 연구

2.1 온톨로지 매핑 기법의 분류

온톨로지 매핑 기법은 사용되는 정보에 따라 어휘 유사성, 구조 유사성, 의미 유사성을 각각 이용하는 기법과, 이 방법들을 결합하는 하이브리드 기법이 있다.

어휘 유사성 기법은 온톨로지에 존재하는 어휘를 대상으로 그 어휘들 간의 유사성을 측정하는 기법이다. EditDistance[4]는 어휘의 유사도를 측정하는 대표적인 기법으로, 비교 대상인 두 어휘 중 한 어휘를 기준으로 정하고, 다른 어휘가 기준 어휘와 똑같이 되기 위한 이동, 삽입, 삭제, 변경 등의 최소 횟수를 수치로 나타낸다. 여기서 수치가 낮을수록 유사하다고 판단한다. <그림 1>에서 어휘 유사도의 계산 예를 제시하고 있다.

기준어휘	: *abcd*	*conference*
변화될어휘	: *abce*	*reference*
EditDistance	: 1	3

그림 1. 어휘 유사도 계산 예
Fig. 1. Calculation example using lexical similarity

이 방법은 두 온톨로지의 어휘들 간에 동의어와 동형이의어가 있는 경우, 제대로 매핑하지 못하거나 잘못된 매핑을 할 우려가 있다. 동의어란 유사한 뜻을 나타내지만 그 형태가 다른 어휘를 뜻하고, 동형이의어는 같은 형태의 어휘지만 다른 뜻을 가진 어휘를 뜻한다.

구조 유사성 기법은 매핑의 대상이 되는 두 개념의 주변 정보(상위 개념, 하위 개념, 연결된 관계 속성 등)를 이용하여 유사성을 측정하는 기법이다.

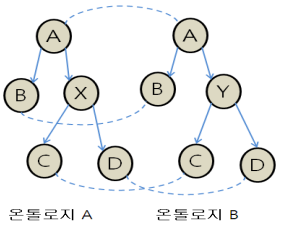

그림 2. 구조 유사성에 의한 매핑 예
Fig. 2. Mapping example using structural similarity

예를 들어 <그림 2>에서 제시된 온톨로지 A와 B에 각각 존재하는 X, Y 개념이 매핑 후보라고 가정할 때, 개념 X와 Y는 어휘가 다르기 때문에 어휘 유사성 기법으로는 매핑이 되지 않으나, 각각의 상위 개념이 A이고, 하위 개념이 C, D이며, 형제 개념이 B로 동일하므로 구조 유사성 기법에서는 매핑이 될 수 있다. 물론 상기 예에서는 주변 정보가 완전히 일치하였지만, 실제로는 부분적으로만 일치하는 경우도 발생할 수 있으므로, 주변 정보의 일치 정도를 수치화하고 임계치를 두어서 매핑 여부를 결정하여야 한다. 구조 유사성 기법은 매핑이 되어야 할 동의어와 매핑이 되지 말아야 할 동형이의어의 구분이 가능하다. 하지만 주변의 개념이 사전에 매핑되어 있어야 적용이 가능하기 때문에 어휘 유사성 기법이 선행되어야 할 필요가 있다.

온톨로지에 포함된 정보만으로 매핑이 어려운 경우에는 의미 정보가 포함된 외부의 자원이 활용된다. 영어를 대상으로 의미 유사성 기법을 적용하고자 할 때 많이 사용되는 외부 자원은 워드넷(WordNet)이 있다. 워드넷은 한 언어의 단어 집합에 대해 같은 의미를 갖는 어휘들의 모음인 동의어집합(synonym set, synset, 신셋)을 명사, 동사, 형용사, 부사의 각 품사별로 별도로 정의하고, 신셋 간의 의미관계를 표현한 어휘의미망이다. 의미관계로는 상위어(hypernym), 하위어(hyponym), 전체어(holonym), 부분어(meronym) 등이 있다. 이 기법은 온톨로지 내 어휘를 외부 자원에 먼저 매핑하여야 비교 대상인 어휘 간 의미의 유사성을 비교할 수 있는데, 이를 위해서는 어휘 의미 중의성 해소(word sense disambiguation)라는 어려운 과정이 필요하게 된다. 또한 외부 자원의 구축 상태와 그 양에 따라 매핑 성능에 영향을 미치며, 언어의 종류에 따라 외부 자원의 확보가 쉽지 않을 수도 있다.

하이브리드 기법은 위 세 가지 기법들을 상호 보완하여 매핑의 성능을 향상시키는 방법이다. 각 기법들은 직렬 또는 병렬로 실행될 수 있는데, 각 기법의 성능과 어떤 평가 지표(정확률, 재현율 등)를 우선하는가에 따라 적용 방법이 달라질 수 있다. 대부분의 매핑 기법들이 여기에 속한다.

2.2 매핑 기법별 기존 연구

[5]는 어휘와 의미 정보를 하이브리드 기법으로 사용하였는데, 1단계는 온톨로지 내 어휘 그대로 비교하고, 2단계는 복합어인 어휘에 속해 있는 관사, 전치사, 접속사를 제거하고, 나머지 단어 간의 문자열을 비교한다. 3단계는 워드넷의 동의어 집합에 단어가 포함되어 있는지를 검사한다. 마지막 4단계는 SUMO(The Suggested Upper Merged Ontology)[6,7]와 MILO(the Mid-level Ontology)[8] 등의 글로벌 온톨로지를 이용한다. 구조 정보의 이용없이 어휘 정보와 동의어 정보만을 이용하므로 동형이의어와 어휘적으로 알기 힘든 축약어 등은 매핑을 할 수 없다.

[9]는 도메인 온톨로지를 구조적인 방법으로 사용하였다. 먼저 매핑 대상 어휘를 A, B라 할 때, 도메인 온톨로지에서 A, B와 최단 거리에 있는 공통 부모 노드를 찾고, A, B와 최단 거리에 있는 공통 자식 노드도 찾는다. 두 거리 중 가까운 값을 유사도 값으로 결정하여 사용한다. 이는 도메인 온톨로지가 확보되고, 어휘를 도메인 온톨로지에 제대로 매핑하기 위하여 어휘 의미 중의성 해소 과정이 추가로 필요한 문제가 있다.

[10]은 어휘 유사성 기법(Edit Distance)과 의미 유사성

기법(워드넷의 동의어 집합 이용)을 각각 계산하여 나온 유사도 값에 정규화 상수를 곱하여 더한 후, 임계 치 이상의 값을 매핑 매트릭스에 저장한다. 이 과정을 매트릭스가 더 이상 변화가 없을 때까지 반복하는 방법이다.

[11]은 다단계 방법으로 온톨로지 매핑을 수행하는데, 1단계로 어휘 유사성 검사를 하여 매핑하고, 그 후 매핑되지 않은 것을 대상으로 워드넷을 활용한다. 상위어, 하위어, 전체어, 부분어, 설명문에 속한 단어들을 슈퍼 워드 셋으로 추출한 후, 이를 상호 비교하여 유사도를 계산한다. 또한 속성 매핑 정제 과정을 통해 매핑된 결과를 후처리하였다. 이 연구는 슈퍼 워드 셋을 통해 어휘 의미의 중의성을 어느 정도 해결하고자 하였다. 하지만 실제 온톨로지에서 자주 등장하는 축약어나 일상적으로 잘 사용하지 않는 어휘가 나타나는 경우에는 이 기법을 적용할 수가 없다.

[12]는 온톨로지를 작은 세그먼트 단위로 나누고, 반복적으로 매핑하는 Anchor-Flood 알고리즘을 제안하여 효율성과 확장성을 꾀하였다. 먼저, 워드넷을 이용한 어휘 유사도를 계산하여 기초 데이터를 형성한다. 이 가운데 한 쌍을 선택하여 앵커(anchor)로 정하고 주변 정보를 그룹화 한다. 주변 정보는 계층체계 상에서 조상 집합, 자손 집합, 이웃 집합 등을 구하여 사용한다. 구조적 유사성 검사는 먼저, 주변 개념만을 가지고 유사도 계산을 하고, 그 다음 주변 속성만 가지고 유사도 계산을 한 후, 이 값들을 평균 내어서 사용하였다. 이 방법은 일 대 다 매핑을 효율적으로 처리하지 못하고, 워드넷을 의미 유사도를 계산하는 데에 제대로 활용하지 못하였다.

3. 제안 시스템

3.1 제안 시스템 구조

본 논문에서 제안하는 온톨로지의 다단계 매핑 시스템의 전체 구조는 <그림 3>과 같다.

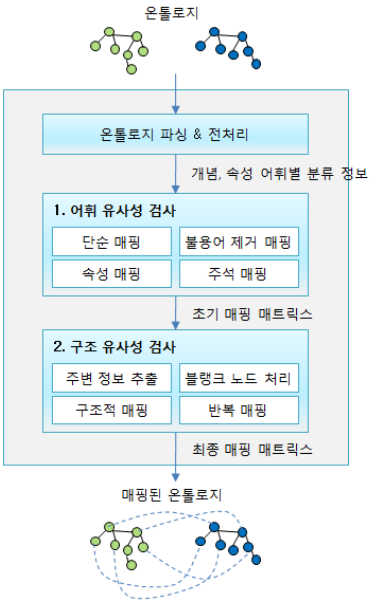

그림 3. 온톨로지 매핑 시스템
Fig. 3. Ontology mapping system

매핑할 두 개의 온톨로지를 입력으로 받아 파싱을 하고 전처리 과정을 거쳐 개념(Concept)과 데이터 타입 속성(DatatypeProperty), 객체 속성(ObjectProperty)[1], 주석(Comment)으로 분류한다. 이렇게 분류된 데이터는 어휘 유사성 검사 과정에서 사용하게 된다. 이 단계에서는 어휘 정보만을 이용하여 초기 매핑 매트릭스를 구축한다. 구조 유사성 검사 단계에서는 구축된 매트릭스의 정보를 이용하여 다중 매핑을 수행한다. 이 과정에서는 더 이상 매트릭스 정보의 변화가 없을 때까지 반복적으로 매핑을 수행한다.

3.2 어휘 유사성 검사 단계

이 단계에서는 Edit Distance를 이용하여 어휘 유사도를 검사한다. 어휘 자체를 그대로 비교하는 “단순 매핑”, 매핑에 불필요한 불용어(stopwords)를 제거하고 매핑하는 “불용어 제거 후 매핑”, 데이터 타입 속성과 객체 속성을 병합하여 매핑하는 “속성 교차 매핑”, 그리고 어휘를 설명하는 주석문의 내용으로 매핑을 시도하는 “주석 매핑”의 총 4단계 과정을 거친다.

어휘 유사성 검사 단계에서는 최대한의 정확률을 얻기 위해 완전히 일치하는 경우에만 매핑을 하게 되며, 매핑되지 않은 어휘들은 다음 절에서 설명하는 구조 유사성 단계에서 처리하게 된다. <그림 4>는 어휘 유사성을 검사하는 단계를 제시하고 있다.

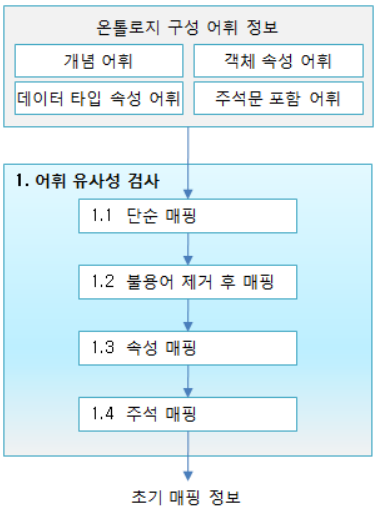

그림 4. 어휘 유사성 검사 방법
Fig.4. Lexical similarity checking

3.2.1 단순 매핑

고유명사에서 사용된 대문자를 소문자로 바꾸었을 때, 그 의미가 바뀌는 경우도 있기 때문에, 단순 매핑 과정에서는 어휘의 대소문자를 구분하여 비교한다. 이 단계에서는 개념, 데이터 타입 속성, 객체 속성 별로 매핑을 시도하며, 매핑된 결과는 매핑 매트릭스에 저장한다. 매핑할 온톨로지가 A, B라고 가정할 때, 매핑 매트릭스는 온톨로지 A의 어휘들을 가로로, 온톨로지 B의 어휘들을 세로로 나열한 테이블이며, 교차되는 셀에는 해당 어휘들의 매핑 여부가 저장된다. 어휘가 완전히 일치하는 경

1) http://www.w3.org/TR/owl-guide/#DefiningProper ties

우에만 매핑 정보가 기록된다.

3.2.2 불용어 제거 후 매핑

이 단계에서는 속성 'Name'과 'IsName' 같이 전 단계에서 불용어로 인해 매핑되지 않았던 것들을 매핑시키기 위하여 불용어들을 제거하는 과정을 거친다. 개념, 데이터 타입 속성, 객체 속성 별로 매핑을 하며, 새로이 매핑된 정보는 1-1 단계에서 구축된 매트릭스에 추가된다.

3.2.3 속성 교차 매핑

속성 교차 매핑 단계는 같은 의미의 속성이지만 데이터 타입 속성과 객체 속성으로 다르게 분류된 속성을 매핑하는 단계이다. 예를 들어 'firstname'이 데이터 타입 속성으로 분류된 온톨로지도 있고, 객체 속성으로 분류된 온톨로지도 있기 때문에 이를 매핑하기 위한 과정이다. 1-2 단계와 같이 불용어를 제거한 후, 데이터 타입 속성과 객체 속성을 병합하고 매핑한다. 이 과정에서 새로이 매핑된 값들도 매트릭스에 추가된다.

3.2.4 주석 매핑

이 단계에서는 온톨로지 어휘를 설명하는 주석문 정보를 이용한다. 어휘는 다르게 표현되었지만 어휘의 주석문 내용이 같으면 같은 개념으로 볼 수 있기 때문이다. 개념 어휘, 데이터 타입 속성 어휘, 객체 속성 어휘별로 매핑을 하게 되며, 새로이 매핑된 값들도 매트릭스에 추가된다.

3.3 구조 유사성 검사 단계

구조적인 유사성을 검사하기 위해 고려해야 할 사항으로는 "주변 정보를 어디까지 볼 것인가?", "속성의 방향성을 고려할 것인가?", "주변 정보 가운데 중복된 어휘 정보를 허용할 것인가?" 등이 있으며, 어떠한 선택을 하는가에 따라 다양한 방법이 존재할 수 있다. 본 연구에서는 다양한 실험을 수행하고 이 가운데 최상의 성능을 보인 정보의 조합을 선택하였다.

3.3.1 주변 정보 추출

구조 유사성 검사를 하려면 매핑 대상 어휘의 주변 정보를 추출하여야 한다. 너무 많은 정보를 추출하면 일반화에 문제가 있고, 너무 적은 정보를 추출하면 변별력이 떨어지기 때문에 적당한 양의 정보를 추출할 필요가 있다.

본 연구에서는 2단계 확장을 통하여 주변 정보를 추출하였다. 2단계 확장이란 매핑 대상과 직접 연결된 것과 링크를 통해 2차적으로 연결된 것을 모두 추출하는 것을 말한다. <그림 5>에서 HigherEducationInstitution 개념을 기준으로 주변 정보를 추출한다고 가정할 때, 점선으로 표기된 그룹이 1단계 확장이고, 실선으로 표기된 그룹이 2단계 확장에 해당한다.

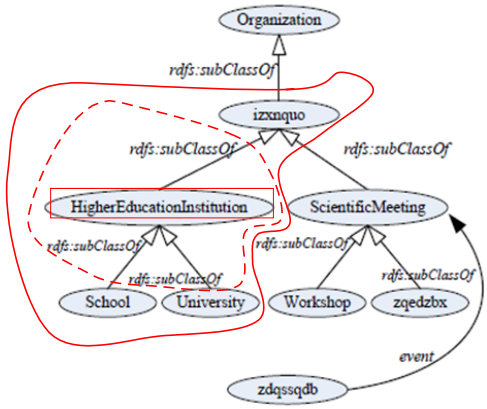

그림 5. 주변 정보의 2단계 확장 추출
Fig. 5. Two-step extended extraction of environmental information

3.3.2 구조 유사성 검사

<그림 6>의 온톨로지 어휘들은 A1, A2, a, b와 같은 기호로 표기[2]되어 있다. <그림 6>에서 C1, C2가 매핑 대상이라고 할 때, 주변 정보(개념과 속성[3] 등)들을 보면 의미를 알 수는 없지만 같은 어휘들이 있음을 알 수 있다. 온톨로지 A와 B를 대상으로 어휘 유사성 검사를 하여 개념 A1, A2, B1과 속성 a, b, e 등이 매핑되었다고 가정하자. 구조 유사성 검사를 위해서는 먼저 매핑 대상 C1, C2의 주변 정보를 모아서 그룹 G(C1), G(C2)를 생성한다. 생성된 그룹에는 2단계 확장 범위 안에 속하는 모든 개념과 속성이 포함되며, 중복된 어휘가 포함될 수 있다.

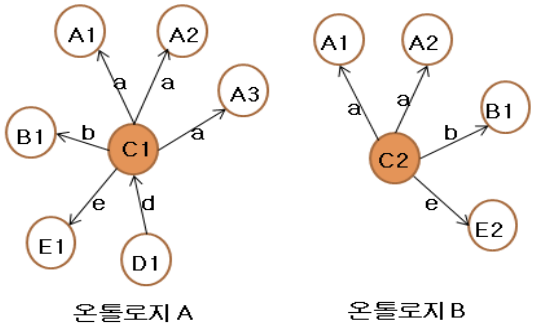

그림 6. 온톨로지 예
Fig. 6. Ontology examples

본 논문에서 제안하는 구조 유사도 계산 수식은 다음과 같다.

$$S_{struct}(C_1, C_2) = \frac{|G(C_1) \cap G(C_2)|}{\min(|G(C_1)|, |G(C_2)|)} \quad (1)$$

C_i : 매핑 대상 개념 또는 속성

2) 구별을 위해 개념 어휘는 대문자로, 속성 어휘는 소문자로 표기하였다.
3) 데이터 타입 속성이거나 객체 속성일 수 있다.

$G(C_i)$: C_i의 주변 정보를 2단계 확장하여 만든 그룹 중복되는 데이터가 존재할 수 있으며, 속성의 경우에는 들어오는 링크인지, 나가는 링크인지 구분하여 저장함

$|G(C_i)|$: $G(C_i)$ 그룹 내의 데이터 개수

<그림 6>에서 제시된 온톨로지 A, B를 대상으로 계산해 보면 G(C1)의 개수는 12[4]이고, G(C2)의 개수는 8[5]이므로, 구조 유사도 수식 S_{struct}의 분모는 8이 된다. G(C1)과 G(C2) 두 그룹에 공통으로 존재하는 어휘의 개수는 7[6]이므로, 구조 유사도 수식은 7/8이 된다. 유사도 값이 특정 임계값을 넘는 경우에 매핑을 하게 된다.

속성간 매핑에서도 동일하게 수식 (1)을 적용한다. <그림 7>에서 온톨로지의 매핑 대상이 p1, p2 일 때 각 대상의 주변 정보를 모아서 그룹 G(p1)[7], G(p2)[8]를 생성한다. subPropertyOf 속성의 경우, 각 그룹에서 두 번씩 나타나는데, 매핑 대상인 p1, p2로 들어오는 subPropertyOf 링크는 상위 속성으로부터의 링크이고, p1, p2에서 나가는 subPropertyOf는 자식 속성으로 가는 링크이므로 다른 의미를 가지고 있다. 그러므로 구조 유사성 검사 단계에서 속성의 방향성을 고려할 필요가 있다.

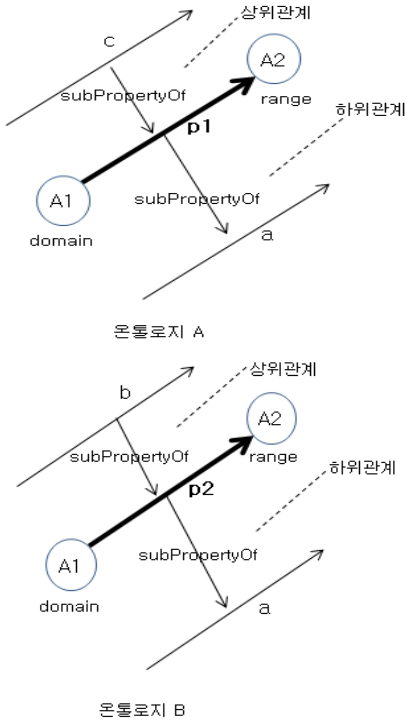

그림 7. 속성 어휘 매핑의 예

Fig. 7. Mapping example between properties

4) A1, A2, A3, B1, D1, E1, a, a, a, b, d, e

5) A1, A2, B1, E2, a, a, b, e

6) A1, A2, B1, a, a, b, e

7) a, c, domain, range, A1, A2, subPropertyOf, subPropertyOf

8) a, b, domain, range, A1, A2, subPropertyOf, subPropertyOf

<그림 7>에서 구조 유사도를 계산해 보면 G(p1)의 개수는 8이고, G(p2)의 개수도 8이므로 분모의 값은 8이 되고, G(p1)와 G(p2) 두 그룹에 공통으로 존재하는 어휘의 개수는 7이므로 유사도 값은 7/8이 된다.

만약 주변 정보 그룹에 블랭크 노드(이름 없는 노드)가 포함되어 있을 경우[9]에는, 블랭크 노드를 기준으로 추가로 2단계 확장을 하게 된다.

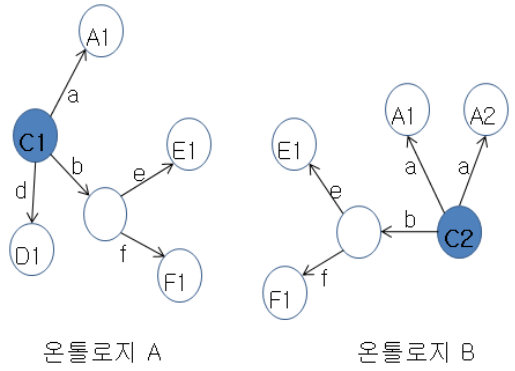

그림 8. 블랭크 노드가 포함된 온톨로지

Fig. 8. Ontologies including blank nodes

<그림 8>의 C1에서 속성 b로 연결되어 있는 이름 없는 노드가 블랭크 노드이다. 설명의 편의를 위해 온톨로지 A의 블랭크 노드를 B1이라 지칭하고, 온톨로지 B의 블랭크 노드를 B2라 가정한다.

블랭크 노드를 기준으로 나가는 링크의 속성(위 예에서는 e, f)만을 대상으로 2단계 확장 그룹을 구한 후, 수식 (1)을 적용하여 S_{struct}(B1, B2)를 계산한다. 임계치를 초과하는 경우, B1과 B2는 매핑이 된다. 매핑이 된 경우에는 S_{struct}(C1, C2) 수식의 분자 부분에 포함되어 계산에 사용된다. 이 과정은 어휘 유사성 검사 단계에서 누락된 블랭크 노드에 관한 매핑 정보를 추가로 계산하여 구조 유사성 검사를 진행하는 것이다.

4. 실험

4.1 데이터 집합

실험은 OAEI(Ontology Alignment Evaluation Initiative)에서 제공하는 데이터 집합을 이용하였다. OAEI는 2004년부터 워크샵을 개최하고 있으며, 온톨로지 매핑을 위한 데이터 집합을 제공하고 있다. 데이터 집합은 benchmarks, anatomy, conference, fishery gears, directory, library 등 여러 가지가 있다.

본 연구에서는 benchmarks 데이터 집합을 이용하며, 이 데이터 집합은 하나의 온톨로지를 여러 가지 형태로 인위적으로 변형하여 110 여 가지의 서로 다른 온톨로지로 만든 것이다. OWL-DL 형식으로 만들어졌으며 RDF/XML 포맷을 사용한다. #101 온톨로지를 소스 온톨로지라고 하는데, 33개의 이름 있는 개념과 24개의 객

9) 실제로 웹 상에 공개되어 있는 온톨로지에 블랭크 노드가 포함되어 있는 경우가 많으며, 시스템 내부적으로는 임의의 식별자를 부여하여 저장한다.

체 속성, 40개의 데이터 타입 속성, 56개의 이름 있는 인스턴스, 그리고 20개의 이름 없는 인스턴스를 가지고 있다[13]. <표 1>은 본 연구의 실험에서 사용한 benchmarks 데이터 집합의 구성에 대해 보여주고 있다.

표 1. benchmarks 데이터 집합의 구성
Table 1. Benchmarks data set

분류	설명	개수
1xx	OWL-Lite 의 제약조건을 수용하는 온톨로지	4
2xx	여러 가지 형태로 제약조건을 만들어 적용한 온톨로지	66
3xx	실제로 사용되고 있는 온톨로지	4

성능 평가 지표로는 OAEI 워크샵에서 사용한 정확률, 재현율, F1-measure를 이용하며, 데이터 집합에 포함된 온톨로지 간 매핑 성능을 평균하여 제시한다.

4.2 실험 결과

<표 2>에 어휘 유사성 검사 단계별로 측정된 정확률과 재현율, F1-measure를 정리하였다. 어휘 유사성 검사에서는 확실한 것만 매핑하였기 때문에 정확률이 대략 91%가 나왔다. 하지만 아직 매핑되지 않은 어휘가 많이 있어서 재현율은 대략 60%가 나왔다. 각 단계를 거칠 때 마다 정확률은 0.5% 미만으로 떨어지지만 재현율은 4.5% 향상, F1-measure는 약 3.5%의 성능 개선이 나타났다. 어휘 유사성 검사를 통해 생성한 초기 매핑 매트릭스를 가지고, 구조 유사성 검사를 수행한 결과는 <표 3>에 나타나 있다.

표 2. 어휘 유사성 검사 실험 결과 (단위:%)
Table 2. Experimental results using lexical similarity

어휘 유사성	정확률	재현율	F1
단순 매핑	91.68	56.09	69.60
불용어 제거 후 매핑	91.65	57.09	70.36
속성 교차 매핑	91.63	57.96	71.00
주석 매핑	91.34	60.69	72.92

표 3. 구조 유사성 검사 실험 결과 (단위:%)
Table 3. Experimental results using structural similarity

임계치	정확률	재현율	F1
0.5	79.27	85.18	82.12
0.6	83.99	85.66	84.82
0.7	87.24	85.53	86.38
0.8	87.68	84.31	85.96
0.9	88.62	82.69	85.96

<표 3>은 어휘 유사성 검사와 구조 유사성 검사를 순차대로 수행한 후의 결과를 나타내는 것으로, 임계치의 값이 0.7일 때 F1-measure의 값이 가장 좋음을 알 수 있다. 임계치가 높을수록 정확률은 오르며 재현율은 하락하는 것을 볼 수 있다. 블랭크 노드에 대한 처리를 하지 않았을 때는 정확률 95.01%, 재현율 76.16%, F1 83.63%의 성능을 보였다. 블랭크 노드를 구분하여 추가 확장하였을 때, F1이 2.75% 정도 향상됨을 알 수 있다.

<표 4>에서 OAEI 2009에 참가한 여러 시스템 중에서 본 연구와 비슷한 접근법(어휘 유사성 기법과 구조적 유사성 기법)을 취한 시스템과 본 연구에서 제안한 시스템의 성능을 비교하였다. 제시된 모든 시스템이 OAEI 2009에서 제공한 동일한 온톨로지 데이터 집합을 사용하였으므로, 객관적인 비교가 가능하며, 본 연구에서 제안한 시스템이 가장 좋은 성능을 보임을 알 수 있다.

표 4. OAEI 2009 시스템과의 성능비교
Table 4. Comparison with relevant OAEI 2009 systems

시스템	정확률	재현율	F1
제안 시스템	87.24	85.53	86.38
AgrMaker[11]	96.00	79.00	85.81
DSSim[14]	97.00	76.33	84.63
GeRoMe[15]	86.67	77.00	81.30
kosimap[16]	88.33	68.67	76.33

5. 결론 및 향후 연구과제

차세대 웹이라 불리는 시맨틱 웹의 핵심 구성요소인 기존 온톨로지들을 연동하여 재사용하기 위해, 어휘 정보와 구조 정보를 이용하여 온톨로지를 다단계로 매핑하는 방법을 제안하였다. 어휘 유사성 검사 방법은 유사어와 동형이의어의 구분을 하지 못하고, 구조 유사성 검사 방법은 주변의 매핑 정보가 없을 경우, 적용하기 어려운 문제가 있다. 이러한 단점들을 상호 보완하기 위하여, 어휘 유사성 검사를 높은 정확률로 먼저 수행하여, 구조 유사성 검사를 위한 초기 매핑 데이터를 만들고, 이 데이터를 기반으로 어휘 유사성 검사에서 매핑이 불가능하였던 것을, 블랭크 노드 처리와 제안한 구조 유사도 수식을 통해 매핑을 수행하였다. 실험 결과, F1-measure 기준으로 86.38%의 성능을 보여 본 논문과 유사하게 어휘와 구조 정보를 사용한 기존 연구보다 높은 성능을 보였다. 향후에는 워드넷을 이용한 다양한 기법들[17-18]을 응용하여 온톨로지 어휘의 의미 중의성 해소 기법을 개발하고자 한다.

참 고 문 헌

[1] 김중태, *웹 2.0 시대의 기회 시맨틱 웹*, 서울 : 디지털미디어 리서치, 2006.
[2] T. Berners-Lee, J. Henderler. & O.Lassila . *The*

semantic web, scientific American, 2001.
[3] 안성준, 김우주, 박상언, “최적 온톨로지 매핑 방법론에 관한 연구”, 한국정보시스템학회 학술대회, pp. 457-462, 2007.
[4] Jerome Euzenat, Pavel Shvaiko. *Ontology Matching*, springer, 2007.
[5] John Li. "LOM: Lexicon-based Ontology Mapping Tool", in Proceedings of the performance Metrics for Intelligent Systems, pp. 1-5, 2004.
[6] Ian Niles, and Adam Pease. "Toward a Standard Upper Ontology", in Proceedings of the International Conference on Formal Ontology in Information Systems, pp. 2-9, 2001.
[7] Adam Pease, and Ian Niles, "Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology", in Proceedings of the IEEE International Conference on Information and Knowledge Engineering, pp. 412-416, 2003.
[8] Ian Niles, and Allan Terry. "The MILO: A General Purpose, Midlevel Ontology", Proceedings of the International Conference on Information and Knowledge Engineering, pp. 15-19, 2004.
[9] 황상규, 변영태, “의미거리측정방법을 활용한 분산 온톨로지 간 자동 정렬 방법 연구”, *정보관리 학회지*, 26(4), pp. 319-335, 2009.
[10] 안진현, 최기선, “반복적 알고리즘을 이용한 온톨로지 매핑”, 한글 및 한국어 정보처리 학술대회, pp. 14-18, 2009.
[11] 곽정애, “의미 관계 유사성 기반의 온톨로지 매칭과 온톨로지 교차매칭의 정제 기법”, *이화여대 대학원 컴퓨터공학과 박사학위 논문*, 2010.
[12] MD. Hanif Seddiqui, Masaki Aono, "An efficient and scalable algorithm for segmented alignment of ontologies of arbitrary size", *Web Semantics: Science, Services and Agents on the World Wide Web*, pp. 344-356, 2009.
[13] OAEI. *Results of the Ontology Alignment Evaluation Initiative 2009*, OAEI, 2009.
[14] Miklos Nagy, Maria Vargas-Vera, and Piotr, Stolarski, "DSSim Results for OAEI 2009", The Fourth International Workshop on Ontology Matching, pp. 160-169, 2009.
[15] Christopher Quix, Sandra Geisler, David Kenache and Xiang Li, "Resuits of GeRoMeSuite for OAEI 2009", The Fourth International Workshop on Ontology Matching, pp. 170-176, 2009
[16] Quentin Reul and JeZ. Pan, "KOSIMap: ontology alignments results for OAEI 2009", The Fourth International Workshop on Ontology Matching, pp. 177-185, 2009.
[17] 강신재, 강인수, “워드넷과 구글에 기반한 온톨로지 개체의 일반화,” *한국지능시스템학회 논문지*, 제19권, 3호, pp. 363-370, 2009.
[18] 강신재, 강인수, 남세진, 최기선, “WordNet을 매개로 한 CoreNet-SUMO의 매핑,” *한국지능시스템학회 논문지*, 제21권, 2호, pp. 276-282, 2011.

저 자 소 개

황세찬(Se-Chan Hwang)
2009년 : 대구대학교 컴퓨터IT공학부 공학사
2011년 : 대구대학교 대학원 컴퓨터정보공학과 석사

관심분야 : 자연어처리, 시맨틱 웹, 온톨로지 매핑
Phone : 053-850-4464
E-mail : scsunbi@naver.com

강신재(Sin-Jae Kang)
한국지능시스템학회 논문지, 제 21권 제 2호 참조